EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLOGICO NACIONAL DE MEXICO

INSTITUTO TECNOLÓGICO DE CIUDAD MADERO

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

DOCTORADO EN CIENCIAS DE LA INGENIERÍA

TESIS

APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING PARA LA PREDICCIÓN DE GELIFICACIÓN DE ORGANOGELADORES BENZOATO USADOS EN REMOCIÓN DE CONTAMINANTES EN SISTEMAS EFLUENTES

Que para obtener el grado de:
Doctora en Ciencias de la Ingeniería

Presenta
M.C. Virginia Loredo Pong
D00070411
No. CVU de CONACyT 660121

Directora de Tesis
Dra. María Lucila Morales Rodríguez
No. CVU de CONACyT 211781

Co-directora de Tesis
Dra. Nancy Patricia Díaz Zavala

Ciudad Madero, Tamaulipas marzo 2023

Instituto Tecnológico de Ciudad Madero
Subdirección Académica
División de Estudios de Posgrado e Investigación

Ciudad Madero, Tamaulipas, 13/diciembre/2022

OFICIO No. : U.168/22
ASUNTO: AUTORIZACIÓN DE IMPRESIÓN DE TESIS

C. VIRGINIA LOREDO PONG
No. DE CONTROL D00070411
P R E S E N T E

Me es grato comunicarle que después de la revisión realizada por el Jurado designado para su Examen de Grado de Doctorado en Ciencias de la Ingeniería, se acordó autorizar la impresión de su tesis titulada:

"APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING PARA LA PREDICCIÓN DE GELIFICACIÓN DE ORGANOGELADORES BENZOATO USADOS EN REMOCIÓN DE CONTAMINANTES EN SISTEMAS EFLUENTES"

El Jurado está integrado por los siguientes catedráticos:

PRESIDENTA:	DRA.	MARÍA LUCILA MORALES RODRÍGUEZ
SECRETARIA:	DRA.	NANCY PATRICIA DÍAZ ZAVALA
PRIMER VOCAL:	DR.	NELSON RANGEL VALDEZ
SEGUNDO VOCAL:	DR.	LUCIANO AGUILERA VÁZQUEZ
TERCER VOCAL:	DRA.	NOHRA VIOLETA GALLARDO RIVAS
SUPLENTE:	DR.	JUAN JAVIER GONZALEZ BARBOSA
DIRECTORA DE TESIS:	DRA.	MARÍA LUCILA MORALES RODRÍGUEZ
CO-DIRECTORA:	DRA.	NANCY PATRICIA DÍAZ ZAVALA

Es muy satisfactorio para la División de Estudios de Posgrado e Investigación compartir con usted el logro de esta meta. Espero que continúe con éxito su desarrollo profesional y dedique su experiencia e inteligencia en beneficio de México.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Por mi patria y por mi bien"®

MARCO ANTONIO CORONEL GARCÍA
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE CIUDAD MADERO
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

c.c.p.- Archivo
MACG 'LFCS'

Av. 1° de Mayo y Sor Juana I. de la Cruz S/N Col. Los Mangos C.P. 89440 Cd. Madero, Tam.
Tel. 01 (833) 357 48 20, ext. 3110, e-mail: depi_cdmadero@tecnm.mx
tecnm.mx | cdmadero.tecnm.mx

DECLARACIONES DE ORIGINALIDAD, PROPIEDAD INTELECTUAL, CESION DE DERECHOS Y/O CONFIDENCIALIDAD

Yo, Virginia Loredo Pong, en mi calidad de autor manifiesto que este documento de tesis es producto original de mi trabajo y que no infringe derechos de terceros, tales como derechos de publicación, derechos de autor, patentes y similares. Por lo tanto, la obra es de mi exclusiva autoría y soy titular de los derechos que surgen de la misma. Asimismo, declaro que en las citas textuales que he incluido y en los resúmenes que he realizado de publicaciones ajenas, indico explícitamente los datos de los autores y las publicaciones. En caso de presentarse cualquier reclamación o acción por parte de un tercero en cuanto a los derechos de autor sobre la obra en cuestión, asumiré toda la responsabilidad y relevo de ésta a mi director de tesis, así como al Tecnológico Nacional de México, Instituto Tecnológico de Ciudad Madero y a sus respectivas autoridades. Cd. Madero, Tamaulipas. marzo 2023.

Declaro que este documento de tesis es un trabajo original, indicando de manera explícita en las referencias bibliográficas los datos de las publicaciones, así como sus autores, de aquellos trabajos a los cuales se hizo cita y referencia.

Acepto total responsabilidad en caso de infringir con las leyes de derechos de terceros, así como cualquier reclamación derivada de este documento de tesis que este en relación con los derechos de propiedad intelectual, exonerando de cualquier responsabilidad tanto a mi director de tesis como al Instituto Tecnológico de Ciudad Madero

A mi amada familia.

Entre la primera palabra escrita en esta tesis y la última existe un mundo, el que existía al inicio, y el que ahora se reconstruye con las partes de quién la escribe, mientras se reinventa una vez más.

Tras los sueños puestos en la primera página, al comienzo de una aventura que desconocía, no escribe la misma persona, quién escribe ahora se conoce más, conoce sus fuerzas, limitaciones, el dolor capaz de soportar, y la entereza que le dan sus raíces para, aún rota, seguir escribiendo una historia con sueños nuevos cada día, para tener la fe en el corazón, de que la vida siempre vuelve a su cauce, a llenarnos de nuevas ilusiones.

Mi familia, la que comenzó este sueño a mi lado, no lo termina completa conmigo, sin embargo, todo su amor está en cada página, en cada idea, cada lágrima involucrada en lograr este reto, por ellos, para ellos, siempre. Por mis padres, mis raíces, por su herencia de sostenerme siempre firme no importa la fuerza de los vientos, por ellos que ayudaron a construir mis alas, sin saber que un día también se sostendrían de ellas, con el amor de mi corazón.

Para mi madre que me da vida cada día, y un motivo fuerte para ser mejor, más integra, tan fuerte como ella, quien me recuerda cuando lo olvido, que no debo dejar de ser yo, más cuando parece que todo está en contra.

Para mi Padre, mi cómplice y mi espejo, quién ahora no sólo está cuando lo veo, quién ahora es mi sombra y mi guardián.

Para Petra, la compañera de mi vida, la que fiel estuvo en cada página escrita a mi lado, y en mi corazón en las últimas.

Que el amor siempre nos permita ser parte uno del otro, aún rotos, aún en la ausencia, cuando la esperanza es grande, y cuando la oscuridad invade.

A mi familia elegida, y compañeros de camino.

Con la fe de que siempre me acompañan, de la manera en que la vida los mantenga a mi lado, cierro este capítulo con más que un crecimiento profesional, también uno espiritual, personal, y con la misma incertidumbre que al inicio, esperando que Dios me lleve al mejor lugar siempre, sin importar lo duro que parezca, dónde yo sea necesaria para dar lo que la vida pida de mí. Y que ahí, siempre ponga los hermanos de amor que en cada aventura coloca en mi camino.

Hermanos de vida, hermanas por sororidad, que me dieron la luz cuando la mía se pausó. Un pedacito de mi corazón es de cada uno de ellos. Anna Laura, Jessy Bustos, Jessy Lozano, Nicole Garza, Rosy, Ariana Vázquez, Yazmín Gorrochotegui, Lorena Velázquez, Miguel Martínez, Gaby Moreno, Samuel Zapién, Ruth Torres, Ángeles Fang, Lucy Loredo, Mónica Briones, Jaime Sosa, Mariela Báez, Eli Hernández, Nora Céspedes, Miriam López.

A mis Maestros.

Por su paciencia, por su guía, por el acompañamiento y la comprensión en todo momento, gracias por ser parte de esta etapa en mi vida, por ayudarme a desarrollar habilidades que no sabía que era capaz de aprender. Gracias por enseñarme a retarme, a exigirme, a ser disciplinada.

Doctores Lucila Morales, Nancy Díaz, Rebeca Silva, Nohra Gallardo, Nelson Rangel, Luciano Vázquez, Ricardo Alamilla, Aarón Melo.

AGRADECIMIENTOS

Gracias a CONACyT por el apoyo a través de la beca No 660121, así como al Instituto Tecnológico de Ciudad Madero, mi alma mater.

Señor, hazme un instrumento de Tu Paz
Donde hay odio, que lleve yo el Amor.
Donde haya ofensa, que lleve yo el Perdón.
Donde haya discordia, que lleve yo la Unión.
Donde haya duda, que lleve yo la Fe.
Donde haya error, que lleve yo la Verdad.
Donde haya desesperación, que lleve yo la Alegría.
Donde haya tinieblas, que lleve yo la Luz.

Maestro, haced que yo no busque tanto ser consolada, sino consolar;
ser comprendida, sino comprender;
ser amada, como amar.

Porque es:
Dando, que se recibe;
Perdonando, que se es perdonado;
Muriendo, que se resucita a la
Vida Eterna.

Resumen

Durante este trabajo se estudió una familia de moléculas derivadas de oxialquilbenzoatos (OABs) con respecto a su comportamiento durante pruebas de gelificación con un conjunto de solventes polares y no polares. Se evaluaron una serie de propiedades estructurales y fisicoquímicas de los OABs y los solventes con el fin de desarrollar una serie de corpus a partir de esta información. Para la designación de los atributos, los corpus diseñados fueron divididos en dos tipos: cualitativos y fisicoquímicos; dependiendo del tipo de características usadas como atributos. Los corpus fueron aplicados en algoritmos de inteligencia artificial para diseñar un modelo que prediga el estado de agregación producido a partir de la combinación de un OAB y solventes específicos. Los productos de cada OAB con cada solvente se designan como las clases a predecir. El algoritmo de machine learning seleccionado para la evaluación de los corpus fue kNN (k neirest neighbours). Se probaron diferentes configuraciones tanto de los corpus como del algoritmo. Las configuraciones evaluadas comprenden la cantidad de atributos, su tipo y valor numérico o alfanumérico, así como la cantidad de ejemplos en los conjuntos y subconjuntos formados. Por último, con el fin de evaluar la capacidad predictiva de los modelos y su configuración óptima, se diseñaron una serie de conjuntos de prueba basados en los datos provenientes de dos moléculas nuevas. Como producto, se obtuvieron una serie de modelos con la capacidad de clasificar correctamente hasta en un 100% moléculas nuevas de la misma familia derivada de OABs. Algunos parámetros presentaron mayor relevancia sobre otros, estos parámetros son; el valor de k vecinos en el algoritmo kNN, la configuración del algoritmo y la estructura y contenido del corpus de información. La caracterización fisicoquímica de las especies estudiadas mediante los parámetros de solubilidad de Hansen, contribuyó mediante sus valores a que el modelo fuese capaz de realizar una asociación con las clases de los ejemplos de entrenamiento y de prueba, tal es así que, los atributos con los que se obtuvo más alta clasificación fueron las interacciones dispersivas, las interacciones polares y las interacciones de puentes de H., tanto de solventes y OABs. La estructura molecular caracterizada por rasgos como la cantidad de carbonos en las cadenas Éter y Éster de los OABs complementa la definición de la capacidad de gelificar de estas moléculas.

Abstract

Along this investigation, a family of molecules derived from oxyalkylbenzoates (OABs) was studied concerning their behavior during gelation tests with a set of polar and nonpolar solvents. A series of structural and physicochemical properties of the OABs and solvents were evaluated to develop a series of corpora from this information. The designed corpora were divided into qualitative and physicochemical, depending on the type of features used as attributes. The corpora were applied in artificial intelligence algorithms to design a model that predicts the aggregation state produced by combining an OAB with specific solvents. The products of each OAB with every solvent are designated as the classes to predict. The machine learning algorithm selected for evaluating the corpora was kNN (k nearest neighbors). Different configurations of both the corpora and the algorithm were tested. Configurations evaluated include the number of attributes, type, numeric or alphanumeric value, and the number of examples in the sets and subsets assigned. Finally, to assess the predictive skills of the models and optimal configuration, a series of test sets were designed based on data from two new molecules. The final product is a series of models with the ability to correctly classify up to 100% new molecules of the same family derived from OABs obtained. Some parameters presented higher relevance than others: the value of k neighbors in the kNN algorithm, the configuration of the algorithm, and the structure and content of the information corpus. Physicochemical characterization of the studied species using Hansen solubility parameters contributed with their values to the ability of the model to associate with the classes of the training and test examples. The attributes with the highest classification obtained were: dispersive interactions, polar interactions, and H-bond interactions, both for solvents and OABs. The molecular structure characterized by features such as the number of carbons in the Ether and Ester chains of the OABs complements the definition of the ability to gel these molecules.

Índice General

Resumen VIII
Abstract IX
Índice Tablas XIII
Índice de Figuras XVII
Nomenclatura XIX
1 Introducción 1
1.1 Planteamiento del problema. 2
1.2 Objetivos. 3
1.2.1 Objetivo general. 3
1.2.2 Objetivos específicos. 3
1.3 Justificación y beneficios. 4
1.4 Alcances y limitaciones. 4
1.5 Organización de la tesis. 5
2 Estado del Arte 7
2.1 Predicción computacional de gelificación molecular. 7
2.2 Diseño de un modelo de predicción para fenómenos fisicoquímicos. 10
2.2.1 Selección de atributos para modelos de clasificación. 10
2.2.2 Selección de conjuntos por tamaño y clase de datos para un modelo de clasificación. 14
2.3 Análisis de la influencia de las propiedades fisicoquímicas y estructurales de los componentes de un gel. 16
2.3.1 Parámetros de Solubilidad de Hansen como atributos fisicoquímicos en un modelo de clasificación de gelificación. 17
2.3.2 Longitud de cadenas alquílicas como propiedad estructural promotora de la gelificación. 20
3 Marco Teórico 24
3.1 Definición de gel. 24
3.2 Tipos de geles 26
3.2.1 Organogeles y organogeladores. 27

3.3 Solventes, el medio de gelificación de un organogelador. 28
3.4 Proceso de gelificación. 29
3.4.1 Interacciones intermoleculares. 30
3.5 Modelos de predicción de gelificación. 31
3.5.1 Parámetros moleculares con habilidades predictivas. 31
3.6 Ciencia de datos aplicada al pronóstico de gelificación. 35
3.6.1 Definición de ciencia de datos. 35
3.6.2 Diseño de un modelo predictivo de gelificación mediante machine learning. 40
4 Metodología 48
4.1 Recolección de datos para el diseño de un corpus predictivo. 48
4.2 Diseño de corpus de datos para modelos predictivos. 50
4.2.1 Atributos de un corpus de datos predictivo. 51
4.3 Evaluación de los corpus de datos. 52
4.3.1 Validación simple y validación cruzada. 52
4.3.2 Evaluación de Prueba. 53
4.4 Configuración y ajustes de los modelos de clasificación. 54
5 Desarrollo 60
5.1 Diseño de corpus con atributos cualitativos (categórico). 60
5.2 Diseño de Corpus de atributos fisicoquímicos. 61
5.3 Composición de subconjuntos de datos. 65
5.3.1 Subconjuntos a partir de corpus cualitativos. 65
5.3.2 Subconjuntos a partir de corpus fisicoquímicos. 67
5.4 Hipótesis. 78
6 Resultados y discusión de Validación con Corpus cualitativo. 80
6.1 Corpus cualitativo con atributos numérico ordinales. 80
6.1.1 Atributos del corpus y su impacto en la clasificación en kNN. 81
6.1.2 Evaluación de desempeño con base en identificación numérica de solventes. 84
7 Resultados y discusión de Validación y Prueba con Corpus fisicoquímico. 90

7.1 Validaciones cruzadas y simples.....90
7.1.1 Corpus A, validación cruzada.....91
7.1.2 Corpus A, validación simple.....93
7.1.3 Corpus A, validación cruzada con dobles ejemplos de entrenamiento.....97
7.1.4 Corpus B, validación cruzada.....99
7.1.5 Corpus C, validación simple.....100
7.2 Análisis de la composición de los conjuntos de entrenamiento y prueba de los corpus A y B correspondientes a las 3 iteraciones de la validación cruzada.....101
7.3 Relevancia de la exactitud de clasificación.....104
7.4 Prueba de Clasificación de las moléculas 1_{14} y 2_{14}.....105
7.4.1 Prueba con conjuntos de entrenamiento con clases No distribuidas. 105
7.4.2 Prueba con conjuntos de entrenamiento con clases distribuidas. 106
7.4.3 ANOVAS de conjuntos de entrenamiento.....111
8 Conclusiones y recomendaciones 118
8.1 Influencia de la clasificación a partir de la validación y prueba con corpus cualitativos. 118
8.2 Influencia de la clasificación a partir de la validación y prueba con corpus fisicoquímicos. 120
8.3 Influencia de hiperparámetros en un corpus de clasificación, y en la configuración del algoritmo. 121
8.4 Recomendaciones para el diseño de un modelo de clasificación.....123
Glosario.....125
Bibliografía 129

Índice Tablas

Tabla 1. Estado del arte sobre una variedad de moléculas y sus atributos estudiados para predecir su gelificación en diversos solventes mediante herramientas computacionales. 8
Tabla 2. Resultados de la predicción de una base de datos aplicada a una serie de algoritmos de IA. 13
Tabla 3. Factores que influyen sobre el desempeño de un modelo de clasificación de solubilidad molecular. 16
Tabla 4. Clasificación de solventes por polaridad y su estructura. 50
Tabla 5. Configuraciones del algoritmo kNN aplicadas en las evaluaciones de los corpus cualitativos y fisicoquímicos. 55
Tabla 6. Descripción del total de modelos predictivos desarrollados. 55
Tabla 7. Atributos de OABs y solventes 61
Tabla 8. Atributos correspondientes al corpus A. 63
Tabla 9. Contribución de componentes de los HSP por grupo estructural 64
Tabla 10. Valores de HSP para los nueve OABs de las familias octil, decil y dodecil éter. 65
Tabla 11. Distribución de número de ejemplos por clase en corpus A, B y C para cada familia éter. 71
Tabla 12. Distribución de número de ejemplos por clase con las 3 familias éter juntas, para los corpus A, B y C con concentración constante. 71
Tabla 13. Distribución de número de ejemplos por clase en corpus A, B y C por cada familia éter, con concentración como atributo. 72
Tabla 14. Distribución de número de ejemplos por clase con las 3 familias éter juntas, para los corpus A, B y B+, con concentración como atributo. 72
Tabla 15. Composición de los conjuntos con clases No distribuidas usados para la prueba de clasificación de las moléculas 1_{14} y 2_{14}. 76
Tabla 16. Composición de los conjuntos clases distribuidas usados para la prueba de clasificación de la molécula 2_{14}. 77

Tabla 17. Resultados experimentales de pruebas de gelificación en laboratorio con una serie de nuevos OABs derivados de oxialquilbenzoato. 77
Tabla 18. Atributos cualitativos de oxialquilbenzoatos (OABs) y solventes de tipo y su equivalente de clase numérico ordinal. 81
Tabla 19. Resultados expresados en porcentaje de clasificación obtenidos en la validación con 5 atributos, tipo alfanumérico vs numérico ordinal. 82
Tabla 20. Resultados de clasificación con el conjunto de validación expresados en porcentaje de acuerdo con valores de *k*, para modelos con variables alfanuméricas vs modelos con variables numéricas ordinales. 83
Tabla 21. Valores designados para identificación de solventes en experimentaciones con kNN. 85
Tabla 22. Resultados a partir de validación de acuerdo con los modelos con variación de valores de ID del atributo solvente. 85
Tabla 23. Solventes probados en experimentación y su identificación numérica, estructura y grupos funcionales. 86
Tabla 24. Relación de estados de agregación producto de la interacción experimental entre solventes (columna izquierda) y OABs (fila superior). 88
Tabla 25. %CA de clase con cruzada en kNN con corpus A. Peso: uniforme y distancia, métrica: Euclídea. 91
Tabla 26. %CA de clase con validación simple en kNN con corpus A. Peso: uniforme y distancia, métrica: Euclídea. 94
Tabla 27. Comparativo de %CA de clase, validación cruzada 30 ejemplos de entrenamiento únicas (azul) ejemplos dobles (rosa). Peso: distancia, métrica: Euclídea. 97
Tabla 28. %CA de clase con cruzada en kNN con corpus B. Peso: uniforme y distancia, métrica: Euclídea. 99
Tabla 29. %CA de clase con validación cruzada en kNN con corpus C. Peso: distancia, métrica: Euclídea. 100
Tabla 30. %CA de clase con validación simple en kNN con corpus C. Peso: distancia, métrica: Euclídea. 101

Tabla 31. Distribución de los ejemplos en los conjuntos de entrenamiento y prueba de la iteración 1. 102
Tabla 32. Distribución de los ejemplos en los conjuntos de entrenamiento y prueba de la iteración 2. 102
Tabla 33. Distribución de los ejemplos en los conjuntos de entrenamiento y prueba de la iteración 3. 102
Tabla 34. Resultados de clasificación para conjunto de Prueba con OAB 1_{14}, conjuntos de entrenamiento clases No distribuidas. Configuración de kNN: métrica Euclídea, k=5, peso uniforme. 105
Tabla 35. Resultados Prueba con OAB 2_{14}, Conjuntos de entrenamiento clases No distribuidas. Configuración de kNN: métrica Euclídea, k=5, peso uniforme. 106
Tabla 36. Resultados de clasificación para conjuntos de Prueba con OAB 1_{14}, Conjunto de entrenamiento con clases distribuidas. Configuración de kNN: métrica Euclídea, k=5, peso uniforme. 107
Tabla 37. Resultados Prueba con OAB 2_{14}, Conjuntos de entrenamiento con clases distribuidas. Configuración de kNN: métrica Euclídea, k=5, peso uniforme. 107
Tabla 38. Resultados de clasificación para conjuntos de Prueba con OAB 1_{14}, Conjunto de entrenamiento con clases distribuidas. Configuración de kNN: métrica Euclídea, k=2, peso uniforme. 108
Tabla 39. Resultados Prueba con OAB 2_{14}, Conjuntos de entrenamiento con clases distribuidas. Configuración de kNN: métrica Euclídea, k=2, peso uniforme. 109
Tabla 40. Resultados de %CA para las moléculas 1_{14} y 2_{14} bajo 3 distintas composiciones de conjuntos y 5 tipos de estructuras de corpus, a concentración constante. 109
Tabla 41. Resultados de %CA para las moléculas 1_{14} y 2_{14} bajo 3 distintas composiciones de conjuntos y 5 tipos de estructuras de corpus, a concentración variable. 110
Tabla 42. Resultados de %CA para las moléculas 1_{14} y 2_{14} bajo 3 distintas composiciones de conjuntos y 5 tipos de estructuras de corpus, a concentración variable. 111
Tabla 43. Resultados de análisis de varianza para cada estructura de corpus, con valores críticos F. 112

Tabla 44. Resultados de análisis de varianza para cada estructura de corpus, con valores críticos F. 112
Tabla 45. Resultados de análisis de varianza para cada estructura de corpus, con valores críticos F. 112
Tabla 46. Resultados de análisis de varianza para cada estructura de corpus, con valores críticos F. 113
Tabla 47. Atributos ordenados de acuerdo con su valor crítico *F* en cada configuración de conjuntos de entrenamiento 113
Tabla 48. Porcentaje de exactitud en la clasificación producida de acuerdo con cada uno de los 15 solventes experimentados.....115
Tabla 49. Porcentaje de exactitud en la clasificación producida de acuerdo con cada uno de los 15 solventes experimentados.....116
Tabla 50. Porcentaje de exactitud en la clasificación producida de acuerdo con cada uno de los 15 solventes experimentados.....117
Tabla 51. Comparativo de resultados más altos obtenidos con un corpus cualitativo vs un corpus fisicoquímico en kNN. 120

Índice de Figuras

Figura 1. Categorías de los factores influyentes en el desempeño de un modelo de clasificación (imagen fundamentada en el trabajo de Cihan y col. [8]).....15
Figura 2. Representación de moléculas peptídicas estudiadas por Haldar [14]. Se especifican los grupos funcionales aromático y péptido con las cadenas alifáticas “n” y “m” de longitud variada.....21
Figura 3. Clasificación de tipos de geles a partir de su origen y medio de formación. 26
Figura 4. Ciencia de datos su origen y aplicaciones [25]. 36
Figura 5. Diagrama de flujo de funcionamiento de un algoritmo de aprendizaje automático. 38
Figura 6. Validación cruzada con 4 iteraciones [34]. 45
Figura 7. Estructura de los derivados de OABs pertenecientes a las familias metil, propil y butil [2]. 49
Figura 8. Distribución de ejemplos en los conjuntos de entrenamiento y prueba para validación cruzada. 53
Figura 9. Distribución de atributos en las estructuras de corpus de tipos A, B y C. 62
Figura 10. Distribución de atributos en las estructuras de corpus de tipos A, B y C. 66
Figura 11. Configuración de conjuntos de entrenamiento y validación aplicadas durante la validación cruzada. 69
Figura 12. Composiciones de conjuntos formados a partir del corpus de atributos fisicoquímicos.....70
Figura 13. Composición de conjuntos de entrenamiento.....73
Figura 14. Composiciones de conjuntos de entrenamiento diseñados con la totalidad de los OABs, con la cantidad de ejemplos original, y con ejemplos distribuidos homogéneamente en cantidad de acuerdo con su clase. 74
Figura 15. Configuración de kNN que produjo el desempeño más alto del modelo de clasificación. 75

Figura 16. Composición base de los conjuntos de prueba formados con ejemplos de las moléculas nuevas 1_{14} y 2_{14}.....76
Figura 17. Comparación de % de Exactitud de clasificación para la validación cruzada de corpus A, distancia vs peso uniforme en la métrica Euclídea. 93
Figura 18. Comparativo del porcentaje de Exactitud de clasificación para la validación simple con corpus A, aplicando la distancia y el peso uniforme a los ejemplos en la métrica Euclídea. 95
Figura 19. Comparativo del % de Exactitud de clasificación para la validación simple vs validación cruzada con corpus A, usando la distancia entre los ejemplos en la métrica Euclídea. 96
Figura 20. Comparativo del porcentaje de Exactitud de clasificación para la validación simple vs validación cruzada con corpus A, usando la distancia entre los ejemplos en la métrica Euclídea. 98
Figura 21. Distribución de cantidad de ejemplos por clase (G, P y S). 103
Figura 22. Valores críticos F para cada atributo por cada composición de conjuntos de entrenamiento.....114
Figura 23. Configuración con mayor exactitud clasificatoria para kNN. 123
Figura 24. Configuración con mayor exactitud clasificatoria para un corpus de datos basado en moléculas tipo OAB. 124

Nomenclatura

HSP: Parámetros de solubilidad de Hansen

LMWG: Low molecular Weight gels

OAB: Oxialquilbenzoato

G: gel

S: solución

P: precipitado

HE: hexano

TO: tolueno

ADE: acetato de etilo

CCH: ciclohexano

THF: tetrahidrofurano

ACN: acetonitrilo

ACE: acetona

DMF: dimetil formamida

ET: etanol

DMS: dimetil sulfóxido

MET: metanol

HEX: hexano

PEN: pentano

CL: cloroformo

ISOP: isopropanol

Capítulo 1

1 Introducción

La interdisciplinariedad en el campo científico fusiona varias ciencias con el fin de cubrir las necesidades metodológicas emergentes y lograr objetivos comunes. Bajo esta premisa la química y la computación se unen para despejar las incógnitas experimentales encontradas en una diversa gama de fenómenos químicos, de los cuales el enfoque de esta investigación es la obtención de materiales conocidos como organogeles.

Las habilidades que los geles son capaces de desarrollar los hacen materiales de alta importancia estructural, dinámica y reológica. Este tipo de características les permiten ser utilizados en diversas aplicaciones en múltiples campos como el de la medicina como vehículos de sustancias activas en fármacos, en la industria de los polímeros aportando propiedades físicas y estructurales a plásticos, o en la remediación de suelos y mantos acuíferos para la captura y remoción de contaminantes ambientales.

En un sistema gel, las características del solvente son primordiales en el proceso de gelificación por la correlación existente con las propiedades del gelador. Para la selección del solvente apropiado se suelen someter a experimentación variedades de estos, considerando las propiedades que los hagan compatibles con la molécula candidata a gelificar.

Por otro lado, el diseño de una molécula que sea capaz de gelificar involucra la selección de rasgos como la polaridad, el tipo y longitud de sus cadenas alquílicas, la variedad de grupos

funcionales presentes, por mencionar algunos de los más determinantes. Debido a que la precisión de sus características determinará su competencia para gelificar, la etapa de síntesis suele ser exhaustiva de manera experimental, por lo que surge la necesidad de una herramienta de apoyo que permita predecir si esta será viable como gelador. Con el fin de desarrollar modelos predictivos, se propone el uso de la ciencia de datos a través de la analítica predictiva, que permite hacer uso de la inteligencia artificial para predecir comportamientos usando datos fisicoquímicos y estructurales de los componentes de un sistema de gelificación, mediante algoritmos de Machine Learning (aprendizaje máquina).

1.1 Planteamiento del problema.

El diseño de un gel representa un reto por su proceso de obtención *per se*, y a la fecha es un fenómeno que sigue siendo ampliamente estudiado por la variedad que se puede desarrollar de estos dependiendo de la aplicación deseada. Este caso de estudio está enfocado en el desarrollo de OABs como candidatos a organogeladores para su uso como agentes de remoción de combustibles. Investigaciones previas han demostrado la viabilidad de algunas familias de moléculas complejas de bajo peso capaces de formar entramados que puedan contener tanto solventes orgánicos como combustibles. Algunos rasgos promotores de la gelificación encontrados en este tipo de materiales con presencia de cadenas alquílicas solvofílicas flexibles, solvofóbico rígidos dependientes de un núcleo amido, amino o aromático [1].

Los OABs estudiados en este proyecto fueron desarrollados durante la investigación de Jaime Sosa y colaboradores [2]. Estos OABs son una serie de moléculas de bajo peso y complejidad estructural sencilla, planeadas a través de la modulación de diseños que sustentan la probabilidad de gelificación. Su estructura está basada en un núcleo aromático del cual dependen un grupo éter y uno éster con cadenas alquílicas de longitudes variables. Algunas de estas moléculas demostraron capacidad para formar geles, sin embargo, otras produjeron estados no deseados como soluciones y precipitados.

Existen factores determinantes intervienen en la formación o supresión de la gelificación, tales como el diseño de la estructura de los OABs, las pequeñas modificaciones en las longitudes de sus cadenas alquílicas, y la selección del solvente. También intervienen algunos estímulos externos como la temperatura, concentración, volumen de la fracción líquida, ciertos tipos de luz como la UV.

La cantidad de valores pertenecientes a las variables mencionadas para encontrar su conexión con la clase de producto obtenido requiere de un análisis intensivo. Un adecuado corpus de información aplicado en softwares de aprendizaje automatizado puede tener la capacidad de encontrar patrones en conjuntos de datos, como en este caso. Tomando en cuenta esto último, se propone el diseño de una serie de corpus que caractericen la producción de los distintos estados de agregación de los OABs con un conjunto de solventes seleccionados, aplicados en algoritmos de machine learning.

1.2 Objetivos.

En esta sección se presenta el objetivo general de este estudio, así como de los objetivos específicos que se realizarán para lograrlos.

1.2.1 Objetivo general.

Diseñar un modelo de predicción de gelificación de moléculas oxialquilbenzoato en sistemas monofásicos desarrollado a partir de algoritmos de machine learning.

1.2.2 Objetivos específicos.

1. Elaborar una base de datos a través de observaciones experimentales mediante pruebas de inversión de vial con moléculas oxialquilbenzoato en sistemas monofásicos con solventes polares y no polares.
2. Caracterizar estructural y fisicoquímicamente los solventes y las moléculas oxialquilbenzoato para la generación de un corpus que pueda utilizarse con técnicas de machine learning.

3. Realizar predicciones para nuevas moléculas utilizando los modelos de clasificación con mejores resultados.
4. Realizar predicciones para nuevas moléculas utilizando los modelos con mejores resultados.

1.3 Justificación y beneficios.

Un estudio a través de algoritmos de inteligencia artificial, con el fin de producir un modelo capaz de deducir las características del solvente y los cambios en la estructura de los OABs que orienten a la formación de un gel no es una tarea trivial, requiere comprender el proceso de gelificación y conocer la influencia de la gran diversidad de fuerzas atractivas y repulsivas que existe entre geladores y solventes, mismas que dotan de una naturaleza compleja a la tarea de obtener un gel, y son la base para la construcción de los cuerpos de datos que servirán para entrenar a los algoritmos. Las redes que se forman al ensamblarse las moléculas de un gelador en un solvente, están sometidas a los mínimos cambios en la estructura molecular de estas y a las variables externas involucradas como la concentración (del gelador), temperatura o volumen (de la fracción líquida), para lo cual atributos tales como la longitud de las cadenas alquílicas de los grupos éter y éster de los OABs, la presencia de las interacciones de tipo π-π del anillo aromático y el tipo de polaridad de los solventes requieren ser analizados para definir su influencia en el modelo de predicción, lo cual permitirá establecer la adecuada selección de atributos que aporten al modelo la información pertinente que, al ser procesada mediante el algoritmo de inteligencia artificial preciso, genere respuestas de clasificación mediante las cuales se pueda conocer las necesidades fisicoquímicas y estructurales de una molécula para ser capaz de gelificar un solvente dado.

1.4 Alcances y limitaciones.

La presente investigación está centrada en el estudio de una serie de moléculas pertenecientes a la familia de compuestos químicos oxialquilbenzoatos, y las características estructurales y fisicoquímicas que les aportan propiedades que las clasifican como geladores. La

información producida durante las pruebas experimentales en laboratorio y a través de los modelos de clasificación diseñados, se limita a la aplicación presente y futura de moléculas de estructura química con similitud a los grupos funcionales presentes en la familia estudiada.

En tanto, la estructura organizacional, los parámetros relevantes de un modelo de clasificación, y el diseño de la información que resulten de los más altos valores de predicción producidos, pueden ser empleados mediante la adecuada adaptación, para su uso con otra clase de moléculas pensadas para formar geles orgánicos o hidrogeles.

1.5 Organización de la tesis.

La composición de este texto comprende inicialmente el problema a resolver durante este proyecto, el objetivo general del mismo y los objetivos específicos para lograrlo, su justificación y beneficios, así como alcances y limitaciones, todos estos subtemas contenidos en el capítulo 1.

Durante el capítulo 2, se desarrolla el estudio del estado del arte para orientar a una adecuada selección de las características de solventes y OABs que posean aptitudes para clasificar correctamente el estado de agregación que se produce al combinar ambos componentes, y cuáles son las técnicas y configuraciones más convenientes para procesar la información.

Posteriormente, se incluye el capítulo 3 con el marco teórico donde se explican los conceptos más importantes utilizados durante el desarrollo del trabajo, como el tipo de propiedades estructurales y fisicoquímicas analizadas para formar la base de datos para los modelos de clasificación, los pasos para la creación de este, y las bases del funcionamiento y configuración de los algoritmos de IA.

El capítulo 4 contiene la metodología llevada a cabo para el diseño de los corpus de información, las propiedades de solventes y OABs estudiadas como atributos, y las diferentes configuraciones probadas durante el modelado.

La sección 5 plantea el desarrollo para producir los distintos modelos de clasificación; la estructura de los componentes: OABs y solventes; la composición de los diferentes corpus diseñados de acuerdo con su cantidad de ejemplos, la subdivisión en conjuntos de diferente tamaño, tipo y cantidad de atributos; y la evaluación y resultados de estos en los algoritmos propuestos bajo distintas configuraciones.

La discusión de los resultados para los diferentes corpus se plantea en los capítulos 6 para los corpus con atributos cualitativos, y 7 de los corpus con atributos fisicoquímicos respectivamente.

Por último, el capítulo 8 plantea las conclusiones y recomendaciones derivados de los resultados obtenidos.

Capítulo 2

2 Estado del Arte

Para explorar las partes vitales que componen esta investigación, se desarrolló una búsqueda bibliográfica acerca de las variables y las técnicas de estudio que usan la caracterización de los componentes de un sistema gel para diseñar modelos de clasificación. De este modo, durante esta sección se abordan trabajos previos que examinan la capacidad de ciertos atributos fisicoquímicos para relacionarse a través de herramientas, modelos y algoritmos computacionales, con la finalidad de predecir y conocer a profundidad el fenómeno de la gelificación.

2.1 Predicción computacional de gelificación molecular.

En la actualidad pese a que el campo de estudio que produce los geles es vasto, la existencia de modelos que aporten información *a priori* acerca de cuándo una molécula podrá formar un gel en un solvente específico es mínima y no universal, y la mayoría de los geladores aún se descubren empíricamente o por cambios estructurales cercanos a uno ya conocido [3].

Con base en el estudio de las interacciones entre solventes y geladores se sabe que el equilibrio entre estas es lo que orienta al autoensamblaje, por lo que, las ligeras modificaciones en el diseño de la molécula o la selección del solvente puede inhibir la gelificación [13]. Con el fin de seleccionar las características necesarias de solventes y geladores que promuevan la gelificación, se hace uso de herramientas computacionales para procesar propiedades fisicoquímicas y estructurales, y datos experimentales para predecir la capacidad de gelificación de moléculas específicas. En la Tabla 1 se muestra un concentrado

de trabajos de una variedad de investigadores, centrados en el diseño de modelos de clasificación con diversos tipos de propiedades fisicoquímicas y estructurales vía herramientas de computación. En ellos se observan resultados satisfactorios de predicción en la clasificación de moléculas con presencia de fuerzas fisicoquímicas similares a las contenidas en los oxialquilbenzoatos estudiados durante este trabajo. Además, para la creación de los modelos, los investigadores involucrados en el estado del arte que se presenta, coinciden en el uso de algoritmos de IA (Inteligencia Artificial) para el procesamiento de datos, así como en la caracterización de los componentes estudiados mediante los HSP y una serie de propiedades básicas y estructurales.

Tabla 1. Estado del arte sobre una variedad de moléculas y sus atributos estudiados para predecir su gelificación en diversos solventes mediante herramientas computacionales.

Autores	Tipo de molécula	Herramienta	Atributos	Metodología	Resultados
Bouteillier y col., 2011	Ocho moléculas de diferente estructura: 1. Cuatro de ellas con interacciones de H como las principales 2. Cuatro de ellas con solamente interacciones π y dipolares	Excel	Parámetros de Solubilidad de Hansen (HSP) de los solventes	Crearon esferas espaciales con los tres HSP como puntos cardinales, para clasificar los solventes de acuerdo con los productos formados con cada molécula.	Clasificación de solventes mediante sus valores de HSP como adecuados o no para gelificar una molécula de acuerdo con la estructura de esta
Diehn y col. 2013	Dibencil Sorbitol (DBS)	Matlab	Parámetros de Solubilidad de Hansen (HSP) de los solventes	Crearon esferas espaciales con los tres HSP como puntos cardinales, para clasificar los solventes de acuerdo con los productos formados con cada molécula.	Clasificación de solventes mediante sus valores de HSP como adecuados o no para gelificar una molécula de acuerdo con la estructura de esta.

Gupta y col., 2016	Péptidos	QSPR mediante machine learning Algoritmos: 1. SVM 2. RF 3. ANN	1. Propiedades básicas 2. Propiedades Estructurales 3. Propiedades fisicoquímicas	✓ Matriz con valores de propiedades, procesada con QSPR ✓ Se probó sólo agua como solvente	Exactitud en la clasificación o %CA 1. SVM: 70.8% 2. RF: 95.8% 3. ANN: 87.5%
Li y col., 2019	Dipéptidos	QSPR mediante machine learning Algoritmos: 1. Random Forest 2. Gradient Boosting Tree 3. Logistic Regression	1. Propiedades químicas 2. Propiedades Estructurales 3. Propiedades fisicoquímicas 4. Propiedades cuánticas	Diseñaron una matriz QSPR con propiedades de diversa naturaleza para 2000 dipéptidos de bibliografía, con fines de predecir su comportamiento fisicoquímico.	Precision Recall (PR) y Receiver operating characteristics curves (ROC) Entre 50 a 62%
Delcbecq y col., 2020	Amidoaminas	Leave one out y Cross validation con machine learning Algoritmos: 1. kNN 2. Naive Bayes 3. Rendom Forest 4. ANN 5. SVM 6. Decision Tree 7. Logistic Regression	Hansen Solubility Parameters (HSP) sólo de los solventes	Se estudió una molécula a la vez, combinada con cada solvente, madiante cada uno de los algoritmos Leave one out deja sólo un ejemplo de prueba y un conjunto de 26 solventes como ejemplos de entrenamiento.	Exactitud en la clasificación o %CA De 18 a 87%
Cihan y col., 2021	Compuestos extraídos de una base de datos	QSPR mediante machine learning Algoritmos: 1. ANN 2. Random Forest 3. XGB	1. Propiedades básicas 2. Propiedades Estructurales 3. Propiedades fisicoquímicas	Leave one out y cross validation para evaluar un QSPR diseñado con las propiedades de una serie de compuestos, para conocer su potencial de solubilidad en agua.	Exactitud en la clasificación o %CA 94%

Durante el siguiente apartado del capítulo, se plantean algunas de las propiedades fisicoquímicas y estructurales mencionadas en la tabla superior, tales como los HSP, y algunos rasgos de estructura molecular como la longitud de los tallos alifáticos, que han sido estudiadas en una diversidad amplia de trabajos encontrados en el estado del arte, con buenos resultados al momento de obtener información para predecir el comportamiento de agregación de moléculas con estructuras variadas. De manera similar, se mencionan algunos autores y sus correspondientes descubrimientos acerca del diseño de modelos predictivos basados en datos de caracterización con propiedades básicas, estructurales y fisicoquímicas de moléculas de distintas especies.

2.2 Diseño de un modelo de predicción para fenómenos fisicoquímicos.

A partir de los datos adecuados que caractericen a los componentes de un gel, es posible diseñar bases de datos para luego diseminarlas en subconjuntos que puedan ser aplicados en una herramienta para procesar la información de tal forma que se obtenga una respuesta informativa sobre los productos que pueden ser obtenidos tras la interacción de una molécula y un solvente. El desarrollo de esta clase de metodologías se aborda en los siguientes subtemas.

2.2.1 Selección de atributos para modelos de clasificación.

En esta sección se incluye la búsqueda de investigaciones que exploran los mecanismos de formación de geles a través de la caracterización de solventes y moléculas con propiedades de distinta índole.

Existe una enorme variedad de sistemas de organogeles, sin embargo, los estudios acerca de la correlación estructura-solvente que dan pie a la gelificación aún son limitados y con enfoques desarrollados dentro de moléculas de un mismo tipo de familias. Danielle Zurcher y Anne McNeil [4] dedican una de sus colaboraciones a la examinación de las interacciones intermoleculares gelador / solvente con el fin de conocer la importancia de ambos componentes. En una de sus principales conclusiones refieren que para que un gel se forme debe existir un equilibrio en la solubilidad del gelador, no presentar extremos entre la

insolubilidad y la solubilidad. Sobre esta conclusión, encontraron que el solvente puede determinar cambios sustanciales tanto en la entalpía de disolución como en la capacidad de gelificación, lo cual modifica el mencionado equilibrio. Bajo este contexto y tomando en cuenta la importancia del rol del solvente durante la gelificación, se plantea la caracterización de este componente bajo diferentes propiedades básicas y fisicoquímicas, que puedan ser ligadas a características estructurales de las moléculas con las que interactúan.

A partir de este preámbulo y con la finalidad de predecir la estructura necesaria de una molécula dada para que sea capaz de gelificar, Gupta y col. [5] desarrollaron un modelo de clasificación, utilizando una serie de características fisicoquímicas, básicas y estructurales de diversas moléculas sometidos a pruebas de gelificación en agua. Estas propiedades fueron puestas a prueba en modelos computacionales como parte de la caracterización de un conjunto de 34 compuestos como entrenamiento, 9 compuestos para validar y 21 como prueba.

En el desarrollo de sus modelos, Gupta y col. optaron por incluir grupos funcionales específicos a las moléculas con el fin de promover su autoensamblaje unidimensional en un medio acuático utilizado como solvente. Una de las partes trascendentales de este estudio analizado es el uso de descriptores moleculares desarrollados en algoritmos de Machine learning para analizar las especies químicas. Estos descriptores son conocidos como *Relaciones Cuantitativas Estructura-Propiedades, o QSPR por sus siglas en inglés (Quantitative structure–property relationships),* y basan su funcionamiento en la obtención de un código numérico mediante matrices formadas con los valores de las propiedades básicas, fisicoquímicas y estructurales, en este caso de una serie de moléculas, procesados a través de métodos matemáticos.

Los QSPR se basan en el principio de que *"los puntos finales medibles experimentalmente son una función de las propiedades moleculares"* [6]. Estos modelos son aptos para vincular las propiedades contenidas en los descriptores con la capacidad de gelificación de un compuesto, en este caso específico. Para procesar los descriptores es necesario el uso de un

algoritmo capaz de procesar la información. En la investigación estudiada, se utilizaron una variedad de algoritmos de machine learning: un modelado bayesiano, random forest, máquinas de vectores de soporte, vecinos más cercanos y redes neuronales fueron empleados para vincular los descriptores con el punto final medido, que fue el estado del producto experimental.

La evaluación de estos modelos desarrollados en algoritmos de IA puede desarrollarse de diferentes formas, dependiendo de la necesidad de análisis que se requiere de los datos. Para el caso estudiado el tipo de evaluación inicial que seleccionaron los investigadores fue la de *Validación Cruzada*, método por el cual se someten a prueba conjuntos de ejemplos contenidos dentro de los utilizados para entrenar los algoritmos, con el fin de evaluar la capacidad de clasificación de los datos, y la selección de los hiperparámetros del algoritmo con más alto desempeño. Posterior a los ajustes hechos con la información proveniente de la validación, aplicaron al modelo un conjunto de prueba, creado con un grupo de moléculas que el modelo no conocía. Es importante tomar en cuenta que "las moléculas nuevas que sean evaluadas en cada modelo de clasificación tengan la misma base estructural que las que fueron usadas para entrenarlos, esto se conoce como *Dominio de aplicabilidad"* [7]. El dominio de aplicabilidad se refiere a los ejemplos en las que el modelo predictivo puede ser usado con confianza, al haber sido comprobada su habilidad obteniendo respuestas ciertas.

De acuerdo con la investigación de Gupta y col. [5], su estudio estuvo basado en la evaluación del desempeño de cada uno de sus modelos creados en los diferentes algoritmos, y bajo la interpretación de 14 geladores que no pertenecieron a los conjuntos de entrenamiento y se utilizaron como prueba, tres de los algoritmos evaluados probaron obtener buenos resultados clasificatorios. En la Tabla 2 se muestran los resultados que obtuvieron.

Tabla 2. Resultados de la predicción de una base de datos aplicada a una serie de algoritmos de IA.

Método	Equilibrio de exactitud	Calidad de predicciones
SVM	0.798	BUENO
RF	0.958	BUENO
kNN	0.7941	DEFICIENTE
NN	0.875	BUENO
PLS	0.625	DEFICIENTE
NB	0.791	DEFICIENTE
C5.0	0.583	DEFICIENTE

En la bibliografía existen una diversidad de estudios con el uso de modelos QSPR para clasificación de moléculas de diversa índole. Dentro de los que se consultaron para este trabajo de investigación, se mencionarán durante esta sección los llevados a cabo por Li y col. [7] y Cihan y col. [8].

El trabajo desarrollado por Li y col. [7], trata una variedad amplia de 2000 moléculas obtenidas de una base de datos de tipo dipéptido como candidatas a formar hidrogeles. A partir de las propiedades básicas, estructurales, cuánticas y fisicoquímicas de estas moléculas desarrollaron modelos QSPR aplicando machine learning para conocer su comportamiento para autoensamblarse en agua. Una de las aportaciones más significativas del trabajo de Li y col. fue la relación encontrada entre la estructura molecular del gelador y sus propiedades para gelificar. Debido a la amplia gama de moléculas que caracterizaron con propiedades de distinta naturaleza, les fue posible ligar el tipo de característica que funciona mejor de acuerdo con los rasgos estructurales de los geladores. En cuanto a los algoritmos que probaron, estos fueron desde clasificadores lineales, regresión logística hasta redes neuronales, encontraron que los más altos rendimientos se presentaron usando random forest, gradient boost y regresión logística. Debido a que la base de datos que utilizaron sólo contenía un 4% de moléculas capaces de gelificar, la evaluación de los modelos la basaron en la precisión de los datos, produciéndose así valores de 54% (random forest), 57% (regresión logística) y 62% (gradient boost).

2.2.2 Selección de conjuntos por tamaño y clase de datos para un modelo de clasificación.

En busca de la selección adecuada de la información para diseñar un modelo capaz de predecir el potencial de solubilidad de una serie de compuestos de una base de datos, Cihan y col. [8] basaron su investigación en la validación de la calidad y cantidad de información adecuada en un modelo óptimo de predicción de solubilidad. Diseñaron un modelo con una vasta cantidad de propiedades de distinta naturaleza y a partir de él seleccionaron la cantidad y tipo de datos con mayor rendimiento, mediante pruebas de validación estadística procesada en algoritmos de machine learning. La información utilizada por los investigadores fue tomada de una amplia base de datos de solubilidad molecular disponible en literatura.

De acuerdo con el estado del arte relativo al estudio de Cihan, los factores con mayor relevancia sobre el desempeño de un modelo de predicción se pueden agrupar en cuatro categorías; con referencia a los datos se mencionan el tamaño de los conjuntos y su calidad; y sobre el modelo, la relevancia de los descriptores químicos y la capacidad del algoritmo [9]. La Figura 1 plantea las categorías y variables subyacentes con mayor afectación sobre el desempeño de un modelo de clasificación basado en descriptores químicos.

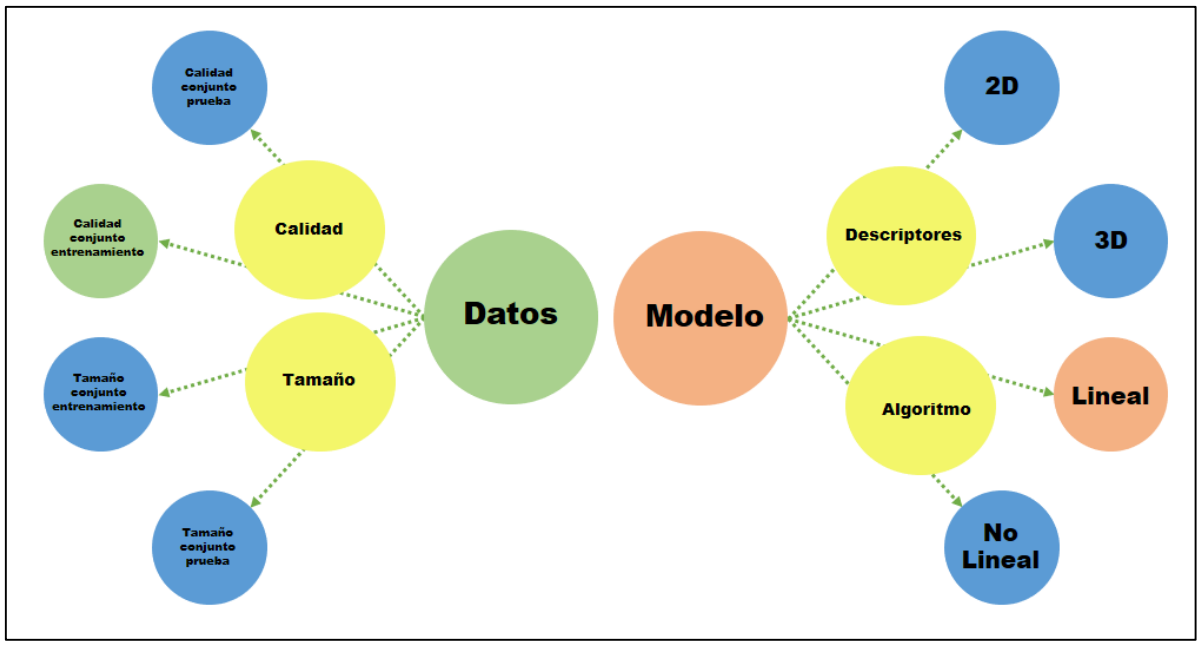


Figura 1. Categorías de los factores influyentes en el desempeño de un modelo de clasificación (imagen fundamentada en el trabajo de Cihan y col. [8]).

Para analizar la influencia de la calidad y cantidad de los datos sobre el modelo de predicción, Cihan y el equipo de investigadores desarrollaron diferentes composiciones de conjuntos de subdatos de acuerdo con los atributos usados, pero con la misma cantidad de instancias y evaluaron mediante la técnica de Validación Cruzada. Tras esta primera evaluación, comprobaron la influencia de la robustez variando la cantidad de instancias y utilizando la composición de atributos con mayor rendimiento en la validación cruzada. Concluyeron que el tamaño y la calidad de los conjuntos de entrenamiento se relacionan con la exactitud del modelo de manera positiva. La correcta selección de los datos, minimiza su robustez, pero aumenta la calidad del conjunto. Observaron que la precisión de la clasificación disminuyó cuando la cantidad de datos era menor en los conjuntos de entrenamiento. En tanto, los atributos de mayor calidad contribuyen a elevar el desempeño cuando los conjuntos de datos contienen una mayor cantidad de instancias. De acuerdo con las observaciones obtenidas por el mencionado grupo de investigadores, la Tabla 3 contiene una serie de factores y su influencia sobre el desempeño de un modelo de clasificación.

Tabla 3. Factores que influyen sobre el desempeño de un modelo de clasificación de solubilidad molecular [9].

Factor	Relevancia sobre el desempeño del modelo de clasificación
Tamaño de los datos	El incremento en la cantidad de los datos tiene un efecto positivo en la exactitud de los modelos. El tamaño de los conjuntos de entrenamiento y de prueba tienen diferentes impactos en los resultados. El tamaño del conjunto de entrenamiento afecta la exactitud del modelo, y el del conjunto de prueba interfiere en la evaluación de este.
Calidad de los datos	La exactitud en la medición de los valores de los datos que forman los conjuntos de entrenamiento y prueba para un modelo influye directamente en la medida del error de clasificación.
Descriptores Químicos	Proveen una representación matemática de la información química de un compuesto, que es usada como entrada en modelos predictivos. De manera general, existen dos tipos: 2D los cuáles son calculados con absoluta precisión, y los 3D que contienen errores metodológicos que se deben ajustar. Los descriptores 3D contienen información susceptible debido al tipo de información que procesan, como distancias atómicas y energía molecular.
Capacidad del algoritmo	De acuerdo con la naturaleza del procesamiento del algoritmo, su capacidad está ligada fuertemente con la clase, calidad y tamaño de los datos que deberá trabajar.

El siguiente apartado, contiene un análisis acerca del uso de parámetros fisicoquímicos moleculares usados como atributos para diseñar modelos de clasificación.

2.3 Análisis de la influencia de las propiedades fisicoquímicas y estructurales de los componentes de un gel.

En el estudio que comprende el desarrollo de un modelo de datos capaz de brindar información *a priori* acerca de los productos que se tendrán a partir de la interacción de una molécula y un solvente, se conocen dos grandes grupos de características, que a través del estudio del estado del arte han demostrado un alto impacto caracterizando a una diversidad de componentes de sistemas gel, con el fin de diseñar modelos de clasificación. Estos grupos

comprenden las propiedades estructurales de las moléculas, y en específico los llamados Parámetros de Solubilidad de Hansen encargados de evaluar el comportamiento fisicoquímico de dos entidades al interactuar en un medio.

Los siguientes apartados abordan estas propiedades y su capacidad como atributos de predicción de gelificación.

2.3.1 Parámetros de Solubilidad de Hansen como atributos fisicoquímicos en un modelo de clasificación de gelificación.

En la literatura que toca el tema de la clasificación o predicción de moléculas y solventes de acuerdo con sus habilidades gelificadoras, se pueden encontrar los primeros esfuerzos basados en propiedades fisicoquímicas tales que definan el comportamiento molecular de una manera clara y distintiva. Existe un conjunto de características que ha sido estudiado muy ampliamente por presentar capacidad de clasificación al relacionarlo con productos como geles y polimerización en solventes, estos son los parámetros de solubilidad de Hansen (HSP).

Para poder caracterizar una molécula a partir de sus HSP, esta requiere tener características estructurales sencillas, pues el cálculo de estos parámetros se torna poco preciso para moléculas complejas.

Dos de los estudios citados durante el presente estado del arte, tienen como base de datos fisicoquímicos los HSP de una serie de solventes. Los trabajos desarrollados por Bouteillier [10] y Diehn [11] respectivamente. En sus respectivas investigaciones, los HSP son usados como atributos para la clasificación a partir del diseño de espacios esféricos mediante el uso de cada uno de los tres parámetros como puntos cardinales, donde posteriormente son situadas las moléculas candidatas a gelificar, de acuerdo con los productos formados con cada solvente. La variante entre ambas investigaciones fue la herramienta utilizada para el procesamiento de datos y la creación de los espacios.

Para detallar la forma en que cada investigador aplicó estos parámetros, se describe a continuación el enfoque de los modelos de clasificación. En su trabajo, Bouteiller y col. [10], se enfocaron en explicar los estados de agregación producidos a partir de pruebas de gelificación de un conjunto de solventes polares y no polares con varias moléculas con distintos tipos de interacciones físicas en su estructura. Para esta labor, caracterizaron los solventes mediante sus HSP, y fueron estos parámetros lo utilizados para definir las aptitudes de cada solvente para formar geles, relacionándolos con las interacciones presentes en los geladores de acuerdo con su estructura.

Bajo este mismo enfoque, Diehn y col. [11] estudiaron el equilibrio de fuerzas que induce a la gelificación de la molécula 1,3: 2,4-dibenciliden (dibenzylidene) sorbitol (DBS) mediante la clasificación de una serie de solventes y los productos que se originan de su interacción con esta molécula, utilizando los HSP para caracterizarlos. En sus resultados califican a cada solvente por su capacidad de interactuar con el DBS mediante sus fuerzas de dispersión, dipolo – dipolo y enlaces de H. En ambas investigaciones, se comprobó la competencia de los HSP como atributos usados para discernir la gelificación de moléculas de índole variada.

El impacto de pequeñas modificaciones a moléculas con el fin de aumentar su capacidad de gelificación, ha sido estudiado por una variedad de investigadores, tal es el caso de estudio de Delbecq y col. [12], quienes investigaron el impacto de las variaciones de la longitud de las cadenas alifáticas de moléculas tipo amido amina, con relativa buena capacidad para gelificar en solventes como tolueno y agua. Una de las moléculas que modificaron, tuvo un reemplazo para elongar una de sus cadenas alifáticas de 18 carbonos por una más larga de 22.

Parte del trabajo de estos investigadores incluyó la creación de un modelo predictivo a partir de los HSP de los solventes en los que fueron probadas las moléculas. Tomaron en cuenta que, estas propiedades han sido usadas ampliamente para clasificar como un compuesto dado reacciona al ser probado en solución con un solvente específico. Probaron con una serie de

algoritmos de machine learning, para evaluar las habilidades predictivas y la utilidad interpretativa de la combinación de los HSP con cada algoritmo clasificando las moléculas.

Los algoritmos que probaron fueron; kNN (nearest neighbors – vecinos más cercanos), Naïve Bayes, Random Forest, Regresión logística, ANN (artificial neuronal networks - redes neuronales), SVM (support vector machines – máquinas de vectores de soporte) y Árbol de decisión. Su forma de evaluar los modelos fue a partir de leave-one out, con dos opciones de salida; Gel / no Gel. Durante esta clase de evaluación, el algoritmo es entrenado con la totalidad de las instancias o ejemplos, menos uno, el cual es dejado para probar la capacidad de predicción del modelo. Así, se evalúa una por una las instancias, mientras que el resto es usado como entrenamiento.

En los resultados, describen como la molécula cuya cadena alifática fue elongada, presenta un aumento en su comportamiento hidrofóbico. En esta molécula se obtuvieron resultados favorables para gelificar en solventes inusuales como en Acetonitrilo y la DMF, además de aparentemente ser una buena candidata a gelificar solventes con regiones más polares.

En cuanto a los algoritmos usados para la clasificación, estos funcionaron de manera similar, pero uno se destacó en particular; las máquinas de vectores de soporte (SVM), con el que el modelo probado proporcionó visualizaciones útiles para ayudar a interpretar los resultados de gelificación molecular.

Como se mostró en los diferentes trabajos analizados durante esta sección, los HSP aportan información útil al momento de determinar si una molécula gelificará un solvente dado o no. Además, en el último trabajo examinado, se observa que la estructura de las moléculas tiene una relación intrínseca con la funcionalización de un gelador. Más específicamente la elongación de las cadenas alifáticas presentes en estas, modula las habilidades de autoensamblaje de una molécula al aumentar su hidrofobicidad. Este rasgo es tratado con mayor detalle en la siguiente sección.

2.3.2 Longitud de cadenas alquílicas como propiedad estructural promotora de la gelificación.

Derivado de una serie de estudios ([12], [13], [14]), se sabe que en parte la capacidad de gelificación de una molécula depende del equilibrio hidrofóbico-hidrofílico de la misma, en este equilibrio interfieren las unidades polares y no polares presentes en la molécula. Las interacciones dadas por las fuerzas de van deer Waals presentes en los grupos alquílicos y que varían en función de la longitud de las cadenas de estos grupos, tienen relación con el comportamiento dispersivo de un componente, que a su vez se vincula con la capacidad de gelificación molecular.

Cuando se relaciona la capacidad de gelificación de una especie con la longitud de sus cadenas alquílicas, se debe tomar en cuenta como un rasgo particularmente relevante los grupos funcionales presentes en la molécula. Grupos funcionales como las amidas, amino amidas, péptidos o dipéptidos ligados a cadenas alifáticas de longitudes adecuadas, son capaces de otorgar a un espécimen propiedades para gelificar en solventes específicos. Tal es el caso estudiado por Haldar y col. [14], en su investigación, modificaron dipéptidos a través de la variación de las cadenas alifáticas presentes en las moléculas, para elucidar el efecto del ajuste hidrofóbico en su estructura, sobre el autoensamblaje en una serie de solventes polares y no polares.

Los dipéptidos estudiados por Haldar, comprendían tres conjuntos de moléculas con la fórmula común $C_mH_{2m+1}C(=O)NH(_L)Val(C=O)NH–(CH_2)_n–(C=O)OB_n$, donde “m” y “n” representan la cantidad de carbonos en las cadenas alifáticas pertenecientes a la cabeza aromática, y a la unidad peptídica respectivamente (Figura 2).

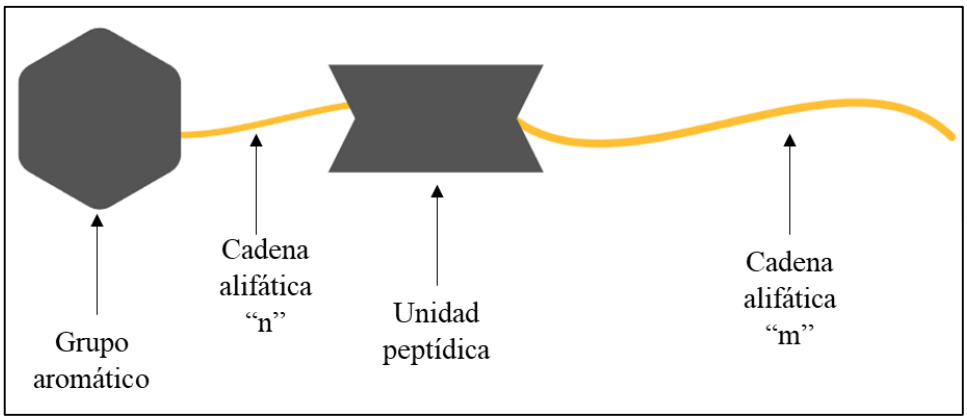


Figura 2. Representación de moléculas peptídicas estudiadas por Haldar [14]. Se especifican los grupos funcionales aromático y péptido con las cadenas alifáticas "n" y "m" de longitud variada.

Los conjuntos de moléculas estudiados en la investigación de Haldar, incluyen tres diferentes sets moleculares con distintas combinaciones de acuerdo con la longitud de las cadenas alifáticas "n" y "m". El set 1 incluía moléculas con n = 2, m = 9, 11, 13, 15, 17, para el set-II con n = 2, 3, 5, m = 13 y el set-III consistía de dos geladores isoméricos (n=2, m=15; n=10, m=7). La base hipotética por la cual seleccionaron las longitudes de las cadenas lineales se fundamentó en la aportación de las fuerzas de van deer Waals presentes en este tipo de grupos, y como la variación en el tamaño de sus cadenas puede modular su capacidad de gelificar en un solvente dado.

En sus resultados, el grupo de investigadores reportan haber encontrado dos tipos de comportamiento durante la gelificación en varios solventes polares y no polares. El efecto de la modulación de las cadenas hidrofóbicas produjo una serie de comportamientos particulares de acuerdo con el solvente usado como medio. El autoensamblaje es menos favorable en solventes como el n-hexano (solvente lineal) que en el Ciclohexano cuando aumenta la longitud de la cadena "m". La conclusión sobre este suceso es que, un solvente de cadena flexible y extendida (n-hexano) en comparación con uno de cadena rígida cíclica (Ciclohexano) es capaz de actuar favorablemente con la cadena n-acilo "m" debido a la interacción de las fuerzas de van deer Waals en ambas entidades.

Otra de las aportaciones de la citada investigación es la orientación dada debido a las diferentes longitudes probadas de las cadenas “n” y “m” por la interacción que permiten entre los grupos aromático y peptídico entre moléculas al momento del autoensamblaje. “El aumento de longitud en la unidad “m” incrementa la interacción molécula-solvente en medios no polares; sin embargo, en medios polares provoca un efecto contraproducente para gelificar.”

La estructura del solvente en el que se espera el autoensamblaje tiene una gran influencia con respecto a las partes flexibles de una molécula, en el caso, las cadenas alifáticas. Una modulación adecuada en la longitud de las cadenas alifáticas de un gelador puede determinar su gelificación en solventes con estructuras que contengan fuerzas de van deer Waals capaces de lograr un equilibrio solvofílico/solvofóbico.

La incorporación de grupos alifáticos en una molécula para funcionalizar su comportamiento como gelador ha sido reportada para distintas familias moleculares. Tal es el caso del estudio desarrollado por Iqbal y col. [13], en el cual modificaron tetrapéptidos mediante la modulación de las cadenas alifáticas en las dos terminaciones de las moléculas estudiadas. Estudiaron el autoensamblaje de la serie de tetrapéptidos diseñados, en una serie de solventes orgánicos, basados en la aportación de las fuerzas de van deer Waals contenidas en los grupos lineales y la posibilidad de ajustar la capacidad de gelificar de una molécula variando la longitud de sus cadenas alifáticas.

La investigación citada proporciona información útil acerca de la relevancia que tienen las cadenas alifáticas polares durante el autoensamblaje. La longitud de la fracción polar tiene afectación sobre la capacidad de gelificación de las moléculas estudiadas, asociándose este con las interacciones intermoleculares de van deer Waals presentes en estos grupos. “Se ha demostrado que la presencia de cadenas de alquilo similares en ambos extremos de la molécula mejora las propiedades de gelificación”.

A partir de la revisión expuesta durante este capítulo, se observan aspectos importantes que aparentemente son de utilidad al momento de desarrollar modelos de clasificación. La adecuada selección de las características de solventes y moléculas para usarse como atributos en una base de datos, es fundamental y está asociada con el algoritmo seleccionado para procesar la información. El uso de algoritmos de machine learning para probar y decidir los parámetros de un modelo que producen clasificaciones correctas más altas implica evaluaciones con conjuntos de datos de entrenamiento y prueba, y dependerá de la naturaleza de los datos y el tipo de atributos el que un algoritmo presente mejores resultados que otro. Los HSP de los solventes usados en las pruebas de gelificación aplicadas a una molécula, son características con alta capacidad para aportar información al momento de predecir si una molécula se convertirá en gelador en un solvente específico o no. A la par con los rasgos fisicoquímicos aportados por los HSP, la longitud de las cadenas alifáticas presentes en un candidato a gelador, orientan el comportamiento de autoensamblaje dependiendo de los grupos funcionales presentes, al modular la polaridad de la molécula, siendo así, una propiedad efectiva al usarse como atributo.

En el siguiente capítulo, se abordarán los conceptos básicos expuestos durante el estado del arte y otros necesarios para comprender y desarrollar un modelo de predicción de gelificación funcional.

3 Marco Teórico

En este apartado se explicarán temas base acerca de los conceptos fundamentales que se involucran en el desarrollo de esta investigación. Que es un gel y los tipos y usos que se encuentran en estos materiales, que son los organogeladores y que fuerzas se entraman con los solventes para dar paso al producto final. El machine learning alternativa adecuada para diseñar un modelo predictivo para la gelificación molecular, así como las diversas opciones y características que sus algoritmos ofrecen.

3.1 Definición de gel.

De las definiciones más convencionales acerca de lo que es un gel, se puede encontrar en la búsqueda bibliográfica, que son sustancias que poseen una estructura continua con dimensiones macroscópicas, permanente en la escala de tiempo en un experimento, y como un sólido en lo que respecta a su comportamiento reológico [15]. En una descripción más simple puede decirse que un gel es un sistema coloidal de aspecto sólido que fluye al ser sometido a esfuerzos relativamente débiles. Los geles poseen las propiedades mecánicas de un sólido como mantener su forma bajo una fuerza aplicada correspondiente a su propio peso,

y bajo una fuerza mecánica, esto demuestra el llamado fenómeno de tensión. Son sistemas coloidales dispersos de al menos dos componentes, donde tanto el componente disperso como el medio de dispersión se extienden a ellos mismos continuamente a través del sistema completo.

Existen ciertas características que contribuyen a poder hacer una clasificación más simple de una sustancia para saber si es o no un gel: (1) si posee una estructura microscópica continua con dimensiones macroscópicas siendo permanente en una escala de tiempo de un experimento analítico y (2) es similar a un sólido en su comportamiento reológico pese a tener mayormente una apariencia de líquido.

Los geles pueden obtenerse a partir de compuestos inorgánicos, proteínas o polímeros [16], y otra clase de geles particulares son los formados a partir de compuestos orgánicos de bajo peso molecular, los cuales reciben el nombre de organogeladores o por sus siglas en inglés LMOGs (low molecular mass organic gelators.

Los geles se forman mediante el autoensamblaje de pequeñas moléculas en estructuras supramoleculares que inmovilizan el solvente a través de fuerzas capilares y de tensión superficial. Esta auto agregación es conducida por interacciones intermoleculares no covalentes. Las interacciones físicas en estos agregados dotan a los geles de las propiedades macroscópicas que los caracterizan. Experimentalmente los geles son obtenidos probando la molécula candidata a gelador con varios solventes, y observando el tipo de agregado que se forma a partir de estas combinaciones.

Las ventajas derivadas de los tipos de interacciones gelador / solvente permiten la formación de geles con estructuras como: fibras sólidas, cristales, varillas, láminas, listones, fibras helicoidales, fibras en forma de estrella, y a partir de estas especies y sus características ser aplicados en usos diversos.

3.2 Tipos de geles.

La clasificación de los geles moleculares se da en torno al solvente en el cual se produce la gelificación. Es así que, tomando en cuenta su naturaleza de origen y el medio en que se conciben, los geles pueden ser categorizados en estos dos grupos que a su vez se subdividen. La Figura 3 contiene un diagrama con la clasificación básica de los geles por su origen y medio de formación.

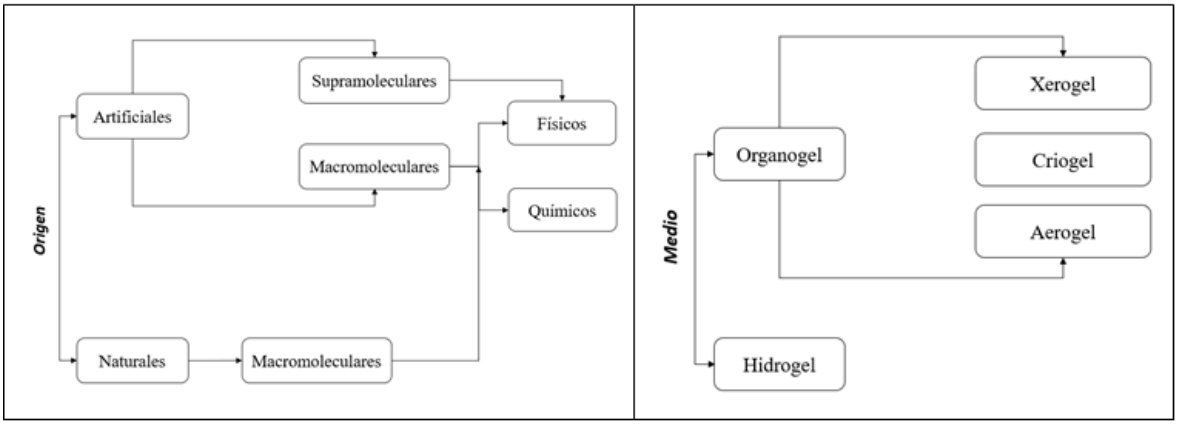


Figura 3. Clasificación de tipos de geles a partir de su origen y medio de formación.

Los geles derivados de compuestos sintéticos por su constitución pueden ser macromoleculares (polímero) o supramoleculares. La formación de geles a partir de compuestos macromoleculares puede ser resultado de reticulación química o interacciones físicas. Esta clase de geles están formados por enlaces fuertes, no se pueden volver a disolver y son térmicamente irreversibles. En tanto los geles formados por interacciones no covalentes débiles (entramados físicos) son reversibles y son conocidos como geles supramoleculares, derivados de compuestos de bajo peso molecular. Su formación sucede por medio de una auto agregación de pequeñas moléculas gelantes que entraman redes fibrilares auto ensambladas (o SAFIN por sus siglas en inglés). Mediante investigaciones recientes, como algunas de las mencionadas durante el estado del arte, se sabe que la obtención de un gelador y el diseño de estos materiales sucede a través de la funcionalización de moléculas base incorporando grupos funcionales en su estructura.

En cuanto a la clasificación derivada del medio en que gelifica un gelador, estas moléculas se dividen en dos tipos que son los siguientes: los hidrogeles; moléculas que poseen la capacidad de retener en sus entramados altas cantidades de agua mediante el hinchamiento de su estructura, sin perder su forma original por medio de la auto asociación, y los organogeles; que son formados mediante la gelificación de moléculas de bajo peso en solventes orgánicos.

3.2.1 Organogeles y organogeladores.

Un organogel será aquél formado por moléculas de bajo peso molecular conocidas como organogeladores que se auto ensamblan en un medio que es un solvente orgánico, dando como resultado una estructura supramolecular que otorga una respuesta a un estímulo químico o físico como el pH, calor, luz, campo magnético o ultrasonido.

Los organogeles poseen características provenientes de sus compuestos de origen, mismas que los diferencian de otros geles. Su compuesto de origen es un organogelador, molécula de bajo peso molecular de alrededor de 300 g/mol, capaces de formar una red tridimensional continua, misma que en los organogeles se presenta debido a enlaces no covalentes (en otra clase de geles estas redes se dan por enlaces covalentes). En los organogeles, el proceso de gelificación ocurre con concentraciones de gelador desde los 0.1 al 1% o más [17].

Comúnmente los organogeladores incluyen moléculas derivadas de ácidos grasos, derivados esteroideos, antracenos, grupos esteroides condensados en anillos aromáticos, aminoácidos y compuestos organometálicos [18]. Los organogeladores poseen una amplia diversidad estructural, lo que da pie al amplio rango de propiedades físicas y características reológicas en los geles que producen. Estos geles pueden ser materiales elásticos y hasta tixotrópicos, visco elásticos o pseudoplásticos.

La funcionalidad y las aplicaciones potenciales de los organogeles los hacen aptos para ser útiles en campos muy diversos, como el de la medicina, siendo usados como sistemas de

administración de medicamentos y componentes bioactivos; la electrónica, como sensores; la petroquímica, como agentes de recuperación de derrames de petróleo; o como nuevos materiales blandos con propiedades mecánicas adaptadas derivados de su versatilidad y diversidad estructural. En la industria alimentaria, los geles moleculares ofrecen productos como materiales estructurantes de aceite comestible para mitigar la migración de aceite en sistemas multifásicos como las confecciones de chocolate, y reemplazar los ácidos grasos saturados y trans por alternativas más saludables.

3.3 Solventes, el medio de gelificación de un organogelador.

Un solvente es el medio en el cual, al ser propicio, se llevará a cabo la formación de un organogel. Esencialmente, buscaremos que las características del solvente le permitan a la molécula asociarse con interacciones fuertes y direccionales entre sí. Esta selectividad permite que las moléculas puedan interactuar entre ellas y no con el solvente, formando así los entramados.

Para saber cuándo una molécula puede comportarse como organogelador en un solvente, se requiere información del medio con el fin de evaluar la compatibilidad de ambas entidades. Ciertos parámetros que caracterizan a un solvente son capaces de aportar información de su comportamiento fisicoquímico, y estos suelen ser utilizados para decidir el medio en el cual una molécula candidata a gelador será probada. Estos parámetros suelen ser clasificados por sus propiedades en tres categorías:

- Propiedades físicas
- Propiedades solvatocrómicas
- Propiedades termodinámicas

De una forma similar, un solvente puede ser categorizado de acuerdo con su polaridad de la siguiente manera:

- Próticos
- Apróticos

- Dipolar apróticos
- Apolar apróticos (baja polaridad)

En cuanto a lo que se refiere a las afinidades de las partes involucradas solvente y gelador en una mezcla, materiales con una estructura química muy diferente pueden sin embargo poseer afinidades muy cercanas [18]. Las propiedades de estos compuestos son capaces de brindar datos que, en conjunto, son propuestos como alternativa dentro de modelos predictivos de gelificación.

3.4 Proceso de gelificación.

La gelificación implica la auto asociación de las moléculas geladoras para formar agregados fibrosos largos, similares a polímeros, que se entraman durante este proceso formando una matriz que atrapa el solvente principalmente por la tensión superficial. Este ensamblaje evita el flujo del solvente bajo la gravedad le proporciona apariencia de sólido al entramado. Varios aspectos del proceso por el cual los geladores se agregan para formar geles, como por ejemplo, en qué solvente gelificará una molécula, qué grupos funcionales son requeridos para hacer de una molécula un organogelador, estos estos y otros detalles son poco conocidos, por lo que, el proceso de formación de gel sigue siendo un área de intenso interés.

La manera más eficiente de “inmovilizar” un gran volumen de solvente con una pequeña cantidad de gelador es que los ensamblajes elementales se vean como barras [19]. La estructura de los agregados lineales en el gel está determinada por la dirección y la fuerza de las uniones asociadas con el proceso de gelificación. Las estructuras ensambladas resultan de un equilibrio entre las fuerzas atractivas y las fuerzas repulsivas entre los grupos principales [20]. En el equilibrio, la configuración óptima se obtiene considerando la influencia que poseen las partes polares de la molécula y la polaridad que aporta la energía libre del solvente. A medida que las moléculas se auto ensamblan, la energía libre total del sistema disminuye, lo cual permite que el organogel asuma un estado estable de baja energía.

Los organogeles usualmente se preparan calentando el gelador en un solvente apropiado y enfriando la solución supersaturada isotrópica resultante a temperatura ambiente. Cuando la solución caliente se enfría, las moléculas comienzan a condensarse y pueden aparecer tres posibles situaciones: (1) agregación altamente ordenada que da lugar a cristales, es decir, cristalización, (2) agregación aleatoria que resulta en un precipitado amorfo, (3) un proceso de agregación intermedio entre estos dos, produciendo un gel.

El tipo de producto que se obtiene durante este proceso, depende en gran medida de las interacciones que se llevan a cabo entre el solvente y la molécula. Estas fuerzas son conocidas como interacciones intermoleculares, y son descritas en el siguiente apartado.

3.4.1 Interacciones intermoleculares.

La materia estimada como supramolecular, a la cual se pertenecen los organogeles, depende de las interacciones físicas que se llevan a cabo entre sus componentes. Organogeladores y solventes interaccionan mediante tres tipos principales de fuerzas intermoleculares. Las más generales son las interacciones no polares. Las interacciones no polares se derivan de las fuerzas atómicas, y también se les conocen como interacciones dispersivas. Dado que las moléculas se forman a partir de átomos, todas las moléculas contendrán esta clase de fuerza atractiva.

Otra energía de cohesión, es el enlace o puente de hidrógeno, mismo que sucede a partir de un intercambio de electrones. Un puente de hidrógeno es una interacción molecular y presenta semejanza con las interacciones polares. Los alcoholes, glicoles, ácidos carboxílicos y otros materiales hidrófilos tienen altos parámetros de puente de hidrógeno.

Existen otras fuentes de energía de cohesión, dependiendo del tipo de molécula como, por ejemplo, los dipolos inducidos, enlaces metálicos, interacciones electrostáticas o cualquier tipo de energía separada. Tal cantidad de tipos de fuerzas físicas no covalentes encargadas de formar los entramados supramoleculares que constituyen los geles, es una de las razones que vuelve trascendente el estudio de esta materia.

Debido a la gran cantidad de información que el fenómeno de gelificación contiene es que se han desarrollado ciencias capaces de procesar de forma eficaz y veloz enormes cantidades de datos con el fin de optimizar esta clase de procesos experimentales.

3.5 Modelos de predicción de gelificación.

Con el fin de procesar los datos adecuados que permitan extraer información útil acerca de qué estructura molecular favorece la gelificación en un solvente dado, es que se ha generado una rama de la ciencia que combina el conocimiento químico de estas especies y su estudio mediante herramientas de alta capacidad para asimilar y conectar información con una respuesta requerida.

En los siguientes apartados, se describen conceptos relacionados con el diseño de un modelo de datos capaz de aportar información del proceso de gelificación, mediante procesos elaborados a través de inteligencia artificial.

3.5.1 Parámetros moleculares con habilidades predictivas.

Cómo una huella digital química que aporta información sobre un compuesto, algunas propiedades fisicoquímicas han sido utilizadas para ligar a un solvente con su capacidad de crear interacciones intermoleculares relacionadas con algunas de las propiedades de la molécula con quien se conjuga en un gel [21].

Uno de los pioneros en hacer uso de las propiedades de solventes con la finalidad de crear patrones de reconocimiento de estados, fue Charles Hansen. De acuerdo con palabras propias de Hansen, su trabajo con respecto a los parámetros de solubilidad de una diversidad de solventes comenzó con la finalidad de definir la afinidad entre solventes y polímeros para poder predecir el grado de unión entre ellos que pudiese controlar la retención del solvente. Sin embargo, no obtuvo una buena respuesta al concluir que no existe una correlación entre

estos parámetros. Los solventes con una estructura molecular más pequeña y lineal se difunden fuera de las películas más rápidamente que aquellos con una estructura molecular más grande y ramificada [18].

3.5.1.1 Parámetros de solubilidad de Hansen.

Estos parámetros han sido utilizados desde 1967 hasta la fecha para hacer correlaciones sistemáticas. Han sido exitosos modelando un comportamiento de los componentes del sistema al momento de la gelificación para una amplia gama de geladores particularmente moléculas de gran tamaño que involucran los tres parámetros. Se derivan de la energía de cohesión que se requiere para convertir un líquido en un gas, por lo cual suelen ser conocidos también como *parámetros de energía de cohesión,* tal terminología es usualmente utilizada cuando se habla acerca de fenómenos de superficie como la manera más apropiada de referirse a ellos.

Para Charles Hansen, el enfoque de los parámetros de solubilidad de un solvente reside en dividir la energía total de vaporización de un líquido en partes individuales, cada una capaz de describir las diferentes fuerzas de energía cohesiva que actúan en este. Así, esta subdivisión se plantea de la siguiente manera: fuerzas de dispersión (atómicas o fuerzas de van deer Waals, δd), fuerzas dipolares o dipolo - dipolo (permanentes, δp) y puentes de hidrógeno (moleculares de intercambio electrónico, δh). Estas energías surgen a partir de las interacciones de las moléculas con otras de su propio tipo.

La comprobación de las correlaciones entre estas propiedades y el estado gel está basada en las caracterizaciones a partir de técnicas como SEM, POM y XRD, que reflejan el impacto de la estructura del gelador en su autoensamblaje. Se sabe que, es posible elevar el desempeño de predicción al combinar parámetros como los de Hansen con propiedades adicionales [22], posiblemente debido al hecho de que una combinación particular podría caracterizar el mecanismo de gelificación más completamente que las propiedades individuales.

A continuación, se detalla cada uno de los tipos de interacción contemplados en la triada de Hansen.

3.5.1.1.1 Fuerzas de van deer Waals - fuerzas de dispersión de London.

Las fuerzas intermoleculares conocidas como *fuerzas de van deer Waals,* involucran interacciones no covalentes de diferentes tipos, las cuales son descritas en breve a continuación.

- *Fuerzas de orientación:* Son fuerzas de tipo atractivo que conducen a la energía de orientación. Están presentes en las agrupaciones moleculares de moléculas con dipolo permanente, y se orientan de acuerdo con sus cargas. Algunos ejemplos pueden verse en las moléculas de HCl, NH_3 y H_2O.
- *Fuerzas de Inducción.* Se presentan entre moléculas con dipolo permanente. Bajo la influencia una de la otra. Se distorsionan y orientan con relación a una de ellas formando un dipolo inducido. Las cargas contrarias quedan orientadas bajo una energía inducida.
- *Fuerzas de repulsión.* Se manifiestan cuando las nubes electrónicas saturadas empiezan a traslaparse. Se conocen como *energías de repulsión de London.*
- *Fuerzas de repulsión.* Las fueras de dispersión explican el comportamiento de los gases nobles y son el principal contribuyente de las fuerzas de van deer Waals

Estas últimas son de nuestro interés al encontrarse presentes en las moléculas de tipo oxialquil benzoato, por los grupos funcionales que las componen. Las fuerzas de dispersión (δ_d) se asocian comúnmente a una gran parte de los tipos de organogeles, hidrogeles y geles poliméricos. Estas interacciones también son conocidas como *fuerzas de dispersión de London*, y surgen entre multipolos temporales. La energía de dispersión se atribuye a la atracción dada por dipolos inducidos.

Son las fuerzas más débiles y ocurren principalmente entre moléculas no polares debido al movimiento de los electrones en los enlaces, los cuales originan pequeñas cargas

superficiales e instantáneas positivas y negativas resultando en la atracción entre moléculas, debido a la densidad electrónica que se mueve alrededor estas [23].

Las fuerzas de dispersión aumentan con la masa molar, es decir, crecen conforme la molécula se hace más grande debido a que las moléculas con mayor masa molar poseen una mayor cantidad de electrones, y estas fuerzas aumentan con el número de electrones. Aunado al número de electrones, una masa molar más grande es sinónimo de un átomo más grande, lo cual facilita la alteración de la distribución electrónica debido a que el núcleo atrae con una menor fuerza a los electrones externos. Las fuerzas de dispersión son las fuerzas de atracción principales entre moléculas no polares [24].

3.5.1.1.2 Puentes de Hidrógeno.

Debido a la polaridad, existe una atracción entre moléculas cuando una de ellas contiene un átomo de hidrógeno como parte positiva, y la otra uno de oxígeno (por ejemplo), como parte negativa. A esta particularidad se le conoce como puentes de hidrógeno, las cuales son interacciones intermoleculares relativamente débiles, sin embargo, pueden formar parte de entramados moleculares complejos contribuyendo a la unión molecular.

3.5.1.1.3 Fuerzas dipolo-dipolo

Esta clase de interacciones sucede al ocurrir una atracción entre el extremo positivo de una molécula polar, y el extremo negativo de otra molécula. Estas fuerzas pueden ser permanentes al darse entre dos moléculas con momentos dipolares internos, o no permanentes, cuando las entidades son polarizables a pesar de no tener momentos dipolares propios.

3.5.1.2 Polaridad.

La polaridad en una molécula es quien le confiere propiedades de solubilidad. En una mezcla, el solvente actúa sobre el soluto solvatándolo y venciendo las fuerzas no covalentes que lo

mantienen unido. Estas fuerzas de cohesión pueden ser las siguientes: interacciones polares, puentes de hidrógeno, fuerzas de London, etc.

En los compuestos orgánicos, la polaridad crece cuando disminuye el tamaño de la cadena de carbonos, también cuando hay presencia de grupos funcionales polares o existe posibilidad de interacción de tipo puentes de hidrógeno dentro del medio de interacción.

3.6 Ciencia de datos aplicada al pronóstico de gelificación.

Hasta hace algunos años, una persona denominada "estadístico" era considerada un recolector de información y nada más que eso. Posteriormente, la minería de datos hace su aparición, lo que lleva a acuñar el concepto de "analítica de los datos". En un campo interdisciplinario a través del método científico, esta ciencia es capaz de extraer conocimiento del proceso de gelificación para un mejor entendimiento de este, lo que nos lleva a la optimización en el diseño y selección de sus componentes. Debido a la capacidad de procesar grandes y complejas cantidades de información, esta ciencia es considerada de gran utilidad al momento de pensar en la creación de un modelo predictivo para un proceso complejo como lo es la gelificación.

3.6.1 Definición de ciencia de datos.

Los datos pueden definirse como información almacenada en formato digital que puede utilizarse como base para llevar a cabo un análisis que contribuya a la toma de decisiones. La ciencia de datos es una ramificación de las áreas de análisis de datos como la estadística o la analítica predictiva, en esta, ambas ciencias se conjugan para crear especialidades como la minería de datos o el machine learning. Actualmente, mucho investigadores se apoyan en la ciencia de datos para desarrollar procesos como modelos, ecuaciones o algoritmos, con el fin de evaluar e interpretar los resultados [25]. La Figura 4 contesta las preguntas básicas acerca de la ciencia de datos.

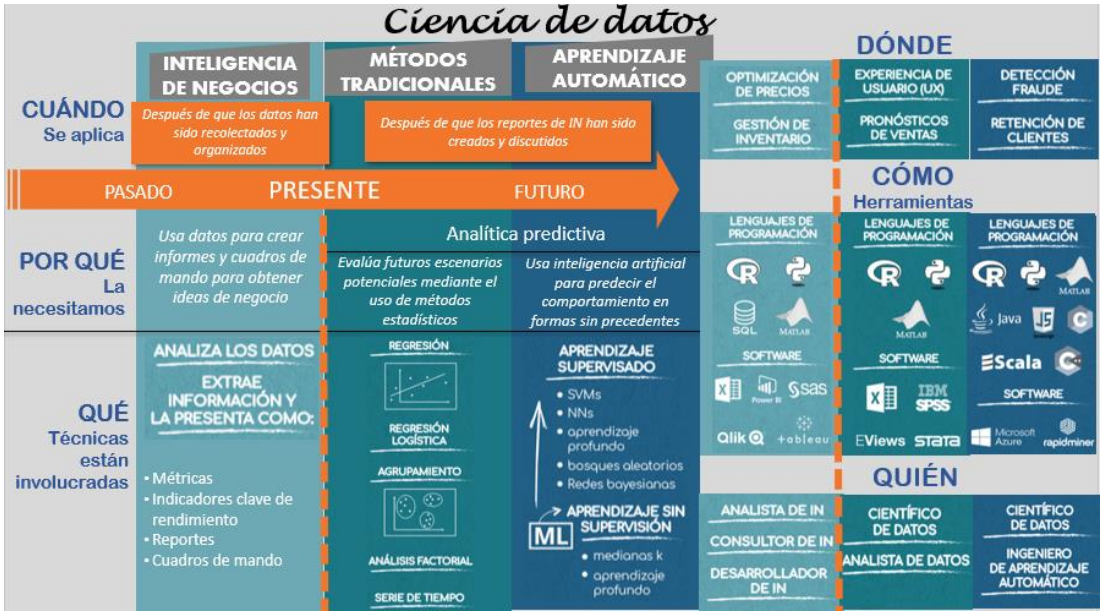


Figura 4. Ciencia de datos su origen y aplicaciones [25].

En el aprendizaje profundo (deep learning), todavía existe un debate sobre por qué los algoritmos utilizados superan el resto de los métodos convencionales. Estos algoritmos además de otros más básicos se encuentran dentro del estudio del machine learning del que se habla a continuación.

3.6.1.1 Machine learning.

El término *machine learning* abarca el conjunto de algoritmos capaces de identificar patrones contenidos en datos y modelar estructuras que los representan.

Dentro de estas entidades, existen diversos tipos de agentes, que poseen atributos que los distinguen de programas convencionales [26]. La cognición y funcionamiento de los agentes que conforman el machine learning viene dado a partir de algoritmos capaces de adaptarse, aprender y predecir mediante la identificación de patrones contenidos en datos. Esta clase de sistemas solo son capaces de memorizar patrones en los datos de entrenamiento, debido a que su capacidad se limita a reconocer sólo lo que han visto previamente. Cuando se usa un sistema entrenado a partir de datos previos con el fin de predecir algo nuevo, la predicción

se basa en que el comportamiento de la nueva entidad será igual a una ya conocida, lo cual no siempre sucede [27].

La categorización del machine learning se divide de la siguiente forma [28]:

• *Aprendizaje supervisado:* el algoritmo de aprendizaje automático tiene como entrada un conjunto de datos con su etiquetado correspondiente. Los datos de aprendizaje deben ser previamente etiquetados.

• *Aprendizaje no supervisado:* No se proporcionan etiquetas al algoritmo de aprendizaje. El algoritmo tiene que encontrar el agrupamiento de los datos de entrada.

• *Aprendizaje de refuerzo:* Un programa informático interactúa dinámicamente con su entorno. Esto significa que el programa recibe comentarios positivos y / o negativos para mejorar su rendimiento.

3.6.1.2 Algoritmos de aprendizaje supervisado.

El aprendizaje supervisado es muy comúnmente utilizado para hacer predicciones con base en comportamientos encontrados en conjuntos de datos previamente alimentados en un algoritmo. Durante este trabajo, el enfoque principal está basado en algoritmos de aprendizaje supervisado, debido a las características que identifican a esta clase de herramientas para la información que será procesada. Los algoritmos de aprendizaje supervisado son capaces de buscar patrones en datos almacenados relacionando todos los campos con uno específico, conocido como campo objetivo [29]. Se realizan predicciones utilizando ejemplos etiquetados (labeled data), es decir el entrenamiento del modelo se realiza con un histórico de datos, donde los algoritmos trabajan con datos de entrada (input data) asignando una etiqueta adecuada, y el sistema aprende a asignar la etiqueta de salida adecuada de acuerdo con el histórico alimentado, prediciendo así el valor de salida.

De manera general, un algoritmo puede ser tan simple o complejo como lo exija el objetivo final que se desea obtener a través de él. Compuestos por series de pasos, un algoritmo

resuelve problemas relativamente simples o más elaborados mediante el uso de series de procedimientos en una computadora. A nivel computacional toman un dato de entrada y también una respuesta ingresada correspondiente, hacen un análisis y encuentran un patrón entre estos, para luego así recibir otros datos de entrada y utilizar lo que aprendieron y dar una salida. El aprendizaje llevado a cabo con los datos de entrada se llama *proceso de formación.* En un nivel básico del machine learning, un algoritmo está formado por códigos contenidos en Java, Python u otro lenguaje de programación.

Los algoritmos de aprendizaje automático se alimentan con los conjuntos de datos provenientes de un corpus de información. El conjunto de entrenamiento señala la respuesta conocida que será de la que aprenda el algoritmo encontrando los patrones existentes con los datos ingresados para aprender sobre el sistema de información que está procesando. Posteriormente con el conjunto de datos de validación o uno de prueba se espera obtener una respuesta del algoritmo basado en la información con la que se entrenó. La Figura 5**¡Error! No se encuentra el origen de la referencia.** ejemplifica la forma esencial en que este proceso sucede.

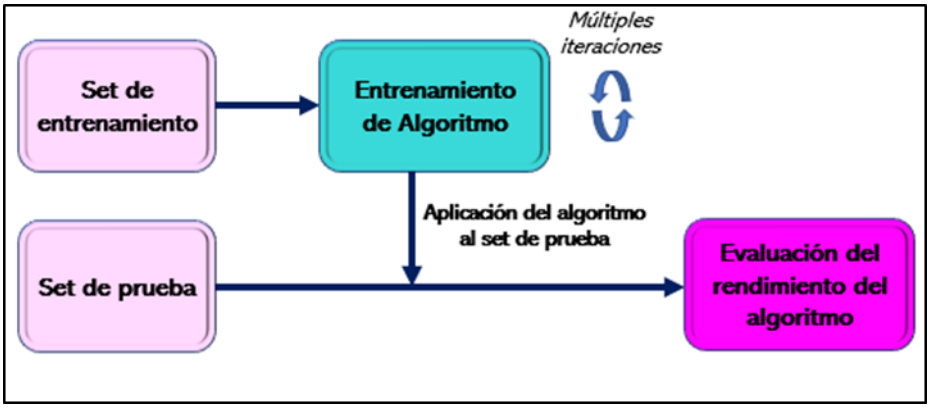


Figura 5. Diagrama de flujo de funcionamiento de un algoritmo de aprendizaje automático.

Una forma común de realizar una mejora de un modelo, es a partir de un modelo de referencia en algoritmos como kNN o Naïve Bayes para datos categóricos, que se usan como referencias

para brindar conocimiento sobre el posible valor predictivo de un conjunto de datos y su diseño.

3.6.1.2.1 kNN.

El algoritmo conocido como vecinos más cercanos o k-Nearest Neighbor (kNN) es uno de los algoritmos de machine learning más sencillos. Su fundamento está basado en identificar los atributos en el conjunto de entrenamiento que se asemejen a los atributos vecinos en el conjunto de datos de prueba, asignándoles como respuesta predictiva la clase predominante entre estas [30]. Pese a su sencillez, en muchos escenarios consigue resultados aceptables.

Este algoritmo usa métodos de aprendizaje no paramétrico, estos métodos permiten que la complejidad de la hipótesis crezca a la par de los datos. El algoritmo kNN está basado en un *aprendizaje por ejemplos* o en memoria, denominado así debido a que construye hipótesis a partir de la información de los ejemplos de entrenamiento. Esta clase de aprendizaje no contiene variables ocultas, por lo cual es necesario que sea un aprendizaje supervisado [26].

El funcionamiento de kNN está basado en la probabilidad de que las propiedades de un ejemplo de entrada particular *n* sean similares a las de los ejemplos cercanos a *n.* Aparentemente es sencillo hasta que surge la necesidad de especificar la *vecindad de los ejemplos.* Si la distancia entre los vecinos es muy pequeña, hay pocas posibilidades de que contenga algún ejemplo igual a los datos de prueba; si es muy grande puede estar sobrepasada con datos que provoquen ruido, dando lugar a una estimación confusa. Lo adecuado entonces es definir una vecindad lo suficientemente grande para que incluya *k ejemplos*, donde *k* es suficientemente grande para asegurar una estimación significativa.

Cuando el conjunto de datos es grande, es necesario aplicar un mecanismo eficiente que encuentre los vecinos más cercanos de un de un ejemplo de prueba *n* [26].

Cuando se aplica el algoritmo kNN se suele utilizar un valor de k vecinos con valor entre 5 y 10, obteniéndose buenos resultados para la mayoría de los conjuntos de datos. Otra manera de seleccionar el valor de k óptimo es mediante la aplicación de una validación cruzada.

Como primer paso para el algoritmo, se aplica su aprendizaje a partir de un conjunto de datos de entrenamiento, se calcula la distancia entre los ejemplos en el conjunto de entrenamiento con el fin de identificar los vecinos más cercanos de un punto. Para la aplicación del algoritmo se requiere de una métrica. La distancia Euclídea es una de las más comúnmente usadas cuando la dimensión del espacio es medida de igual forma, no así cuando la medición es distinta, es decir, se debe de tratar del mismo atributo. Por otro lado, los atributos que caracterizan a los ejemplos de entrenamiento pueden ajustarse a conveniencia con el fin de aprovechar la similitud de los valores de cada ejemplo de entrenamiento según su parecido con los del ejemplo que se desea clasificar. La jerarquía o *peso asignado a los atributos* puede seleccionarse de manera uniforme o priorizando su distancia numérica con respecto a los valores del ejemplo contenido en el conjunto de prueba.

3.6.2 Diseño de un modelo predictivo de gelificación mediante machine learning.

Una herramienta predictiva para fenómenos de índole químico, requiere para su diseño una base de datos con la información adecuada para que un algoritmo pueda aprender la habilidad de encontrar patrones entre las combinaciones de los atributos, calculando la probabilidad de clasificar los puntos de datos pertenecientes a cierto tipo de respuesta. En el caso de la gelificación, debe incluir como salidas factibles los estados de agregación posibles que pueden producirse experimentalmente, es decir; el gel, la precipitación del gelador, la insolubilidad, entre los más comunes. Estos resultados dependen de factores provenientes del medio (solvente), del gelador y de las fuerzas intermoleculares que guían al autoensamblaje. En el descubrimiento de nuevos geladores de manera no fortuita, un paso crucial es identificar cuáles son los parámetros más importantes en el proceso.

Los patrones que se establecen entre las variables de un experimento químico permiten guiar a la optimización de estos procesos, reduciendo las experimentaciones. Durante la última década el uso de redes neuronales y Deep Learning se ha extendido al estudio de la interacción entre una molécula y el medio [28]. A pesar de que estas aproximaciones utilizan descriptores químicos, se ha popularizado la generación de modelos de aprendizaje que

utilizan no solamente esta clase de descriptores, sino también de otros tipos para generar sistemas predictivos mediante modelos de aprendizaje supervisado.

Los problemas desarrollados con algoritmos de machine learning por lo general constan de varias etapas, a continuación, se enlista esta serie de pasos.

a) La definición del problema; ¿qué se pretende predecir?, ¿Qué datos se tienen disponibles? o, ¿qué datos son necesarios? [30].
b) Explorar y entender los datos que se van a emplear para crear el modelo [30].
c) Definir la métrica, como forma apropiada de cuantificar la calidad de los resultados obtenidos [30].
d) Evaluación del modelo. Se debe contar con un conjunto de datos de entrenamiento, uno de prueba y uno de validación (este último suele ser un subconjunto del de entrenamiento). Ninguna información del conjunto de prueba debe participar en el conjunto de entrenamiento del modelo [30].
e) Ajustar un primer modelo capaz de superar resultados mínimos. Por ejemplo, para clasificación, el mínimo a superar es el porcentaje de la clase mayoritaria (moda).
f) Mejorar el modelo gradualmente, optimizando sus hiperparámetros [30].
g) Evaluar la capacidad del modelo final con el conjunto de prueba para estimar la capacidad del modelo al predecir nuevas observaciones [30].

Una vez que han sido seleccionados los datos, se aplica un algoritmo de machine learning para crear un modelo con la capacidad de identificar los patrones presentes en los datos de entrenamiento y aplicarlos a nuevos ejemplos.

La manera más común de desarrollar un modelo ya con los ajustes preestablecidos, comprende una serie de etapas que se mencionan a continuación:

- *Ajuste/entrenamiento.* Se aplican los datos de entrenamiento a un algoritmo para entrenar al modelo.

- *Evaluación/validación.* "El objetivo de un modelo predictivo no es ser capaz de predecir observaciones que ya se conocen, sino nuevas observaciones que el modelo no ha visto" [30]. Para la estimación del error que produce un modelo, se necesitan estrategias de validación, como la aplicación de un conjunto de prueba, o la validación cruzada.
- *Optimización de hiperparámetros.* "Muchos algoritmos contienen en sus ecuaciones uno o varios parámetros que no se aprenden con los datos, a estos se les conoce como hiperparámetros" [30]. Por ejemplo, kNN tiene el hiperparámetro "k" vecinos, que se selecciona de acuerdo con el tipo, robustez y necesidad de los datos. No hay manera de saber *a priori* el valor óptimo de un hiperparámetro que contribuya a producir el mejor modelo, entonces, es necesario aplicar estrategias de validación para probar diferentes valores.
- *Predicción.* Después de haber obtenido el modelo, se aplica para predecir nuevos ejemplos.

3.6.2.1 Corpus de datos para un modelo predictivo.

El corpus de información para un algoritmo de aprendizaje supervisado se divide esencialmente en tres conjuntos diferentes que permiten desarrollar las etapas homónimas previamente mencionadas [31]:

1. *Entrenamiento.* Son los datos con los que se construye el modelo, mismos que se encargan de entrenar al algoritmo.
2. *Validación.* Es una porción de datos utilizada durante la validación de un modelo y prevenir que se sobre o infra ajuste.
3. *Prueba.* Es una última porción de datos que se mantiene aparte y sobre la que se evalúa el modelo. Usualmente se reporta la eficacia del modelo según los resultados en este conjunto.

Como criterio orientativo, usualmente se utiliza el 80% de los datos para entrenar al algoritmo, el 10% de los datos para el conjunto de validación y el 10% sobrante para la

estimación de la precisión del modelo [32]. El reparto debe hacerse de forma aleatoria o aleatoria-estratificada [30].

La importancia de todo modelo estadístico radica en poder entrenarlo con datos que ya hemos visto, y posteriormente usarlo en datos nuevos. Para ello debemos estar seguros de que el modelo no ha simplemente memorizado las muestras de entrenamiento, sino que ha aprendido propiedades del corpus de información. Esta propiedad se llama *generalización* [31].

3.6.2.2 Datos de un corpus.

Otro factor importante dentro del diseño de un modelo en lo que respecta a la información, es el tipo de datos que van a ser aplicados. Aunque el diseño de un modelo adecuado involucra un proceso iterativo, en el cual se van ajustando y probando distintos valores de hiperparámetros, hay ciertos pasos que contribuyen a hacer una selección inicial adecuada [30]:

- Cuando dos datos numéricos están muy correlacionados, estas aportan información redundante al modelo, debido a esto, no es conveniente agregar ambas. En estos casos, es posible excluir el dato que no está realmente asociada con la variable respuesta; o de lo contrario, combinar ambos para utilizar su información en un único dato.
- Cuando un dato tiene varianza igual o próxima a cero sólo aporta ruido al modelo más que información, por lo que es más conveniente excluirlo.

Al entrenar un modelo, es importante incluir los atributos que están realmente relacionados con la variable respuesta, debido a que estas son las que contienen información útil para la predicción. Tener atributos o datos en exceso puede ocasionar que la capacidad predictiva del modelo se reduzca al estar expuesto a nuevos datos [30].

Un punto de suma importancia a tomarse en cuenta cuando se diseña un corpus de datos, es la clase de valores que cada uno poseerá. Los datos pueden ser clasificados de acuerdo con el tipo de valores que toman sus variables. En el siguiente apartado se mencionan los más comúnmente utilizados en modelos de clasificación.

3.6.2.2.1 Datos de tipo cualitativo (categóricos).

Los datos categóricos son también conocidos como datos *cualitativos*. Esta clase de variables permiten clasificar una serie de datos mediante valores que se asocian a una cualidad o categoría. Las variables que pertenecen al tipo categórico clasifican al dato, frecuentemente mediante números enteros tomados como una representación.

Es así que, el manejo de datos categóricos implica trabajar con características ordinales y numerales. Las características ordinales son aquellas de tipo categórico que pueden ser ordenadas (polaridad: alta > media > baja), en tanto que, las características nominales no corresponden a ningún orden (grupo funcional: aromático, éter, alcohol).

Para tener la seguridad de que el algoritmo interpreta adecuadamente las características ordinales, se necesita tener un manejo de valores con números enteros. Por ejemplo, en el caso del atributo polaridad, se pueden asignar las siguientes correspondencias: Alta:2, Media:1, Baja:0 [33].

3.6.2.2.2 Datos de tipo continuo.

Esta clase de datos están representados por variables cuantitativas. Estas variables pueden tomar un número infinito de valores dentro de un intervalo para representar un atributo. A comparación de las variables discretas que sólo pueden ser representadas por números enteros, una variable continua puede componerse por números infinitesimales, por lo cual son capaces de representar un dato con mayor exactitud.

3.6.2.3 Protocolo de evaluación de un modelo predictivo.

El desempeño de un modelo predictivo debe poder medirse. Para este propósito, uno de los métodos más usuales es la aplicación de una validación cruzada. A partir de una validación cruzada se puede establecer lo siguiente.

- El corpus total de datos se debe subdividir en diferentes conjuntos.
- Una forma de puntuación, la cual variará de acuerdo con la naturaleza del problema. Esta puede ser regresión o clasificación.
- Definir en qué algoritmos se probará el corpus de datos.
- Establecer cuál es el modelo con mejor desempeño clasificatorio.
- Optimizar el modelo con mayor desempeño, ajustando sus hiperparámetros.

Una manera de desarrollar una validación cruzada es a través de un número definido de iteraciones. Para esto, los datos del corpus total son divididos en *x* subconjuntos. Es así que, uno de los subconjuntos se usa como conjunto de *prueba*, mientras que el resto de los datos son usados como *entrenamiento.* La validación se repite un número *x* de iteraciones para cada uno de los posibles subconjuntos de datos de prueba. La elección del número de iteraciones es a consideración de la robustez del corpus total de datos. La Figura 6 es un esquema de la repartición de los datos en subconjuntos dentro de una validación cruzada con 4 iteraciones.

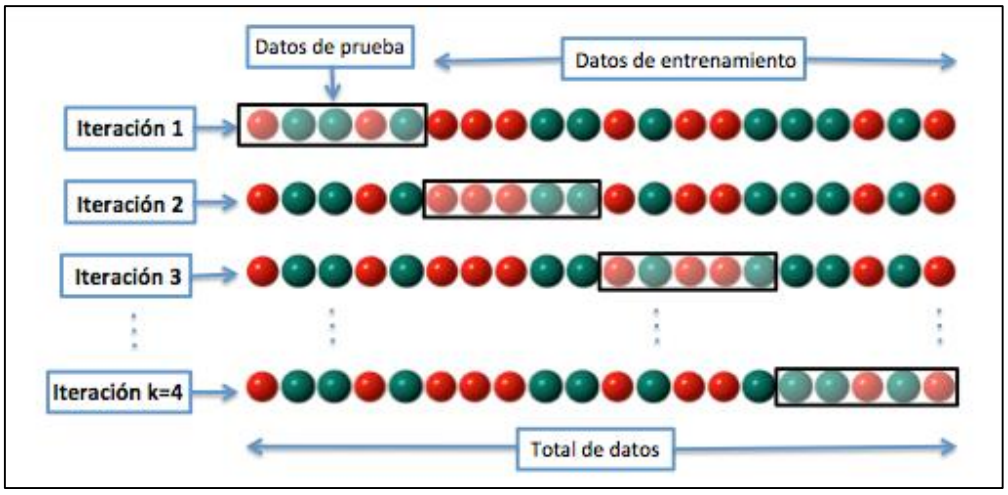


Figura 6. Validación cruzada con 4 iteraciones [34].

Esta metodología es muy precisa debido a que se evalúa a partir de *x* número de combinaciones con los datos, mediante las variaciones en la composición de los conjuntos de entrenamiento y de prueba.

Este procedimiento es útil para comparar los resultados de diferentes procedimientos de clasificación predictiva, es decir, si usamos dos algoritmos diferentes, con una validación cruzada podemos comparar ambos y determinar cuál es el más preciso.

Para obtener resultados significativos mediante la validación cruzada, los conjuntos de validación y prueba deben ser extraídos a partir del mismo corpus total de datos. Por medio de una validación cruzada es posible estimar adecuadamente y optimizar los valores para los hiperparámetros del modelo.

3.6.2.4 Ajuste de la configuración de un modelo predictivo.

Un algoritmo de machine learning pasa por dos tipos de hiperparámetros, el primer tipo son los que se aprenden a mediante la etapa de entrenamiento, y el segundo son los que se ajustan a partir de las evaluaciones del modelo. Es necesario ajustar sus hiperparámetros de un modelo y del algoritmo para obtener el mayor rendimiento predictivo posible.

Regularmente los modelos desarrollados en machine learning utilizan hiperparámetros para aproximar la función con la que se trabaja [35]. Los hiperparámetros son utilizados para organizar y estandarizar la información que se va a ingresar al modelo, esto también se conoce como parametrizar. Estos son herramientas utilizadas para describir la configuración del modelo. En el caso, por ejemplo, de kNN, un hiperparámetro será el valor de *k* vecinos, o la métrica usada. No son usados para modelar los datos, pero tienen gran influencia en la capacidad y características de aprendizaje del modelo [35]. Hay gran dependencia del rendimiento de un modelo predictivo con respecto a los hiperparámetros, pero no se sabe inicialmente cuales son los valores adecuados para estos. Comúnmente se encuentra los

valores óptimos para cada hiperparámetro probando diferentes opciones de estos, hasta encontrar la mayor exactitud al clasificar del modelo [30].

Posterior a las pruebas realizadas con los conjuntos de datos de entrenamiento, se somete a medición al modelo mediante datos no conocidos, a este proceso se le conoce como *generalización*. El propósito es llegar a la más alta capacidad de generalización del modelo [33].

Inicialmente, durante el entrenamiento, los datos de entrenamiento y de prueba están muy relacionados, entre menor sea el desajuste en los datos de entrenamiento, menor será también en los datos de prueba. Esto sucede cuando el modelo está infra-ajustado, lo cual significa que aún hay aprendizaje que desarrollar por parte del modelo, y ajustar sus parámetros relevantes. Después de cierto número de iteraciones con los conjuntos de entrenamiento, se detiene el progreso de generalización, las métricas de validación se paralizan y luego ocurre una degradación. Entonces el modelo esta sobre-ajustado, lo que significa que ha aprendido tan bien los datos de entrenamiento, que sabe pautas muy específicas sobre los datos de entrenamiento que a esas alturas son irrelevantes para datos nuevos.

Existen dos maneras de evitar un sobreajuste del modelo. Una forma puede ser obtener más datos, un modelo entrenado con más datos llevará a cabo una generalización de mejor forma. Si no es posible obtener más datos, entonces se lleva a cabo una regularización, que es el proceso de modular la cantidad de datos que el modelo puede almacenar, o bien, adicionar restricciones con la información que se mantiene desde el inicio.

En el siguiente capítulo, se describen los pasos metodológicos mediante los cuales se desarrollaron una serie de modelos y corpus de datos, que posteriormente fueron evaluados para medir su capacidad de predicción.

Capítulo 4

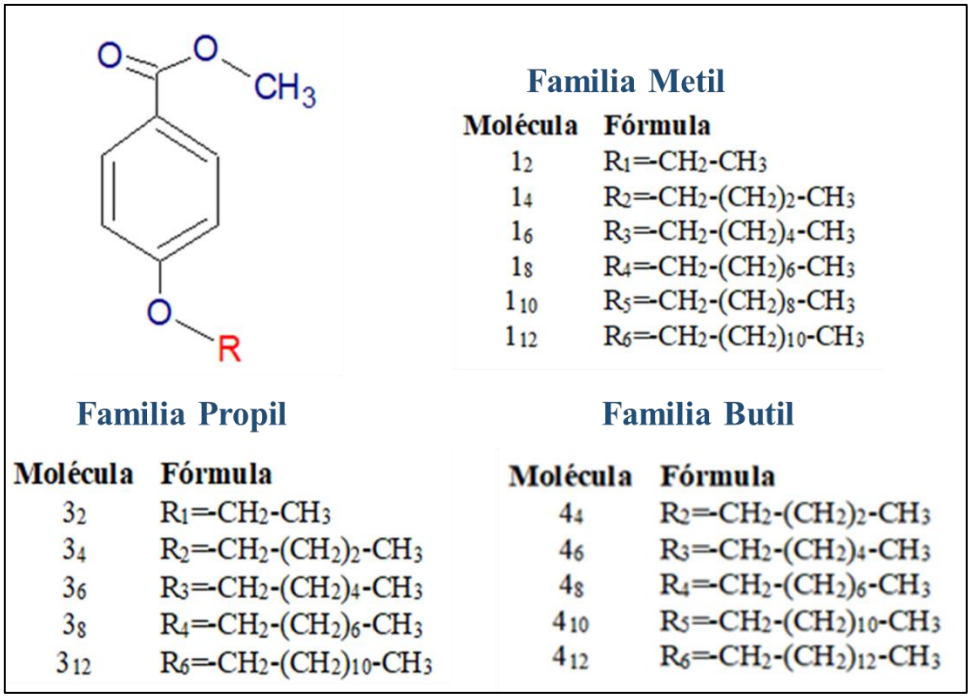
4 Metodología

4.1 Recolección de datos para el diseño de un corpus predictivo.

Para este estudio se analizó la información de una serie de moléculas derivadas de OABs de cadenas éter y éster de longitud variable [2]. Esta información está basada en los resultados de las pruebas de gelificación que comprenden distintos estados producidos mediante la combinación de esta serie de moléculas con un conjunto de solventes polares y no polares de diversas estructuras. Los distintos estados de agregación producidos fueron los siguientes: soluciones, precipitados, insolubles y geles, obtenidos a dos diferentes concentraciones de soluto; 10% y 5% peso / volumen.

Los OABs y los solventes fueron caracterizados estructural y fisicoquímicamente con el fin de diseñar una serie de corpus de datos que pudieran ser utilizados para entrenar y probar el algoritmo propuesto. Mediante estos corpus aplicados a un algoritmo, es posible desarrollar una serie de modelos de predicción capaces de proveer información teórica que facilite la selección adecuada de las características del oxialquilbenzoato y el solvente que promuevan la formación de un gel.

La Figura 7 ilustra la estructura base y los radicales éter y éster de longitud variable pertenecientes a los OABs estudiados.



Familia Metil

Molécula	Fórmula
1_2	R_1=-CH_2-CH_3
1_4	R_2=-CH_2-$(CH_2)_2$-CH_3
1_6	R_3=-CH_2-$(CH_2)_4$-CH_3
1_8	R_4=-CH_2-$(CH_2)_6$-CH_3
1_{10}	R_5=-CH_2-$(CH_2)_8$-CH_3
1_{12}	R_6=-CH_2-$(CH_2)_{10}$-CH_3

Familia Propil

Molécula	Fórmula
3_2	R_1=-CH_2-CH_3
3_4	R_2=-CH_2-$(CH_2)_2$-CH_3
3_6	R_3=-CH_2-$(CH_2)_4$-CH_3
3_8	R_4=-CH_2-$(CH_2)_6$-CH_3
3_{12}	R_6=-CH_2-$(CH_2)_{10}$-CH_3

Familia Butil

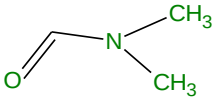
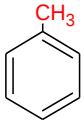
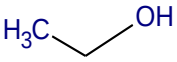
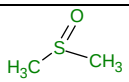
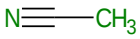
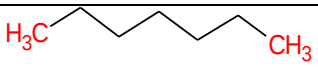
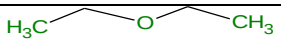
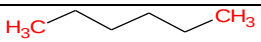
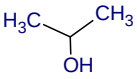
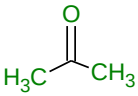
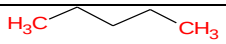
Molécula	Fórmula
4_4	R_2=-CH_2-$(CH_2)_2$-CH_3
4_6	R_3=-CH_2-$(CH_2)_4$-CH_3
4_8	R_4=-CH_2-$(CH_2)_6$-CH_3
4_{10}	R_5=-CH_2-$(CH_2)_{10}$-CH_3
4_{12}	R_6=-CH_2-$(CH_2)_{12}$-CH_3

Figura 7. Estructura de los derivados de OABs pertenecientes a las familias metil, propil y butil [2].

Un total de 16 moléculas y 15 solventes de tipo polar y no polar fueron caracterizadas mediante algunas de sus propiedades cualitativas (químicas y estructurales), y una serie de propiedades fisicoquímicas de carácter continuo, que cuantifican sus interacciones intermoleculares.

Los solventes tienen una alta relevancia durante la gelificación, debido a que es el medio en el que puede ocurrir el autoensamblaje si este posee las características adecuadas. La Tabla 4 muestra el conjunto de solventes estudiados así como sus respectivas estructuras químicas clasificados de acuerdo con su polaridad.

Tabla 4. Clasificación de solventes por polaridad y su estructura.

Polares próticos	Polares apróticos	No polares apróticos
 Metanol	 Dimetil formamida	 Tolueno
 Etanol	 Dimetil sulfóxido	 Ciclohexano
	 Acetonitrilo	 Heptano
	 Acetato de etilo	 Hexano
 Isopropanol	 Acetona	 Pentano
		 Cloroformo
		 Tetrahidrofurano

En la siguiente subsección se explica el diseño de los corpus con datos cualitativos y fisicoquímicos.

4.2 Diseño de corpus de datos para modelos predictivos.

Las características de los componentes de un sistema gel; solvente y gelador, tienen la capacidad de aportar información que contribuya a conocer *a priori* que material se obtendrá a partir de su interacción. Estos rasgos que describen su composición energética y aportan información para clasificarlos distinguiéndolos entre sí. Para su diseño, un corpus de datos se compone de una serie de parámetros, los cuales se describen durante este apartado.

4.2.1 Atributos de un corpus de datos predictivo.

Un corpus formado con datos a partir de las características de un sistema gel puede ser de dos tipos: cualitativo o fisicoquímico, y tiene como finalidad poder ser utilizado durante el diseño de un modelo que prediga los productos a partir de la combinación de una molécula y un solvente dados.

Una variedad de factores pueden causar variaciones durante la gelificación, estos pueden ser; la naturaleza del solvente, la saturación del líquido, la cinética de la interacción de las moléculas, y la estructura molecular de los componentes [36]. Tomando en cuenta estos factores, los corpus cualitativos están compuestos por atributos categóricos que describen la estructura y carácter químico de los OABs y los solventes. Estos atributos se enlistan a continuación.

1. Cantidad de carbonos en parte éster del oxialquilbenzoato.
2. Cantidad de carbonos en parte éter del oxialquilbenzoato.
3. Polaridad del solvente.
4. Heteroátomos en solvente.
5. Linealidad del solvente.
6. Saturación de enlaces en solventes.
7. Grupo funcional predominante en solvente
8. Número de identificación para cada solvente.

Siendo variables categóricas (cualitativas), estas se representan con valores de carácter numérico ordinales (números sin fracciones).

Por otra parte, los atributos para formar un corpus con datos fisicoquímicos requieren ser parámetros que reflejen el carácter, la composición química y la energía de cohesión de las moléculas, así como el comportamiento de los compuestos al combinarse físicamente. Debido a que es posible cuantificar el potencial de las moléculas para interactuar con otras,

se seleccionaron los las interacciones que forman los parámetros de solubilidad de Hansen como atributos de este tipo de corpus, mismos que se enlistan a continuación.

- Interacciones dispersivas (δ_d)
- Interacciones polares (δ_p)
- Puentes de Hidrógeno (δ_h)

Al ser variables fisicoquímicas que representan parte del temperamento cinético molecular de solventes y OABs, sus valores nominales son de tipo numérico continuos.

Con respecto a las condiciones aplicadas durante las pruebas de gelificación, los valores de temperatura (T=25°C), y tamaño del vial fueron constantes en todos los corpus, y estas variables no fueron consideradas como parte de los atributos. En tanto la concentración se toma como un valor constante de 10% v/v para las primeras pruebas y posteriormente se adiciona como un atributo agregando los ejemplos al 5% v/v.

Durante la primera parte de la experimentación, se elaboró una serie de corpus cualitativos, con atributos de tipo ordinal. Se formaron 7 versiones distintas de acuerdo con la cantidad de atributos que describen los 16 OABs y los 15 solventes, lo que resulta en un total de 240 ejemplos por la combinación de cada uno los integrantes de ambos componentes.

4.3 Evaluación de los corpus de datos.

Para estimar el rendimiento de la información y su utilidad interpretativa como parte de modelos de predicción, se utilizaron varios medios de evaluación, mismos que son descritos a continuación.

4.3.1 Validación simple y validación cruzada.

Para la selección de la configuración con mejores resultados de clasificación, inicialmente se aplicaron pruebas de validación de dos tipos: validación simple y validación cruzada. Aplicar una validación consiste en pedir al algoritmo que clasifique un conjunto de ejemplos que ya están contenidos dentro del conjunto de entrenamiento, es decir, ya conocidos por el

algoritmo. Por otro lado, para una validación cruzada o cross-validation, se divide en más de dos subconjuntos el total de los ejemplos, dependiendo de su robustez, utilizando como datos de entrada el número de subconjuntos n-1, mientras se utiliza como prueba uno de ellos, hasta evaluarlos todos de manera aleatoria. Con la validación cruzada se garantiza que los datos son independientes de su partición como entrenamiento y prueba. La Figura 8 muestra la forma de distribución de los ejemplos en los conjuntos para la evaluación de los corpus mediante la validación cruzada.

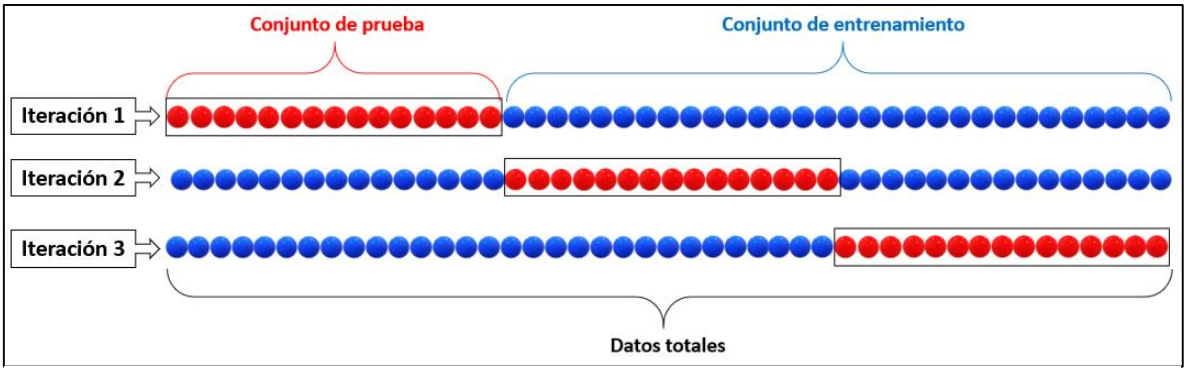


Figura 8. Distribución de ejemplos en los conjuntos de entrenamiento y prueba para validación cruzada.

Por otra parte, durante una validación simple del modelo predictivo en un algoritmo, los ejemplos que forman el conjunto de prueba están contenidas también en el conjunto de entrenamiento, por lo que se evalúa la predicción con ejemplos ya conocidas por el algoritmo. De esta forma, sabemos si las características que representan los atributos descriptivos son capaces de entrenar, en este caso al algoritmo kNN. En Orange Canvas© este método es conocido como *Test on train data.* Hay que señalar que los resultados arrojados a partir de esta prueba no permiten evaluar la efectividad del modelo como predictor, sin embargo, podemos evaluar la funcionalidad de los valores contenidos en los atributos.

4.3.2 Evaluación de Prueba.

Uno de los factores fundamentales en la evaluación de un modelo de predicción es probar con un conjunto de elementos nunca antes visto por el algoritmo, a modo de corroborar su

precisión [5]. Tomando en cuenta esta aseveración, se evaluaron los modelos diseñados mediante pruebas con conjuntos de ejemplos no utilizados para el entrenamiento, pero de la misma familia de moléculas OABs. Tomando en cuenta que, al llevar a cabo la predicción, es vital que el modelo se aplique a moléculas con una química no lejana de la que el algoritmo conoce previamente [37], las moléculas de prueba son homólogas a las utilizadas durante las validaciones previas, teniendo como variante la longitud de las cadenas alquílicas de sus grupos éster.

Para aplicar las diferentes evaluaciones descritas, los corpus con el total de datos se dividieron en subconjuntos; de entrenamiento, de validación y de prueba. Cada subconjunto desarrolla una función específica al momento de la evaluación; los subconjuntos de entrenamiento contienen información a partir de ejemplos caracterizados con la cual el algoritmo aprende para posteriormente clasificar ejemplos desconocidos; los subconjuntos de validación funcionan como un primer ensayo que mide la precisión del modelo para clasificar; y por último los subconjuntos de prueba estiman la función real del modelo para predecir la clase de producto de ejemplos desconocidos.

4.4 Configuración y ajustes de los modelos de clasificación.

Todos los corpus desarrollados, los cualitativos y los fisicoquímicos se utilizaron como parte de las configuraciones de los modelos aplicados al algoritmo kNN. Para la evaluación de los corpus cualitativos, se aplicó la validación simple y la validación cruzada. Para el caso de la validación cruzada el número de subconjuntos varió dependiendo de la cantidad de ejemplos en los corpus. Los subconjuntos se eligieron de forma estratificada por medio de una herramienta en el software.

En tanto, para estimar el desempeño de los corpus fisicoquímicos, se emplearon las técnicas de validación cruzada y prueba. Los valores aplicados a las variables del algoritmo kNN para los corpus cualitativos y para los fisicoquímicos se muestra en la Tabla 5.

Tabla 5. Configuraciones del algoritmo kNN aplicadas en las evaluaciones de los corpus cualitativos y fisicoquímicos.

Corpus	Vecinos (k)	Métrica	Peso de atributos	Evaluación
Cualitativos	3	Euclídea	Uniforme	Validación simple
	5			Validación cruzada
	10			Prueba
Fisicoquímicos	3	Euclídea	Uniforme	Validación cruzada
	5	Chebyshev	Distancia	Prueba

A partir de la combinación de los diferentes corpus diseñados y las configuraciones propuestas para el algoritmo, se formaron una serie de modelos. En la Tabla 6 se describen los diferentes modelos desarrollados. Durante cada evaluación se trabajó con uno o varios corpus de composición diferente. Cada uno de los corpus tanto cualitativos como fisicoquímicos se describen en el recuadro correspondiente a "Corpus y composición", de acuerdo con los atributos contenidos y el tipo de estos.

Tabla 6. Descripción del total de modelos predictivos desarrollados.

Evaluación	Componentes	Composición de corpus
Validación simple	•15 solventes •16 oligómeros **Conjuntos:** •Entrenamiento: 240 ejemplos •Validación: 62 ejemplos	**Corpus cualitativos** Compuesto por 7 atributos categóricos y 2 atributos numérico ordinales. •Identificación de oligómeros - categórico •Nombre de solventes - categórico •Número de éter de oligómero •Número de éster de oligómero •Polaridad de solventes - categórico •Heteroátomos de solventes - categórico •Estructura de solventes - categórico •Saturación de enlace de solventes - categórico •Grupo funcional de los solventes - categórico El número de identificación de cada solvente se asigna del 0 al 15 para cada uno. Etiqueta de Clases: G, I, P y S
Validación simple	•15 solventes •16 oligómeros **Conjuntos:**	**Corpus cualitativos** Compuesto por 9 atributos numérico ordinales. •Identificación de oligómeros •Nombre de solventes •Número de éter de oligómero •Número de éster de oligómero

	•Entrenamiento: 240 •Validación: 62 ejemplos	•Polaridad de solventes •Heteroátomos de solventes •Estructura de solventes •Saturación de enlace de solventes •Grupo funcional de los solventes El número de identificación de cada solvente se asigna del 0 al 15 para cada uno. Etiqueta de Clases: G, I, P y S
Validación simple	•15 solventes •16 oligómeros **Conjuntos:** •Entrenamiento: 240 •Validación: 62 ejemplos	**Corpus cualitativos** Se cambiaron los valores de los atributos que describen a los solventes. El número de identificación de cada solvente se asigna de 10 en 10 para cada uno. Etiqueta de Clases: G, I, P y S
Validación simple y Prueba	• 15 solventes • 16 oligómeros **Conjuntos:** • Entrenamiento: 240 • Validación: 62 ejemplos	**Corpus cualitativos** Se cambiaron los valores de los atributos que describen a los solventes. El número de identificación de cada solvente se asigna de con valores numéricos que van desde el 10 hasta el 6010 sin un intervalo específico, tomando en cuenta el tipo de estructura de cada uno. Etiqueta de Clases: G, I, P y S
Validación simple	• 15 solventes • 16 oligómeros **Conjuntos:** • Entrenamiento: 240 • Validación: 62 ejemplos	**Corpus cualitativos** Se cambiaron los valores de los atributos que describen a los solventes. El número de identificación del solvente se dividió en 3 partes, mismas que representan el peso molecular de los grupos funcionales y el grupo principal de cada solvente. Etiqueta de Clases: G, I, P y S
Validación simple	• 15 solventes • 16 oligómeros **Conjuntos:** • Entrenamiento: 240 • Validación: 62 ejemplos	Se cambiaron los valores de los atributos que describen a los solventes. El número de identificación de cada experimento 6, 7, y 9 se multiplicaron por 100 para ampliar el rango de valores numéricos. Etiqueta de Clases: G, I, P y S

Validación simple	• 15 solventes • 3 oligómeros: familia dodecil éter solamente. **Conjuntos:** • Entrenamiento: 30 ejemplos • Validación: 15 ejemplos	**Corpus fisicoquímico A, B y C compuesto por 5, 6 y 8 atributos respectivamente:** Etiqueta de Clases: G, I, P y S
Validación cruzada	• 15 solventes • 3 oligómeros: familias octil, decil y dodecil éter por separado **Conjuntos:** • Entrenamiento: 30 ejemplos Validación: 15 ejemplos	**Corpus fisicoquímico A compuesto por 5 atributos:** • HSP de solventes (3 valores) • Número de éter de oligómero • Número de éster de oligómero Etiqueta de Clases: G, I, P y S
Validación cruzada	• 15 solventes • 3 oligómeros: familias octil, decil y dodecil éter por separado **Conjuntos:** • Entrenamiento: 30 ejemplos • Validación: 15 ejemplos	**Corpus fisicoquímico B** **Corpus compuesto por 6 atributos:** • HSP de solventes (3 valores) • HSP de OABs (3 valores) Etiqueta de Clases: G, I, P y S
Validación cruzada	• 15 solventes • 3 oligómeros: familias octil, decil y dodecil éter por separado **Conjuntos:** • Entrenamiento: 30 ejemplos Validación: 15 ejemplos	**Corpus fisicoquímico C** **Corpus compuesto por 8 atributos:** • HSP de solventes (3 valores) • HSP de OABs (3 valores) • Número de éter de oligómero • Número de éster de oligómero Etiqueta de Clases: G, I, P y S
Validación cruzada	• 15 solventes • 3 oligómeros: familias octil, decil y dodecil éter por separado	**Corpus fisicoquímicos de estructuras A, B y C** **+ Concentración en cada uno**

	Conjuntos: • Entrenamiento: 60 ejemplos • Validación: 30 ejemplos	
Validación cruzada	• 15 solventes • 9 oligómeros: familias octil, decil y dodecil éter juntas **Conjuntos:** • Entrenamiento: 90 ejemplos • Validación: 45 ejemplos	**Corpus fisicoquímicos de estructuras A, B y C**
Validación cruzada	• 15 solventes • 9 oligómeros: familias octil, decil y dodecil éter juntas **Conjuntos:** • Entrenamiento: 180 ejemplos • Validación: 90 ejemplos	**Corpus fisicoquímicos de estructuras A, B y C** **Etiquetas de clase: YES / NO** **Concentración Variable como atributo**
Prueba	**Conjuntos entrenamiento** • 15 solventes • 9 oligómeros: familias octil, decil y dodecil éter juntas • 135 ejemplos **Conjuntos Prueba** • 15 solventes • Oligómeros 1_{14} y 2_{14} por separado • 15 ejemplos	**Corpus fisicoquímicos de estructuras A, B y C** **Etiquetas de clase: YES / NO** **Concentración constante**
Prueba	**Conjuntos entrenamiento** • 15 solventes • 9 oligómeros: familias octil, decil y dodecil éter juntas • 270 ejemplos	**Corpus fisicoquímicos de estructuras A, B y C** **Etiquetas de clase: YES / NO** **Concentración Variable como atributo**

	Conjuntos Prueba • 15 solventes • Oligómeros 1_{14} y 2_{14} por separado • 15 ejemplos	
Prueba	**Conjuntos entrenamiento** • 15 solventes • 9 oligómeros: familias octil, decil y dodecil éter juntas • 186 ejemplos **Conjuntos Prueba** • 15 solventes • Oligómeros 1_{14} y 2_{14} por separado • 15 ejemplos	**Corpus fisicoquímicos de estructuras A, B y C** **Etiquetas de clase: YES / NO** **Ejemplos distribuidos en cantidades uniformes de acuerdo con su clase** **Concentración constante**
Prueba	**Conjuntos entrenamiento** • 15 solventes • 9 oligómeros: familias octil, decil y dodecil éter juntas • 366 ejemplos **Conjuntos Prueba** • 15 solventes • Oligómeros 1_{14} y 2_{14} por separado • 15 ejemplos	**Corpus fisicoquímicos de estructuras A, B y C** **Etiquetas de clase: YES / NO** **Ejemplos distribuidos en cantidades uniformes de acuerdo con su clase** **Concentración Variable como atributo**

En el siguiente capítulo se presenta el desarrollo de las evaluaciones a todos los corpus, así como los ajustes en sus configuraciones y del algoritmo kNN.

Capítulo 5

5 Desarrollo

En la presente sección se describe el desarrollo del diseño y la evaluación de los diferentes modelos de clasificación, así como su utilidad interpretativa para predecir cuándo una molécula gelificará en un solvente específico.

Inicialmente se describe el diseño y estructura de los corpus de datos compuesto por una serie de variables que caracterizan solventes y OABs. De manera posterior, se explican las configuraciones del algoritmo y sus ajustes para elevar la exactitud de la clasificación de los modelos predictivos.

5.1 Diseño de corpus con atributos cualitativos (categórico).

Se desarrollaron una serie de corpus de información compuestos con una serie de atributos cualitativos que caracterizan tanto a solventes como a OABs. La distribución de todos los atributos de acuerdo con sus valores alfanuméricos se muestran en la Tabla 7.

Tabla 7. Atributos de OABs y solventes

OABs			Solventes					
Éter	**Éster**	**OAB**	**Solvente**	**Polaridad**	**Heteroátomos**	**Estructura**	**Saturación de enlace**	**Grupo funcional**
1	2	1_2	Tolueno	NPA	Bajo	Cíclico	Insaturado	Aromático
		1_4	Pentano					
	4	1_6	Ciclohexano					Alcano
		1_8	Hexano					
3	6	1_{10}	Heptano	PA	Medio			Éter
		1_{12}	Acetato de etilo					Éster
	8	3_2	Acetona					Cetona
		3_4	Metanol					Alcohol
4	10	3_6	Etanol	PP	Alto	Alicíclico	Saturado	Amida
		3_8	Isopropanol					
		3_{12}	Dimetil formamida					Sulfóxido
		4_4	Dimetil sulfóxido					
	12	4_6	Acetonitrilo					Amina
		4_8	Cloroformo					
		4_{10}	Tetrahidrofurano					Halogenuro
		4_{12}						

Debido a que la evaluación de los corpus de datos en los algoritmos propuestos requiere de valores numéricos, con el fin de aplicar los corpus compuestos por propiedades cualitativas, se asignaron una serie de valores a cada una de las características que se enlistan en los atributos, convirtiéndolos en variables categóricas. Por la naturaleza de las ecuaciones aplicadas con las métricas que se evaluaran a través del algoritmo kNN, los valores numéricos comprenden diferentes tipos, esto para dilucidar qué rangos numéricos producen predicciones con una mayor exactitud.

Cada uno de los valores numéricos asignados a los atributos cualitativos de oxialquilbenzoatos (OABs) y solventes se detalla en las tablas correspondientes en la sección de resultados con sus respectivos porcentajes de clasificación producidos (Capítulo 6).

5.2 Diseño de Corpus de atributos fisicoquímicos.

Se seleccionaron los parámetros de solubilidad de Hansen (HSP) para describir el comportamiento fisicoquímico de solventes y oligómeros, así como propiedades estructurales y experimentales. Durante esta subsección, se describe la composición de diferentes estructuras desarrolladas de corpus de acuerdo con los atributos que las forman.

Se desarrollaron tres diferentes estructuras para los corpus de información, variando los atributos que componen a cada una de ellas. Estas estructuras se desarrollan con la finalidad de estimar el comportamiento de la clasificación bajo distintas composiciones de los ejemplos de acuerdo con las características que representan a solventes y OABs. La Figura 9 esquematiza el contenido de cada estructura de acuerdo con sus atributos.

En la literatura, se tienen referencias de modelos de predicción de fenómenos químicos como la gelificación a partir de propiedades químicas de sus componentes. Algunos de los más comunes son los HSP [38-41] Estos parámetros comúnmente han sido utilizados a partir de tablas y ecuaciones para la caracterización de solventes, y es menos común que sean obtenidos para otro tipo de moléculas con el fin de predecir su comportamiento fisicoquímico.

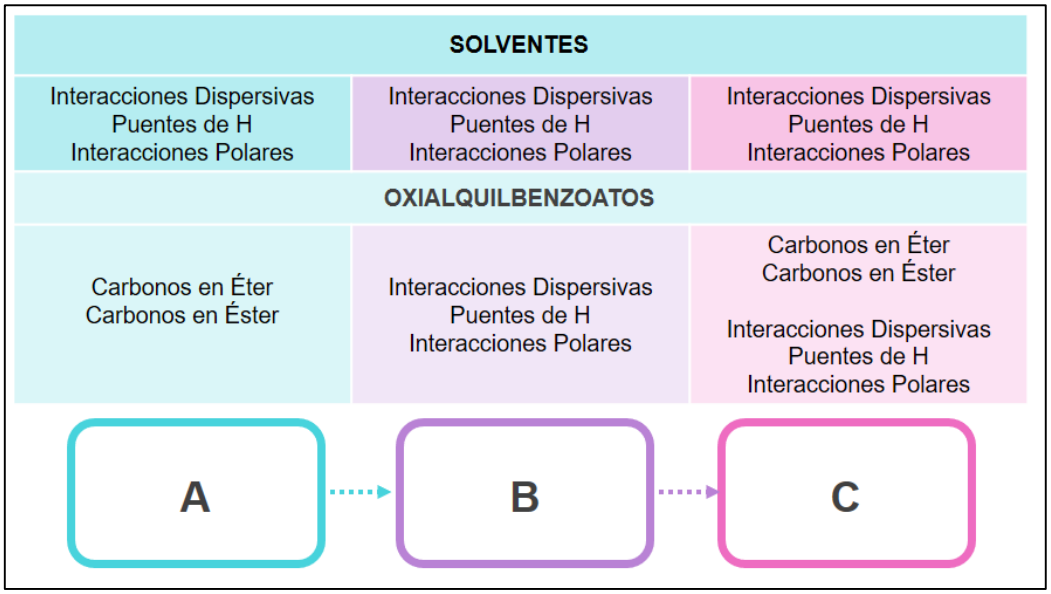


Figura 9. Distribución de atributos en las estructuras de corpus de tipos A, B y C.

Los valores de los atributos que componen al corpus A se muestran en la

Tabla **8**. Las tres interacciones que constituyen los HSP de los solventes fueron obtenidos de un listado del libro Polymer Handbook [42].

Tabla 8. Atributos correspondientes al corpus A.

OAB´s		**Solventes**			
		Nombre	**I. dispersivas**	**I. de H**	**I. polares**
Éter	12	Tolueno	18	2	1.4
		Pentano	14.5	0	0
		Ciclohexano	16.8	0.2	0
		Hexano	14.9	0	0
		Heptano	15.3	0	0
		Acetato de etilo	15.8	7.2	5.3
		Acetona	15.5	7	10.4
Éster	1	Metanol	15.1	22.3	12.3
		Etanol	15.8	19.4	8.8
		Isopropanol	15.8	16.4	6.1
		DMF	17.4	11.3	13.7
	3	DMS	18.4	10.2	16.4
		Acetonitrilo	15.3	6.1	18
	4	Cloroformo	17.8	5.7	3.1
		THF	16.8	8	5.7

Para la obtención de los HSP de los OABs, se desarrollaron una serie de cálculos basados en distintas ecuaciones obtenidas de la literatura.

Cálculo de los HSP mediante Método de contribución grupal de Hoftyzer y van Krevelen.
La interacción de grupos estructurales dentro de las moléculas puede no seguir reglas simples aditivas. Sin embargo, la estimación de los parámetros de Hansen puede resultar muy útil. Según el método de Hoftyzer y van Krevelen [43], los términos se estiman con las siguientes ecuaciones.

$$Ec.1)\ \ \delta_d = \frac{\sum F_{di}}{V}$$

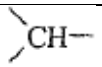
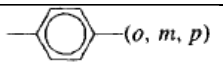
$$Ec.2)\ \ \delta_p = \frac{\sqrt{\sum {F^2}_{pi}}}{V}$$

$$Ec.3)\ \ \delta_h = \sqrt{\frac{\sum E_{hi}}{V}}$$

Los valores de contribución grupal de van Krevelen y Hoftyzer están fundamentados en datos de energía cohesiva de polímeros. Estas técnicas de contribución grupal se basan en el supuesto de que las contribuciones de diferentes grupos funcionales a la propiedad termodinámica son aditivas.

Para el cálculo de los OABs 1_{12}, 3_{12} y 4_{12}, se utilizan los valores presentes en la Tabla 9, aplicados a las ecuaciones 1, 2 y 3. Los volúmenes molares son obtenidos del libro Polymer Handbook, determinados por el método de contribución grupal de Fedors [42].

Tabla 9. Contribución de componentes de los HSP por grupo estructural

Grupo estructural	F_{di} **($J^{1/2}\,cm^{3/2}/mol$)** [43]	F_{pi} **($J^{1/2}\,cm^{3/2}/mol$)** [43]	E_{hi} **(J/mol)** [43]	**V (cm^3/mol)** [42]
$-CH_3$	420	0	0	33.5
$-CH_2-$	270	0	0	16.1
	80	0	0	-1.0
	1270	110	0	52.4
-O-	100	400	3000	3.8
-COO-	390	490	7000	18.0

Las unidades convencionales para los parámetros de solubilidad son (calorías por $cm^3)^{1/2}$, o $cal^{1/2}\ cm^{-3/2}$. Las unidades SI son $J^{1/2}\ m^{-3/2}$, equivalente al $pascal^{1/2}$. Se utilizó el factor de conversión 2,0455 $(MPa)^{0.5}$=1 $(J\ cm^{-3})^{0.5}$=1 $cal^{0.5}cm^{-3/2}$ para expresar los tres parámetros en las mismas unidades.

En la Tabla 10 se presentan los valores correspondientes a los tres parámetros de Hansen obtenidos con la aplicación de las ecuaciones y valores previamente expuestos, para todos los OABs.

Tabla 10. Valores de HSP para los nueve OABs de las familias octil, decil y dodecil éter.

Molécula	I. dispersivas	I. polares	I. de Hidrógeno
1_{12}	36.0074	4.8038	12.9950
3_{12}	35.8525	4.3672	12.3904
4_{12}	35.7852	4.1774	12.1181
1_{10}	35.9500	5.4000	12.0900
3_{10}	35.7900	4.8600	11.4600
4_{10}	35.7200	4.6200	11.1800
1_{8}	36.1700	6.0900	12.8300
3_{8}	35.9600	5.4000	12.0900
4_{8}	35.8700	5.1200	11.7600

5.3 Composición de subconjuntos de datos.

El contenido de esta sección expone cómo a partir de los ejemplos que forman los corpus se integró cada uno de los subconjuntos necesarios para aplicar los datos en el algoritmo; conjuntos de entrenamiento, validación y prueba.

5.3.1 Subconjuntos a partir de corpus cualitativos.

Los primeros estudios estuvieron enfocados en la evaluación de la robustez de los conjuntos que se aplican al algoritmo, los tipos y valores de los atributos de caracterización, y la selección de las configuraciones con mayor exactitud para clasificar. Debido a esto, de inicio los corpus con atributos cualitativos contienen la totalidad de las moléculas OABs estudiadas y el conjunto de solventes en los que se aplicaron las pruebas de gelificación.

La Figura 10 es un compendio breve de la distribución en conjuntos de los corpus de atributos cualitativos.

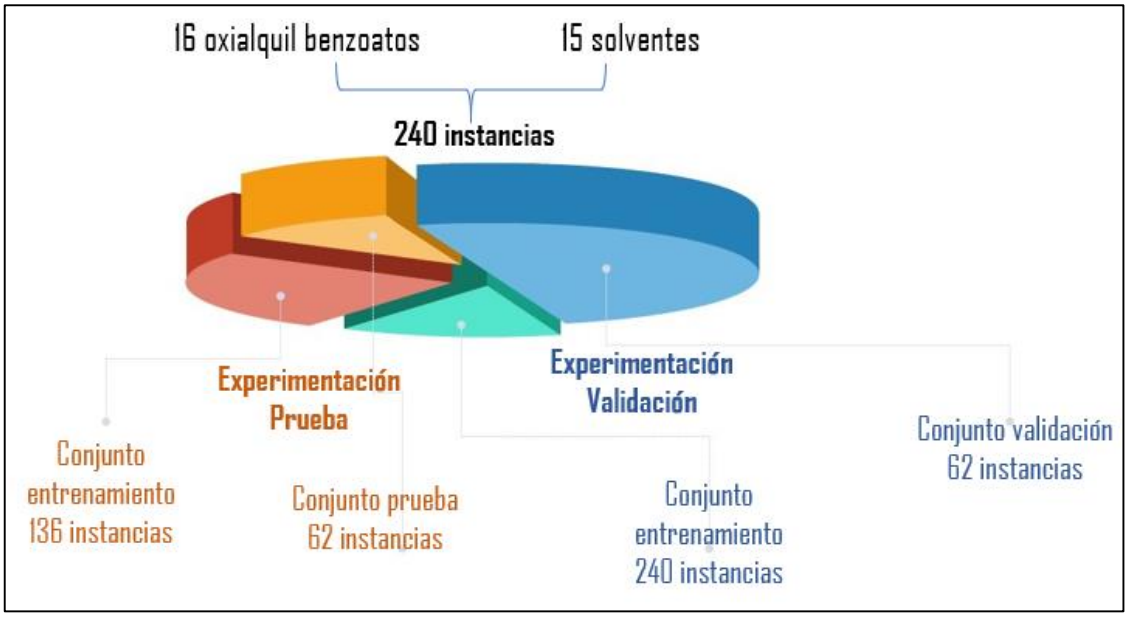


Figura 10. Distribución de atributos en las estructuras de corpus de tipos A, B y C.

A partir de la combinación de cada OAB con cada uno de los solventes, se obtuvo un total de 240 ejemplo o instancias. Estos ejemplos se dividieron en conjuntos para aplicarse en las diferentes técnicas de evaluación en el algoritmo kNN.

5.3.1.1 Validación simple.

En primera instancia, para aplicar la validación simple se dividió el total de corpus de datos en dos tipos de conjuntos: de entrenamiento y de validación. Para la formación de los conjuntos se tomó en cuenta la compatibilidad entre la polaridad de los solventes y la longitud de las cadenas éter y éster en los geladores.

Los conjuntos de entrenamiento se componen del total de los 240 ejemplos, de los cuáles de tomaron 62 ejemplos para validar (conjunto de validación). De manera qué, los ejemplos que se pide clasificar al algoritmo, son contenidos en los datos de entrada. En ambos conjuntos

existen combinaciones gelador-solvente donde se producen 4 estados de agregación, que son las salidas posibles de los modelos; solución (S), gel (G), precipitado (P) e insoluble (I).

5.3.1.2 Evaluación de Prueba.

A partir de los 240 ejemplos de la base de datos, se tomaron al azar un grupo de 136 de manera aleatoria, al que se asignó como conjunto de entrenamiento. Del resto de ejemplos del total, se seleccionaron otros 62 ejemplos para ser clasificados por el algoritmo (conjunto de prueba). Las cantidades mencionadas están basadas en los porcentajes recomendados para la elaboración de conjuntos de entrenamiento y prueba en algoritmos de IA [30].

Las composiciones de estos conjuntos se conservan para las evaluaciones aplicadas a los corpus cualitativos, variando solamente la caracterización de los ejemplos por los atributos que los describen.

5.3.2 Subconjuntos a partir de corpus fisicoquímicos.

Como se mencionó previamente en el capítulo 4, para los corpus formados con atributos de tipo fisicoquímico, se aplicaron diversas evaluaciones, esto con el fin de medir su capacidad como datos de clasificación en un modelo de inteligencia artificial.

Durante este apartado se describen las composiciones de los conjuntos en los que se dividieron los corpus de datos, para utilizarse durante las validaciones y pruebas.

5.3.2.1 Validación simple.

Se aplicó una validación simple sólo para el corpus perteneciente a los ejemplos de los OABs de la familia dodecil éter con los 15 solventes. Los ejemplos de cada conjunto, entrenamiento y validación, se eligieron tomando en cuenta la polaridad de los solventes y la familia éster de cada OAB, a concentración constante de 10%v/v. Esta prueba inicial se aplicó con el fin de tener los primeros datos de clasificación, y hacer una preselección de las configuraciones con desempeño clasificatorio más alto. La composición de los conjuntos queda de la siguiente manera.

- 45 ejemplos totales
- Conjunto de entrenamiento: 45 ejemplos
- Conjunto de validación: 15 ejemplos

Se formaron 3 diferentes conjuntos de validación, a modo de clasificar cada uno de los 45 ejemplos totales. Cada conjunto de validación se aplicó de manera separada, teniendo en todos los casos el mismo conjunto de entrenamiento de 45 ejemplos totales. De este modo, los ejemplos a clasificar, son ya conocidos por el algoritmo al estar contenidos en los datos de entrada.

5.3.2.2 Validación cruzada.

Una evaluación cruzada es una de las técnicas más cercanas a la clasificación de moléculas nuevas a la que se pueda ver sometido un modelo predictivo. Por lo tanto, se desarrollaron diferentes configuraciones para las validaciones cruzadas para estimar las aptitudes de los datos como corpus predictivos.

La Figura 11 muestra un diagrama con las diferentes variables tomadas en cuenta para las configuraciones de todos los conjuntos de datos implementados.

Las variables tomadas en cuenta fueron las siguientes:

- La distribución de los ejemplos de acuerdo con las tres familias éter de los OABs. Se desarrollaron corpus con los ejemplos por familia de manera separada, y con los ejemplos de las tres familias juntas.
- Se tomó en cuenta la concentración de los OABs como una constante al 10%v/v, misma que no se incluyó como un atributo. Después, se agregaron los ejemplos con un 5%v/v de los OAB y se adicionó esta propiedad como un atributo
- A cada conjunto se aplicaron las tres diferentes estructuras de caracterización de acuerdo con los atributos descriptivos de solventes y OABs.

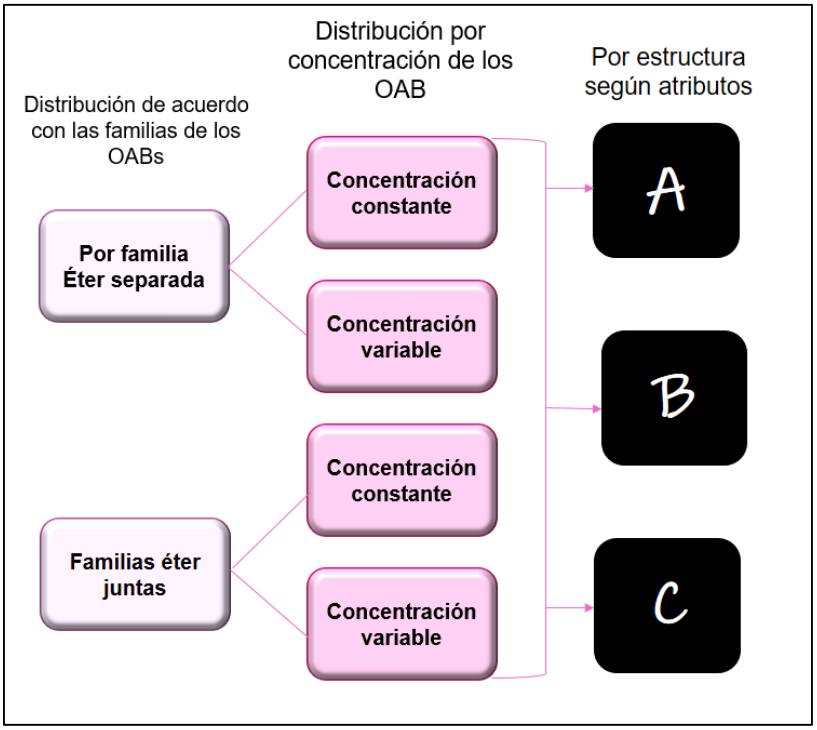


Figura 11. Configuración de conjuntos de entrenamiento y validación aplicadas durante la validación cruzada.

Para las validaciones cruzadas las etiquetas para las clases de los ejemplos fueron G (gel), S (solución) y P (precipitado).

Se dividen los corpus de acuerdo con su familia éter; dodecil, decil y octil donde cada molécula fue probada en pruebas de gelificación con un conjunto de 15 solventes polares y no polares. Las moléculas de la familia dodecil éter son denominadas 1_{12}, 3_{12} y 4_{12} en referencia al incremento en el número de carbonos presentes en el grupo éster (1, 3 y 4) y a los 12 carbonos constantes en el grupo éter de cada una de las tres. De manera similar, se nombran las moléculas de las familias decil éter (1_{10}, 3_{10} y 4_{10}) y octil éter (1_8, 3_8 y 4_8). Se formaron una serie de corpus denominados A, B y C. El corpus A está formado por los HSP de los 15 solventes y los atributos que señalan la cantidad de carbonos en el grupo éter y el éster de los oligómeros, el corpus C agrega los HSP de los solventes, y en el B se caracterizaron solventes y oligómeros sólo en términos de los HSP.

En la Figura 12 se muestran las diferentes composiciones de los conjuntos formados a partir del corpus de ejemplos con atributos fisicoquímicos. De manera general, primero de evaluaron modelos con ejemplos que contienen el total de las moléculas OAB de las tres familias estudiadas, posteriormente, se dividieron los ejemplo por familia.

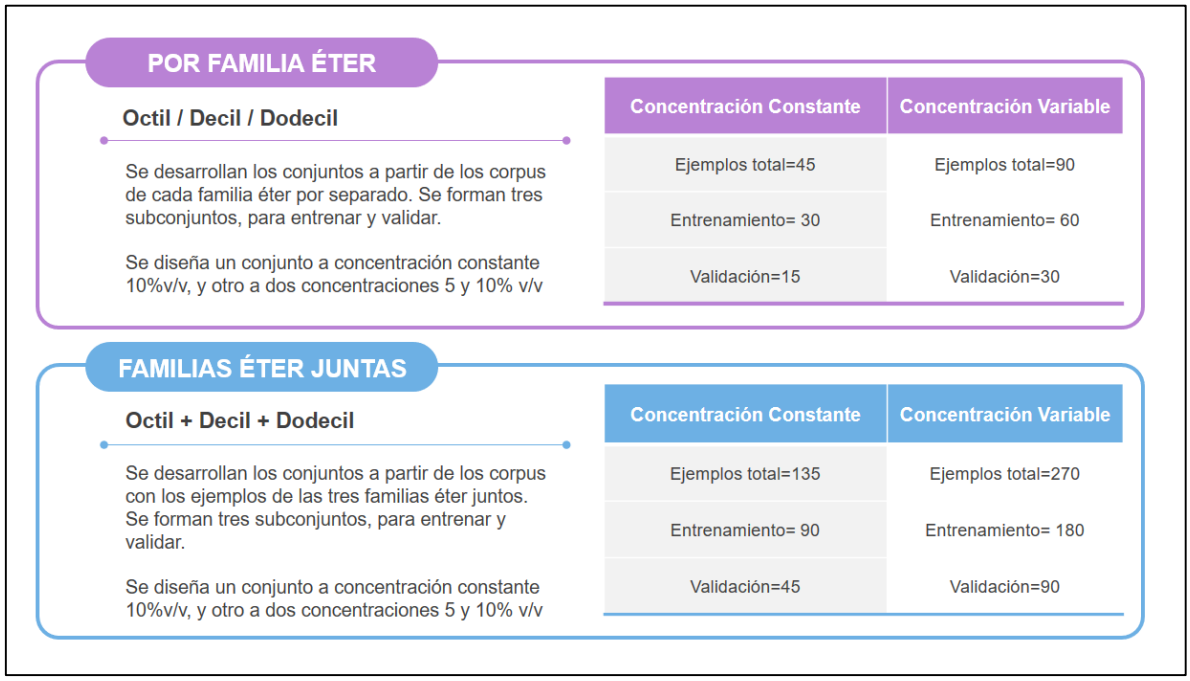


Figura 12. Composiciones de conjuntos formados a partir del corpus de atributos fisicoquímicos.

5.3.2.2.1 Ejemplos a concentración constante.

Inicialmente los ejemplos que se encuentran en cada uno de los corpus corresponden a los datos de pruebas de gelificación a concentración constante del 10%v/v. Al no variar, este parámetro no es tomado en cuenta como un atributo debido a que no contribuye a la distinción entre los ejemplos.

5.3.2.2.1.1 Por familia éter.

Los ejemplos que componen cada corpus para cada familia éter a concentración constante son 45, y están distribuidos como se muestra en **¡Error! No se encuentra el origen de la referencia.**.

Tabla 11. Distribución de número de ejemplos por clase en corpus A, B y C para cada familia éter.

Clase	Ejemplos por familia éter		
	Dodecil	**Decil**	**Octil**
G	16	8	1
S	27	35	42
P	2	0	1
I	0	2	1
Total	45	45	45

5.3.2.2.1.2 Familias éter juntas.

Con el fin de evaluar el comportamiento de corpus robustos por su número de ejemplos, se creó una base de datos formada por los ejemplos de las 3 familias éter, y se diseñó una serie de corpus siguiendo el esquema A, B y C. Cada corpus contiene 135 ejemplos. Los ejemplos de cada corpus A, B y C son 135, distribuidos por su clase como se presenta en la Tabla 12.

Tabla 12. Distribución de número de ejemplos por clase con las 3 familias éter juntas, para los corpus A, B y C con concentración constante.

Clase	Ejemplos
G	25
S	104
P	3
I	3
Total	135

5.3.2.2.2 Ejemplos a concentración variable.

A los corpus previamente diseñados, se agregaron nuevos ejemplos obtenidos de las pruebas de gelificación con una concentración del gelador del 5%v/v. De esta forma, la concentración se adiciona como un atributo con el fin de evaluar su influencia como atributo, y el comportamiento de corpus con una mayor cantidad de ejemplos sobre la clasificación.

5.3.2.2.2.1 Por familia éter.

Se tienen 90 ejemplos en total por familia éter, los cuales se distribuyen por su clase de acuerdo con la Tabla 13. Con base en estos datos, se formaron de forma similar los diferentes corpus de estructura A, B y C.

Tabla 13. Distribución de número de ejemplos por clase en corpus A, B y C por cada familia éter, con concentración como atributo.

Clase	Ejemplos por familia éter		
	Dodecil	Decil	Octil
G	31	15	1
S	57	73	87
P	2	0	1
I	0	2	1
Total	90	90	90

5.3.2.2.2.2 Familias éter juntas.

De manera similar a la base de datos con los ejemplos de las moléculas pertenecientes a las tres familias éter juntas, se formó la base de datos adicionando los ejemplos de concentración de gelador al 5%v/v. La Tabla 14 muestra la distribución por clase de los 270 ejemplos totales.

Tabla 14. Distribución de número de ejemplos por clase con las 3 familias éter juntas, para los corpus A, B y B+, con concentración como atributo.

Clase	Ejemplos sin concentración	Ejemplos con concentración
G	25	47
S	104	217
P	3	3
I	3	3
Total	135	270

5.3.2.2.3 Distribución de ejemplos por clase.

De inicio, los ejemplos en los corpus son los correspondientes a las combinaciones únicas de cada OAB con cada solvente. Debido a que existe una cantidad inferior de ejemplos de tipo

Gel, los conjuntos formados no contenían una distribución de clases homogénea en cantidad, lo que provoca una disminución en el desempeño del modelo al momento de clasificar ejemplos nuevos. Para dar solución a esta problemática, se triplicaron los ejemplos de tipo Gel con el fin de tener una distribución uniforme de los ejemplos de acuerdo con sus clases. Este cambio se aplicó para el corpus formado con la totalidad de los OAB´s (familias éter juntas).

A partir de la distribución de las clases de manera homogénea en cantidad, se formaron nuevos conjuntos de datos que se describen en la Figura 13.

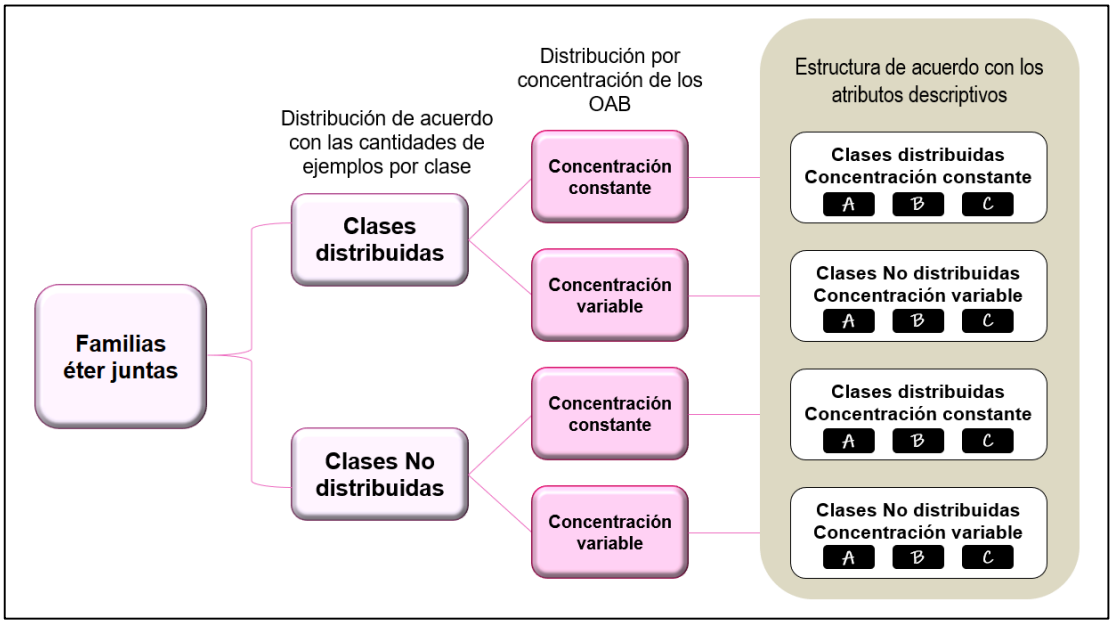


Figura 13. Composición de conjuntos de entrenamiento.

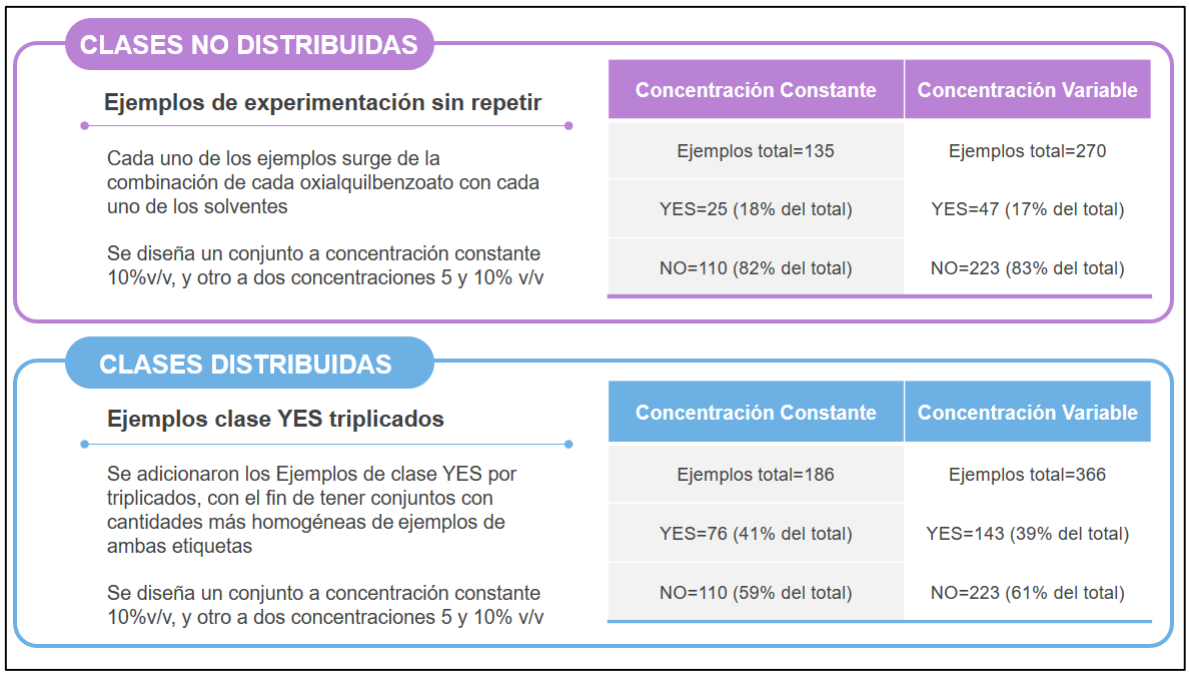


Figura 14. Composiciones de conjuntos de entrenamiento diseñados con la totalidad de los OABs, con la cantidad de ejemplos original, y con ejemplos distribuidos homogéneamente en cantidad de acuerdo con su clase.

En la Figura 14 se muestra la composición de los conjuntos de entrenamiento diseñados con la totalidad de los OABs, con las cantidades de ejemplos únicas y triplicando los que corresponden a la etiqueta Gel, ahora YES, para obtener una distribución imparcial de acuerdo con las clases, con el fin de aumentar el desempeño del modelo.

De acuerdo con los resultados obtenidos durante las validaciones cruzadas, la configuración de kNN que produce el desempeño más alto en el modelo se muestra en la Figura 15.

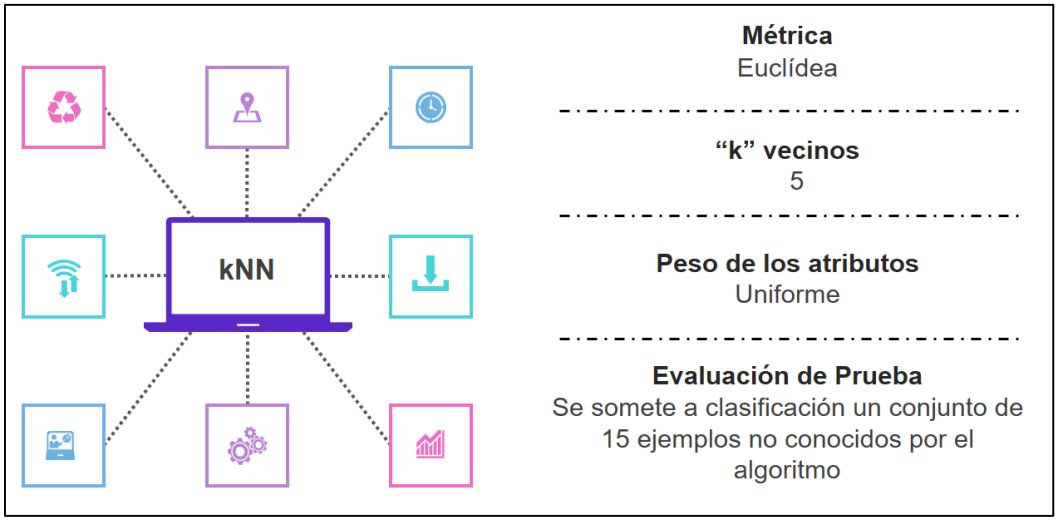


Figura 15. Configuración de kNN que produjo el desempeño más alto del modelo de clasificación.

Entonces, se propone someter a prueba las diferentes composiciones de datos previamente validadas, ahora con un conjunto de Prueba formado con dos OABs desconocidos por el algoritmo. Estos conjuntos se proponen a partir de 15 ejemplos debido a la combinación con cada uno de los solventes analizados en las pruebas de gelificación con las moléculas de entrenamiento.

5.3.2.3 Prueba de clasificación de moléculas 1_{14} y 2_{14}.

Se sometieron a clasificación los conjuntos de Prueba compuestos por los ejemplos formados con la combinación de un conjunto de 15 solventes y las moléculas 1_{14} y 2_{14} respectivamente, como se muestra en la **¡Error! No se encuentra el origen de la referencia.**. De acuerdo con l a concentración del OAB, se manejó un conjunto de entrenamiento a concentración constante del 10%v/v. Para estas pruebas, se aplican algunos cambios de configuración obtenidos a partir de los resultados en la validación cruzada, se decide hacer una distribución de los ejemplos por su clase. Además, se cambiaron las etiquetas con los tipos de productos obtenidos experimentalmente por YES para geles y NO para el resto de los estados de agregación.

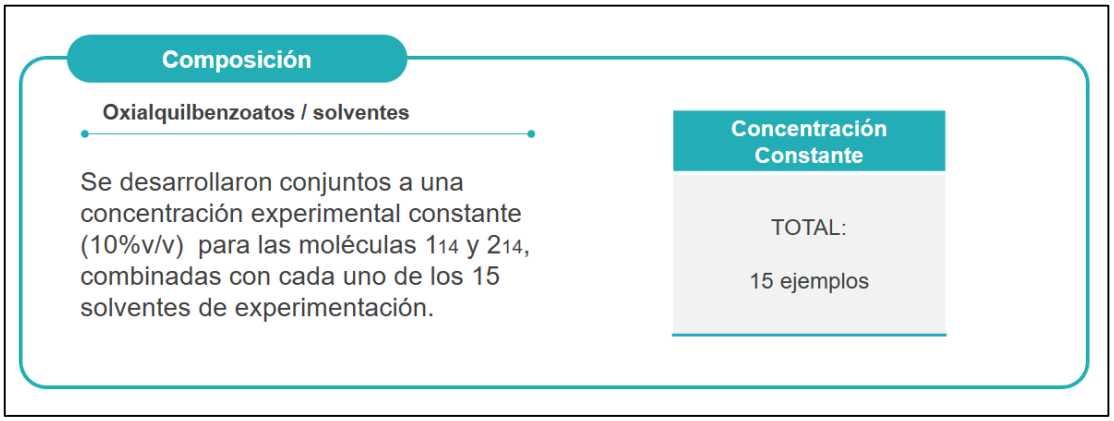


Figura 16. Composición base de los conjuntos de prueba formados con ejemplos de las moléculas nuevas 1_{14} y 2_{14}.

5.3.2.3.1 Prueba con conjuntos de entrenamiento con clases No distribuidas.

La

Tabla **15** muestra los conjuntos utilizados para entrenamiento y para prueba, tanto para la molécula 1_{14} como para la 2_{14}. Se presentan los contenidos de acuerdo con la composición de cada conjunto por cantidad de ejemplos y etiqueta de clase. Los conjuntos de prueba de 1_{14} y 2_{14} tienen la misma estructura en cuanto a cada solvente y tipo de corpus A, B o C, variando solamente en los valores de las características que describen a cada molécula.

Tabla 15. Composición de los conjuntos con clases No distribuidas usados para la prueba de clasificación de las moléculas 1_{14} y 2_{14}.

Conjunto de entrenamiento concentración constante	Conjunto de entrenamiento concentración variable	Conjunto de prueba	
		Concentración constante	Concentración variable
○ Ejemplos Total: 135	○ Ejemplos Total: 270	○ Ejemplos Total: 15	○ Ejemplos Total: 30
○ YES: 25 (18.5% del total)	○ YES: 47 (17% del total)		
○ NO: 110	○ NO: 223		

5.3.2.3.2 Prueba con conjuntos de entrenamiento con clases distribuidas.

La Tabla 19 contiene tanto los conjuntos de entrenamiento como los de prueba evaluados durante este segmento del estudio. En este caso, el entrenamiento se llevó a cabo con conjuntos robustecidos con ejemplos clase YES, de tal modo que el algoritmo pueda

entrenarse con información uniforme en cuanto a las clases, esto con el objetivo de reducir errores de clasificación.

Tabla 16. Composición de los conjuntos clases distribuidas usados para la prueba de clasificación de la molécula 2_{14}.

Conjunto de entrenamiento concentración constante	Conjunto de entrenamiento concentración variable	Conjunto de prueba	
		Concentración constante	Concentración variable
○ Ejemplos Total: 186	○ Ejemplos Total: 366	○ Ejemplos Total: 15	○ Ejemplos Total: 30
○ YES: 76 (41% del total)	○ YES: 143 (39% del total)		
○ NO: 110	○ NO: 223		

Para todos los casos, se probaron en kNN dos valores de *k* vecinos, 5 y 2, debido a una mayor robustez en los conjuntos de entrenamiento.

Resultados experimentales de laboratorio para comparativo.

La Tabla 17. Resultados experimentales de pruebas de gelificación en laboratorio con una serie de nuevos OABs derivados de oxialquilbenzoato.muestra los estados de agregación obtenidos como resultado de las pruebas de gelificación en laboratorio de las moléculas de prueba a una concentración de 10%v/v con cada uno de los 15 solventes. Las pruebas de gelificación se llevaron a cabo a temperatura ambiente en todos los casos (25°C).

Tabla 17. Resultados experimentales de pruebas de gelificación en laboratorio con una serie de nuevos OABs derivados de oxialquilbenzoato.

Solvente	2_{14}		1_{14}	
Hexano	S	NO	G	YES
Ciclohexano	S	NO	S	NO
Heptano	S	NO	G	YES
Acetato de etilo	S	NO	P	NO
Acetona	S	NO	P	NO
THF	S	NO	S	NO
Tolueno	S	NO	S	NO
DMS	G	YES	G	YES
Cloroformo	S	NO	P	NO
Isopropanol	G	YES	P	NO
DMF	G	YES	G	YES
Metanol	G	YES	G	YES
Pentano	S	NO	G	YES
Acetonitrilo	G	YES	G	YES

Etanol	G	YES	G	YES

A partir de esta evaluación, se identifica la eficiencia de los modelos de clasificación de acuerdo con sus configuraciones probadas, determinando también cuál de estas es la que produce valores de clasificación más altos.

5.4 Hipótesis.

De acuerdo con las características descritas, se plantearon una serie de hipótesis sobre los resultados esperados a partir de la aplicación de la información en el algoritmo kNN, mismas que se explican a continuación.

En cuanto al corpus de datos.

1. La experimentación con atributos cualitativos (datos categóricos) pertenecientes a los solventes y los OABs, permite guiar a una configuración óptima de los hiperparámetros que otorguen mayor calidad de clasificación del modelo predictivo.
2. La cantidad de atributos descriptivos es directamente proporcional al aumento en la eficacia en el modelo kNN.
3. La calidad de los atributos en el corpus tiene mayor relevancia que la cantidad de estos para contribuir a una alta clasificación.

4. La calidad de la clasificación varía cuando los atributos descriptivos del modelo son representados por valores numéricos ordinales, y será mayor a comparación de la calidad de clasificación del modelo con atributos alfanuméricos.
5. El aumento en la cantidad de ejemplos y la distribución de estos mismos de acuerdo con su clase, para que se encuentren en cantidades equitativas en el corpus de datos y en los conjuntos de entrenamiento pueden optimizar el desempeño de la clasificación por parte del modelo.
6. Los valores numéricos asignados a cada atributo cualitativo (datos categóricos) impactan directamente a la calidad de la clasificación debido a la métrica utilizada por el predictor.

En cuanto al algoritmo.

1. Para el algoritmo kNN, el valor óptimo en la cantidad de vecinos más cercanos se encuentra entre los valores de k= 5 y 10, y para su selección se debe tomar en cuenta la robustez de los corpus de acuerdo con la cantidad de ejemplos.
2. Entre mayor sea el valor de *k* vecinos, se hace más grande la probabilidad de error en la localización de clases correctas para cada ejemplo, esto debido a ruido.
3. En kNN, la asignación del tipo de peso que se le asigna a los atributos (uniforme o distancia), influye sobre el aumento o disminución del correcto etiquetado de los ejemplos.

Capítulo 6

6 Resultados y discusión de Validación con Corpus cualitativo.

Los datos que conforman el corpus cualitativo fueron trabajados de acuerdo con la descripción de las experimentaciones en la Tabla 6 de la sección 4.4, Configuración y ajustes de los modelos de clasificación**.**. Durante el presente capítulo, se detallan los resultados en dos partes fundamentales, que describen el inicio de la selección de los hiperparámetros del modelo con más alto desempeño inicialmente, mismos que fueron optimizados a lo largo de la experimentación. Se presentan los resultados a partir de un corpus con variables de categoría numérico ordinal, mismos que fueron modulándose en cuanto a los rangos de valores de los atributos, con el fin de aumentar el desempeño clasificatorio del modelo.

6.1 Corpus cualitativo con atributos numérico ordinales.

Se llevaron a cabo experimentaciones con el conjunto de *validación*, traduciendo los atributos cualitativos a numéricos ordinales, es decir siendo representados por valores con números enteros.

Cuando se desarrolla la metodología basada en datos experimentales para crear un corpus que será procesado en él o los algoritmos seleccionados, uno de los pasos trascendentales es la selección analítica de ciertas variables que contribuyen con el buen desempeño del algoritmo al momento de procesar los datos.

Para nuestro caso, los denominados hiperparámetros analizados a través de la experimentación, con el fin de la optimización del modelo son los que se nombran a continuación.

6.1.1 Atributos del corpus y su impacto en la clasificación en kNN.

Para evaluar el impacto de las variables sobre un modelo predictivo en el algoritmo kNN, se llevó a cabo una prueba tomando como ejemplo para el ejercicio el método de validación, aplicando valores de k=3 y 5 atributos. La equivalencia ordinal de los atributos cualitativos se muestra en la Tabla 18. Inicialmente se comenzó probando con 5 atributos, posteriormente se añadieron más al diseño de nuevos corpus.

Tabla 18. Atributos cualitativos de oxialquilbenzoatos (OABs) y solventes de tipo y su equivalente de clase numérico ordinal.

OABs			Solvente		
Éter	**Éster**	**ID**	**Polaridad**		**Nombre de solvente**
1	2	1-2	NPA	0	Tolueno
		1-4			Pentano
		1-6			Ciclohexano
		1-8			Hexano
3	4	1-10	PA	1	Heptano
		1-12			Acetato de etilo
	6	3-2			Acetona
		3-4			Metanol
		3-6			Etanol
		3-8			Isopropanol
	8	3-12			Dimetil formamida
		4-4			Dimetil sulfóxido
4	10	4-6	PP	2	Acetonitrilo
		4-8			Cloroformo
	12	4-10			Tetrahidrofurano
		4-12			

Se conservan atributos alfanuméricos.

Los resultados de esta prueba se compararon contra los resultados obtenidos con su modelo equivalente con datos descritos de tipo alfanumérico. En este caso se tienen resultados significativamente diferentes, mismos que se muestran en la Tabla 19.

Tabla 19. Resultados expresados en porcentaje de clasificación obtenidos en la validación con 5 atributos, tipo alfanumérico vs numérico ordinal.

Modelo	Tipo de atributo cualitativo	
	Numérico ordinal	Alfanumérico
kNN, k=3, 5 atributos	96.77%	79.03%

Con base en las cifras obtenidas, se observa que existe un impacto en la clasificación a causa del cambio en el tipo de variable con la que se representan los atributos cualitativos en un corpus de alimentación de un modelo predictivo en kNN. Se tomó la decisión de llevar a cabo la experimentación con el resto de los atributos numéricos ordinales, representados como se muestra en la Tabla 6 de la sección 4.4.

En la Tabla 20 se aprecian los resultados de la clasificación expresados en porciento, como un comparativo entre las experimentaciones realizadas con atributos cualitativos alfanuméricos vs cualitativos numéricos ordinales, aplicando diferentes valores para k vecinos.

Tabla 20. Resultados de clasificación con el conjunto de validación expresados en porcentaje de acuerdo con valores de *k*, para modelos con variables alfanuméricas vs modelos con variables numéricas ordinales.

Vecinos	**k=3**		**k=5**		**k=10**	
Atributos	***Resultado (%) a. numéricos ordinales***	***Resultados (%) a. alfanuméricos***	***Resultados % clasificación a. numéricos ordinales***	***Resultados % clasificación a. alfanuméricos***	***Resultados % clasificación a. numéricos ordinales***	***Resultados % clasificación a. alfanuméricos***
4	96.77%	90.32%	90.32%	90.32%	87.09%	87.09%
5	93.54%	93.54%	90.32%	90.32%	88.70%	87.09%
6	90.32%	90.32%	87.09%	91.93%	87.09%	91.93%
7	93.54%	90.32%	91.93%	91.93%	90.32%	90.32%
8	93.54%	91.93%	93.54%	91.93%	90.32%	90.32%
9	90.32%	91.93%	90.32%	91.93%	93.54%	88.70%

____ Porcentaje de clasificación mayor.

Al analizar la respuesta de acuerdo con la cantidad de atributos, se observa que no existe una relación directa en una mayor calidad de clasificación proporcional a una mayor cantidad de atributos, si no que el porcentaje más alto se presenta con la cantidad mínima de estos, con un valor de vecindad de k=3. En este orden, es importante notar la posibilidad de que sea el tipo de atributo y no la cantidad de estos los que contribuyen a la mejora en la clasificación, por los porcentajes de aumento y disminución observados a través del incremento en la cantidad de atributos.

Ahora bien, referente a los valores de k, sí se tiene una tendencia proporcional a una disminución en los porcentajes de clasificación cuando la vecindad de hace más grande, siendo la más baja cuando k=10, y significativamente notorio el incremento con las variables numéricas ordinales y la k menor de 3. Esto puede ser el reflejo de la contribución de los valores numéricos al aprendizaje del algoritmo y los cálculos del predictor.

Basados en estas se decide continuar con la experimentación variando los valores nominales de identificación de los nombres de solventes, con el fin de evaluar el impacto del rango numérico con respecto a la métrica utilizada por el predictor en el algoritmo kNN, y la trascendencia de ese atributo en especial por su importancia a nivel experimental, y debido a que su conexión con los OABs posee un potencial implícito de información que aportar al algoritmo.

6.1.2 Evaluación de desempeño con base en identificación numérica de solventes.

Tomando como base los valores presentados en la Tabla 18, donde se exponen los valores numéricos para cada atributo del corpus de información, se varían números de identificación de los solventes. El criterio que se utilizó para su asignación se describe a continuación:

1. Se asignan valores de 10 en 10 para cada solvente de acuerdo con el listado original, sin un orden que obedezca a sus características fisicoquímicas. Los valores van del 10 al 150.
2. Los números hacen referencia a la estructura de cada solvente, a su número de carbonos, y de radicales presentes. Los valores van del 10 al 6010, sin seguir una secuencia definida.
3. La estructura química de los solventes se divide en 3 partes: l primera hace referencia al peso molecular de la estructura principal, y las dos siguientes a los pesos moleculares de los radicales y grupos funcionales presentes en el solvente.

Los valores designados para probar en cada experimentación se presentan en la Tabla 21.

Tabla 21. Valores designados para identificación de solventes en experimentaciones con kNN.

Solvente	ID de solventes por experimentación				
	1	2	3		
Tolueno	10	6010	78	15	0
Pentano	20	50	75	0	0
Ciclohexano	30	600	85	0	0
Hexano	40	60	90	0	0
Heptano	50	70	105	0	0
Acetato de etilo	60	4010	30	15	44
Acetona	70	3010	30	28	0
Metanol	80	10	15	17	0
Etanol	90	20	30	17	0
Isopropanol	100	30	45	17	0
Dimetil formamida	110	90	30	28	14
Dimetil sulfóxido	120	100	15	28	32
Acetonitrilo	130	120	30	0	14
Cloroformo	140	401	12	0	95
Tetrahidrofurano	150	4040	40	16	0

La

Tabla 22 muestra los resultados de la clasificación expresados en porcentaje para cada una de las experimentaciones, con la variación del número de identificación del solvente.

Experimento	**1**			**2**			**3**		
Atributos	3 a 8			3 a 8			5 a 10		
k	3	5	10	3	5	10	3	5	10
Clasificación (%)	98.38	91.93	87.09	100	91.93	88.70	98.38	91.93	88.70
Error	1/62	5/62	8/62	0	5/62	7/62	1/62	5/62	7/62

Tabla 22. Resultados a partir de validación de acuerdo con los modelos con variación de valores de ID del atributo solvente.

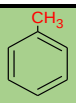
Experimento	**1**			**2**			**3**		
Atributos	3 a 8			3 a 8			5 a 10		
k	3	5	10	3	5	10	3	5	10
Clasificación (%)	98.38	91.93	87.09	100	91.93	88.70	98.38	91.93	88.70
Error	1/62	5/62	8/62	0	5/62	7/62	1/62	5/62	7/62

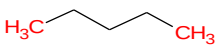
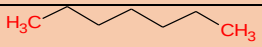
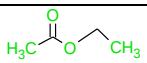
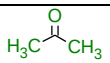
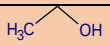
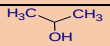
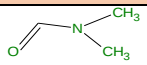
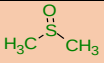
Es muy importante mencionar el reporte en la tabla se refiere a *"Atributos"* de 3 a 8 con sólo un valor como resultado, debido a que el resultado fue el mismo en los modelos de 3, 4, 5, 6, 7 y 8 o 5 al 10 para el caso de la experimentación 3, donde se describe al solvente con 3 atributos. Con esto se presenta la interrogante de la verdadera aportación de cada atributo y sus respectivos valores al modelo predictivo, por su calidad y no por la cantidad de estos.

Los valores asignados a la identificación del solvente en el experimento 2 influyen en la calidad de la clasificación como se aprecia en los porcentajes obtenidos por cada experimento, notoriamente cuando k=3, donde se alcanza el 100% de la clasificación. Por lo que se concluye que los valores de identificación de los solventes tienen relevancia al momento de desarrollarse la regresión logística en el predictor para posteriormente seleccionar los ejemplos más cercanos al punto que se desea clasificar con su respectiva etiqueta de respuesta. Estos valores de identificación contribuyen a diferenciar de una forma significativamente mayor a los solventes que son distintos entre sí, como, por ejemplo, un alcohol de cadena corta como el metanol (10) de un aromático como el tolueno (6010), y a entrenar al algoritmo a procesar estas diferencias, mismas que también se presentan experimentalmente en los estados de agregación que se obtienen a partir de cada solvente.

Se presenta la Tabla 23 para una mejor apreciación de la caracterización numérica de cada solvente con respecto a los números de identificación de la experimentación donde se obtiene el 100% de predicción, sus diferencias y similitudes, y la relación con el estado de agregación obtenido con cada uno en promedio con los diferentes geladores.

Tabla 23. Solventes probados en experimentación y su identificación numérica, estructura y grupos funcionales.

Solvente	**ID solvente (predicción 2, 100% de clasificación)**	**Estructura del solvente**	**Grupo funcional**
Tolueno	6010		aromático

Pentano	50		Alcano
Hexano	60		Alcano
Heptano	70		Alcano
Ciclohexano	600		Alcano
Acetato de etilo	30		Éster
Acetona	2010		Cetona
Metanol	10		Alcohol
Etanol	20		Alcohol
Isopropanol	30		Alcohol
Dimetil formamida	90		Amida
Dimetil sulfóxido	100		Sulfóxido
Acetonitrilo	120		Nitrilo
Cloroformo	103		Halógeno
Tetrahidrofurano	401		Éter

Precipitado
Gel

Solución

Cada color en la tabla distingue un estado de agregación obtenido, mismo que representa una etiqueta de salida, obtenido con el solvente marcado. Se aprecia una tendencia en estados similares con solventes de la misma familia, o con un grupo funcional distintivo, por ejemplo, al poseer un grupo aromático, se presenta un estado de precipitado, o al tratarse de grupos alcohol, se forman geles.

La Tabla 24 presenta una relación más detallada sobre los solventes en combinación específica con los OABs para formar los distintos estados de agregación encontrados en experimentación. La identificación numérica del solvente durante la experimentación 2 donde se obtuvo el 100%, hace énfasis en marcar una diferencia entre los conjuntos de solventes que forman estados distintos, sea dicho, P, G, I y S.

Tabla 24. Relación de estados de agregación producto de la interacción experimental entre solventes (columna izquierda) y OABs (fila superior).

Solvente	**OAB**															
	4_4	**4_6**	**4_8**	**4_{10}**	**4_{12}**	**3_2**	**3_4**	**3_6**	**3_8**	**3_{12}**	**1_2**	**1_4**	**1_6**	**1_8**	**1_{10}**	**1_{12}**
Tolueno	S	S	P	S	S	S	S	P	S	S	S	S	P	S	S	S
Pentano	S	S	S	S	P	S	S	S	S	P	S	S	S	S	S	G
Ciclohexano	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
Hexano	S	S	S	I	G	S	S	S	S	G	S	S	S	S	S	G
Heptano	S	S	I	I	G	S	S	S	S	G	S	S	S	S	G	G
Acetato de etilo	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
Acetona	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
Metanol	S	S	S	S	G	S	S	S	S	G	S	S	S	S	G	G
Etanol	S	S	S	S	S	S	S	S	S	G	S	S	S	S	G	G
Isopropanol	S	S	S	S	S	S	S	S	S	G	S	S	S	S	G	G
Dimetil formamida	S	S	S	S	S	S	S	S	S	S	S	S	S	S	G	G
Dimetil Sulfóxido	S	S	S	G	G	S	S	S	G	G	S	S	S	S	S	G
Acetonitrilo	S	S	S	S	S	S	S	S	S	G	S	S	S	S	S	G
Cloroformo	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
Tetrahidrofurano	S	S	S	S	P	S	S	S	S	S	S	S	S	S	S	S

Tras observar las tendencias, se tiene que las moléculas de cadena éster más pequeña, en este caso la familia metil (*1_x, donde x es el número de carbonos en la parte éter)* existe una mayor obtención de Geles con solventes de cadena alquílica, con grupos oxhidrilo, y en dos casos presencia de Nitrógeno (grupo nitrilo y amida). Para las familias propil y butil, se observa que cuando la cadena éter tiene un contenido de 10 y 12 carbonos (*3_x y 4_x*) existe formación de geles con la misma clase de solventes que la familia metil, sin embargo, en menor proporción. Estas propiedades (átomos de carbono en éster y éter) al ser atributos en el modelo predictivo, representan información trascendental para ligarse a un número de identificación de solventes y permitir al algoritmo aprender de las conexiones entre ellos y los estados de agregación obtenidos.

Es importante saber entrenar al algoritmo con la información adecuada que represente claramente las diferencias y similitudes presentes en la experimentación química, para obtener una predicción de mayor calidad.

En el siguiente capítulo se abordan los resultados obtenidos a partir de la experimentación con corpus de datos fisicoquímicos, donde se aplican algunos de los resultados obtenidos durante los experimentos con datos cualitativos con respecto a los hiperparámetros de los modelos.

Capítulo 7

7 Resultados y discusión de Validación y Prueba con Corpus fisicoquímico.

Como previamente se explicó durante la sección 5.2 donde se detalla el diseño de los corpus de datos formados con atributos fisicoquímicos, durante este capítulo se presentan los resultados obtenidos durante cada una de las evaluaciones con esta clase de atributos, y el progreso en la modulación del diseño de un modelo que produzca la más alta exactitud al momento de clasificar moléculas OAB nuevas.

7.1 Validaciones cruzadas y simples.

Se formaron tres diferentes corpus fisicoquímicos a partir de la variación de los atributos, tal como se describe en la sección **¡Error! No se encuentra el origen de la referencia.**. I nicialmente se probó el corpus tipo A.

Corpus *A*:

Atributos (5)				
Solventes (3)			Moléculas (2)	
Interacciones dispersivas	Interacciones polares	Interacciones de H	# Carbonos en grupo éter	# Carbonos en grupo éster

Este modelo se evaluó cambiando varios criterios para la experimentación en el algoritmo kNN en Orange Canvas.

Inicialmente se presentan los resultados obtenidos con el corpus A. Se experimentó en el algoritmo kNN a través de la validación simple y la cruzada, con el fin de conocer los hiperparámetros en el algoritmo que arrojan una clasificación más exacta. Primero se presenta el reporte de la validación cruzada.

7.1.1 Corpus A, validación cruzada.

Para esta validación, se dividió el total de los datos del corpus en tres subconjuntos diferentes formados al azar (sección 4.3.1). Para la evaluación del algoritmo se utilizaron por separado cada uno de los tres subconjuntos como datos de prueba, mientras los otros dos subconjuntos fueron de entrenamiento. Se evaluaron 3 valores distintos de k, 3, 5 y 10. Fueron asignadas dos formas diferentes de dar peso a los atributos de los ejemplos; uniforme y por distancia (sección 3.6.1.2.1).

Durante esta evaluación, los resultados obtenidos son reflejo de lo que será el comportamiento del modelo predictivo al momento de someterlo a una instancia de prueba desconocida, para una clasificación basada en el conjunto de información con el que se entrena al algoritmo. Los resultados se presentan en la Tabla 25. **%**CA de clase con cruzada en kNN con corpus A. Peso: uniforme y distancia, métrica: Euclídea.

Tabla 25. %CA de clase con cruzada en kNN con corpus A. Peso: uniforme y distancia, métrica: Euclídea.

Valor de k	%CA de clase (G, S o P)						Promedio de 3 iteraciones (%)	
	Iteración 1		Iteración 2		Iteración 3			
	uniforme	distancia	uniforme	distancia	uniforme	distancia	uniforme	distancia
3	73	80	73	73	67	67	71.0	73.3
5	73	80	80	80	67	67	73.3	75.7
10	67	87	67	80	60	67	64.7	78.0

Se observa que existe variación entre los resultados de cada iteración. Con esto se infiere que la distribución de ejemplos distintas en los conjuntos de entrenamiento y de prueba tiene

influencia sobre la clasificación. Sin embargo, la iteración 1 y 2 conservan mayor similitud en sus resultados a comparación de la iteración 3.

Conforme se aumenta la cantidad de vecinos, existe una disminución en el porcentaje de coincidencia en las 3 iteraciones, con excepción de la iteración 2 donde se consigue el porcentaje máximo de esta corrida cuando k=5, este valor de vecinos hace que el promedio de las 3 iteraciones incremente, obteniendo el resultado más alto de 73.3 cuando $k=5$. Esto hace a los conjuntos de prueba y entrenamiento de la iteración 2 los más aptos a comparación de los de la 1 y 3, por lo que se analizará su contenido para dar una descripción más completa de esta observación en las siguientes secciones.

Para la siguiente prueba, el peso que se asigna a los ejemplos se determinó usando la distancia para la métrica Euclídea para que el algoritmo decida la clase de los ejemplos de prueba, en vez de darle peso uniforme a todos los atributos para una selección mediante regresión logística.

Los porcentajes de clasificación en las iteraciones 1 y 2 aumentaron usando la distancia Euclídea a comparación de cuando se usó un peso uniforme para los atributos. Esto sucede para los valores de vecinos de 3, 5 y 10 en la iteración 1 y para 3 y 10 en la iteración 2, el resultado cuando k=5 se mantiene igual.

La iteración 1 presenta los valores más altos de clasificación, conteniendo el porcentaje mayor de esta experimentación cuando k=10. Con este hecho se concluye que la medición de la distancia Euclídea entre los ejemplos permite formar un grupo similar más amplio con los valores de los atributos correspondientes a los ejemplos del conjunto de prueba de la iteración 1, incrementando la precisión de la clasificación cuando se aumenta la cantidad de vecinos a 10.

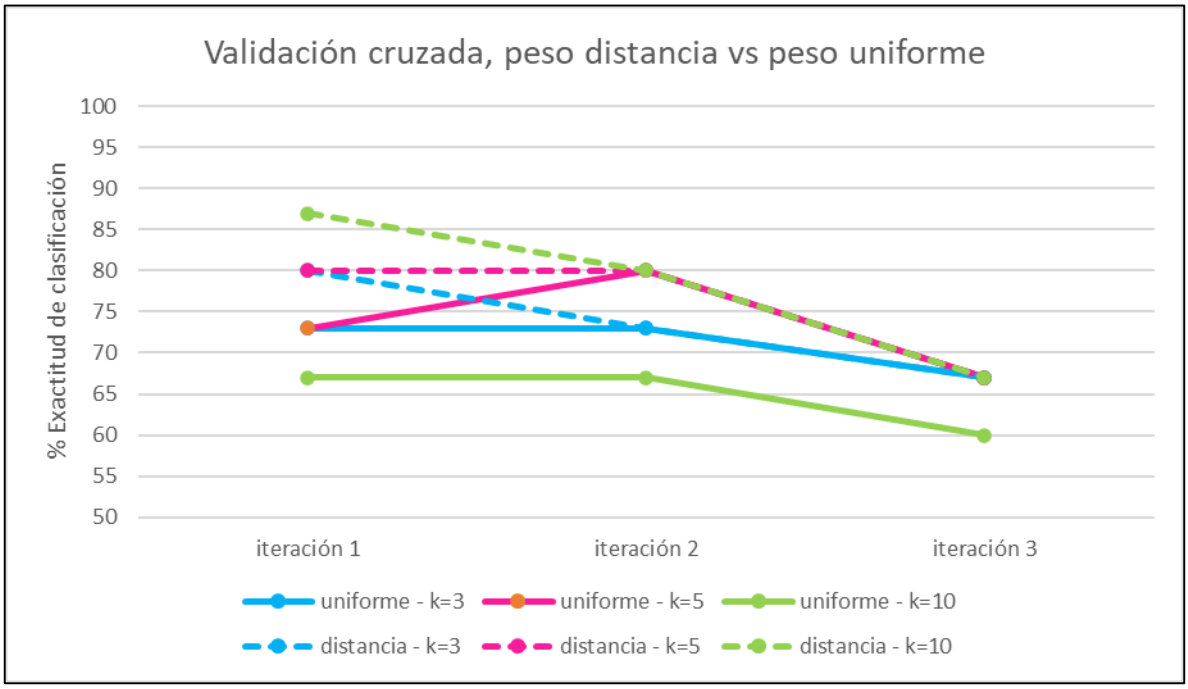


Figura 17. Comparación de % de Exactitud de clasificación para la validación cruzada de corpus A, distancia vs peso uniforme en la métrica Euclídea.

La Figura 17 presenta un comparativo de los resultados haciendo una validación simple al usar un peso uniforme en los ejemplos y asignando la distancia entre las mismas para aplicar en la métrica Euclídea. La distancia favorece los resultados en la clasificación de los ejemplos de prueba de la iteración 1, en tanto para los ejemplos 2 y 3 se obtuvieron los mismos valores.

Para complementar las observaciones, se complementa el estudio de este corpus desarrollando la validación simple, del que se reportan resultados en la siguiente sección.

7.1.2 Corpus A, validación simple.

En este caso, los conjuntos de entrenamiento de las 3 iteraciones conservaron los 45 ejemplos totales de origen, de tal forma que los ejemplos de los conjuntos de prueba de cada iteración se encuentren en su respectivo conjunto de entrenamiento, y de este modo, el algoritmo se entrene con los ejemplos con las que será probado. Tal como en la validación simple, se

evaluaron valores de k de 3, 5 y 10, y dos asignaciones de peso a los atributos; uniforme y distancia.

Para poder tener un panorama comparativo, la tabla Tabla 26 muestra los resultados de la validación simple para las dos formas de asignación de peso a los atributos.

Tabla 26. %CA de clase con validación simple en kNN con corpus A. Peso: uniforme y distancia, métrica: Euclídea.

Valor de k	%CA de clase (G, S o P)						Promedio de 3 iteraciones (%)	
	Iteración 1		Iteración 2		Iteración 3			
	uniforme	distancia	uniforme	distancia	uniforme	distancia	uniforme	distancia
3	86	100	93	100	66	100	81.6	100
5	73	100	80	100	73	100	75.3	100
10	80	100	86	100	73	100	79.6	100

La exactitud de clasificación aumenta a comparación de los resultados obtenidos con la validación cruzada, salvo 4 excepciones, con el peso uniforme de los ejemplos (Tabla 25). La iteración 2 repite siendo el conjunto que alcanza los valores más altos de precisión, siendo el mayor cuando k=3. Nuevamente la iteración 3 refleja los valores mínimos comparados con la 1 y la 2.

De igual forma, se experimentó con la distancia como parámetro de selección de la clase para los ejemplos de prueba. Se obtuvo una clasificación del 100% (Tabla 26). Esto es un indicio de un funcionamiento adecuado de los atributos que alimentan el corpus probado, al permitir un adecuado entrenamiento del algoritmo con ellos para una selección acertada de la clase de cada instancia mediante la distancia Euclídea.

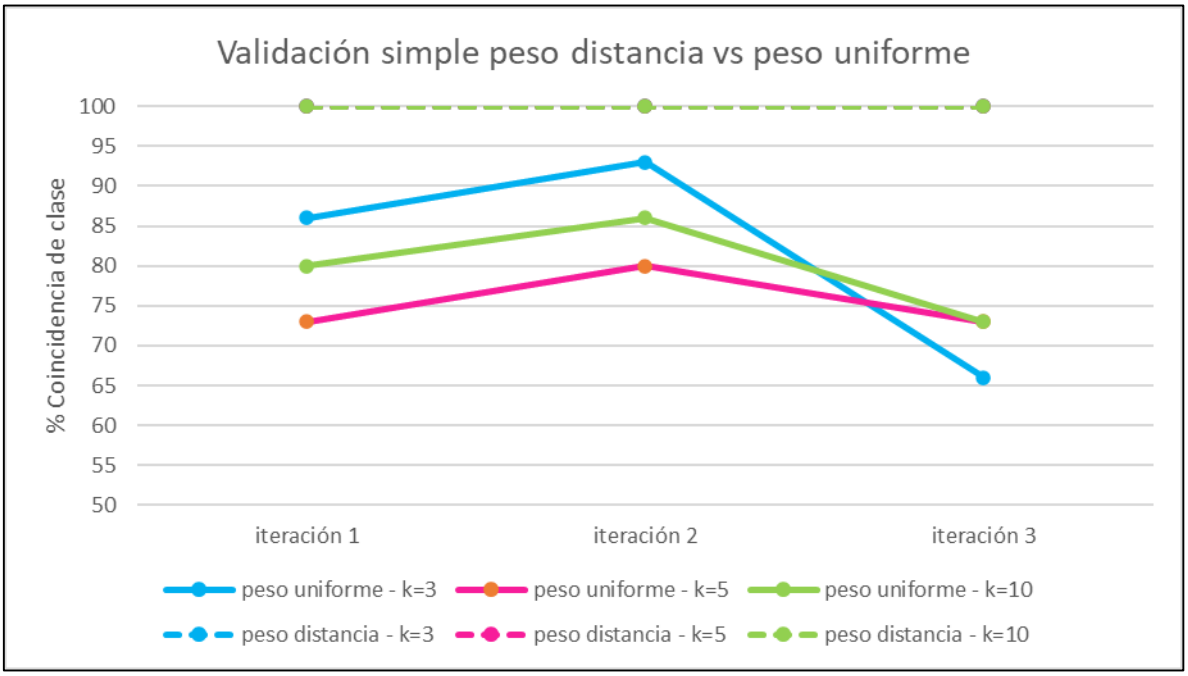


Figura 18. Comparativo del porcentaje de Exactitud de clasificación para la validación simple con corpus A, aplicando la distancia y el peso uniforme a los ejemplos en la métrica Euclídea.

En la Figura 18 se aprecia claramente la influencia en los resultados (contenidos en la Tabla 26) cuando se cambia el peso asignado a los ejemplos. Las líneas punteadas que se superponen pertenecen al 100 obtenido en las 3 iteraciones usando la distancia entre las ejemplos. Más abajo se tienen las líneas que representan los valores obtenidos a través de la asignación de un peso uniforme para cada instancia. Por lo que se concluye que el uso de la distancia con la métrica Euclídea favorece la clasificación.

Conociendo que la aplicación de la distancia a la métrica optimiza la clasificación, se hace un comparativo entre los resultados de la validación cruzada y la simple usando la distancia aplicada a la métrica Euclídea, esto con el fin de concluir la influencia del tipo de validación, que a su vez nos orienta a los detalles en el contenido de los conjuntos que mejoran el modelo.

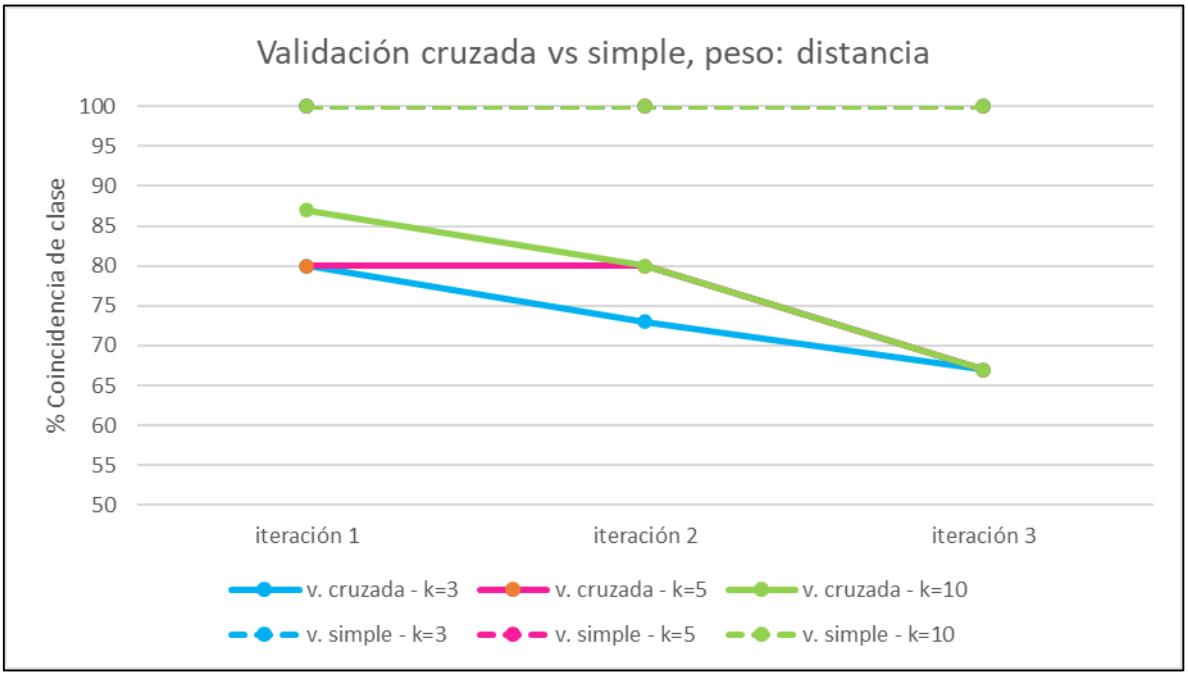


Figura 19. Comparativo del % de Exactitud de clasificación para la validación simple vs validación cruzada con corpus A, usando la distancia entre los ejemplos en la métrica Euclídea.

Con el uso de la distancia Euclídea en la validación simple se presentan los resultados óptimos, en tanto la validación cruzada refleja sus resultados por valores relativamente menores (Figura 19). Es importante tomar en cuenta que la validación cruzada es más fidedigna a lo que será una prueba de clasificación para el modelo al insertar una instancia desconocida, por lo que se necesitan los ajustes pertinentes que eleven la clasificación con este tipo de experimentación.

Con el fin de analizar la influencia de la robustez de los datos con los que se entrena al algoritmo, se modifica el corpus A, duplicando los ejemplos en el conjunto de entrenamiento y se evalúan los resultados, mismos que se presentan en la siguiente sección.

7.1.3 Corpus A, validación cruzada con dobles ejemplos de entrenamiento.

La experimentación llevada a cabo en el punto 7.1.1, se modifica duplicando los 30 ejemplos de entrenamiento de cada iteración, (60 ejemplos cada uno). Esto con el fin de evaluar la influencia de un entrenamiento con un conjunto de ejemplos más robusto. El resto de los parámetros se conserva igual. La Tabla 27 muestra un comparativo entre los resultados obtenidos con un conjunto de entrenamiento con los ejemplos duplicadas, y con los ejemplos únicos para la validación cruzada.

Tabla 27. Comparativo de %CA de clase, validación cruzada 30 ejemplos de entrenamiento únicas (azul) ejemplos dobles (rosa). Peso: distancia, métrica: Euclídea.

Valor de k	%CA de clase (G, S o P)					
	It1 - A	**It1 -doble**	**It2 - A**	**It2-doble**	**It3 - A**	**It3 -doble**
3	80	80	73	67	67	67
5	80	87	80	73	67	67
10	87	80	80	80	67	67

La Figura 20 presenta la visualización de la comparativa de resultados obtenidos con el entrenamiento de ejemplos únicas y de estas siendo duplicadas.

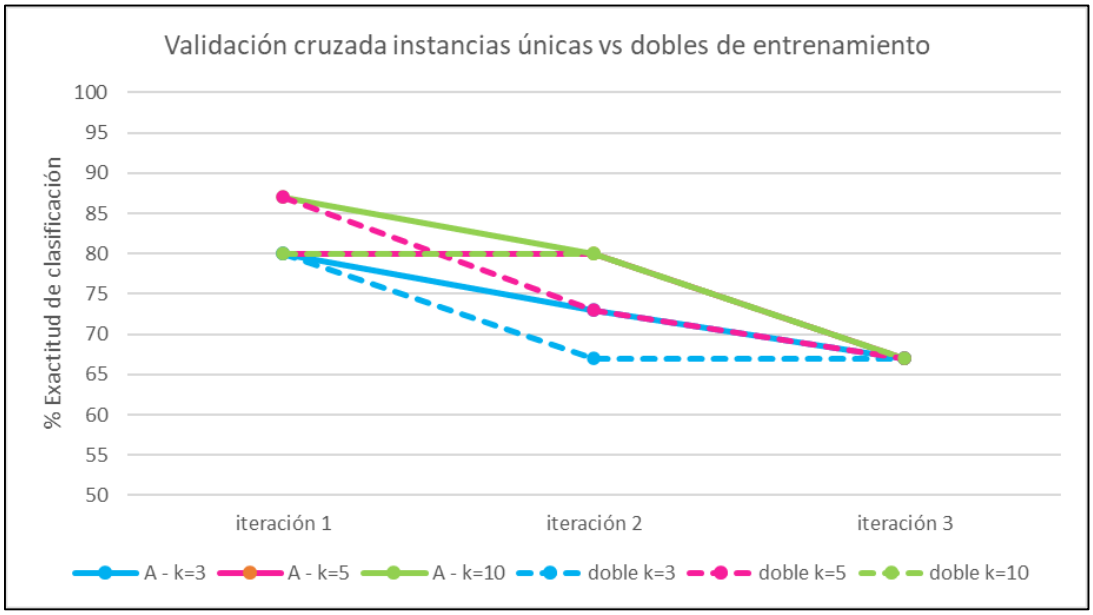


Figura 20. Comparativo del porcentaje de Exactitud de clasificación para la validación simple vs validación cruzada con corpus A, usando la distancia entre los ejemplos en la métrica Euclídea.

La Tabla 27 muestra los resultados de la validación cruzada cuando se duplican las ejemplos en el conjunto de entrenamiento (rosa) y un comparativo con los ejemplos únicos (azul). La iteración 3 produce resultados sin cambio. En tanto, la iteración 1 tiene su valor máximo cuando los ejemplos son dobles asignando una k=5. Si comparamos el resultado más alto obtenido con los ejemplos sin repetir, este se logra cuando k=10, por lo que se puede concluir que en este caso hay una ligera ventaja al aumentar la cantidad de ejemplos (7% más), debido a que localizan en un radio más cercano las de la clase correcta, reflejo de un valor de k menor. En la iteración 2, por el contrario, se tienen resultados ligeramente menores cuando se duplican los ejemplos en los casos de valores de k=3 y 5.

Basados en estas observaciones, podemos concluir que la duplicación de los mismos ejemplos dentro del conjunto de entrenamiento no aporta a un aumento significativo de la precisión clasificatoria. El siguiente paso a consideración es el análisis del contenido de los conjuntos, de acuerdo con las clases presentes en cada uno de ellos, su cantidad y homogeneidad.

Para la formación del corpus B se calcularon previamente los HSP de las moléculas 1_{12}, 3_{12} y 4_{12} (sección **¡Error! No se encuentra el origen de la referencia.**) y estos valores se a gregaron al corpus A. Se presenta a continuación los resultados de la validación cruzada con este corpus nuevo.

7.1.4 Corpus B, validación cruzada.

Con el antecedente de las diferencias obtenidas al cambiar el peso asignado a los ejemplos, se experimentó con las dos posibilidades en esta variable, con el peso uniforme y con la distancia entre los ejemplos. Los resultados se presentan en la Tabla 28.

Tabla 28. %CA de clase con cruzada en kNN con corpus B. Peso: uniforme y distancia, métrica: Euclídea.

Valor de k	%CA de clase (G, S o P)						Promedio de 3 iteraciones (%)	
	Iteración 1		Iteración 2		Iteración 3			
	uniforme	distancia	uniforme	distancia	uniforme	distancia	uniforme	distancia
3	73	87	60	80	67	67	73.3	71.3
5	73	87	73	80	67	67	73.3	75.6
10	67	87	73	67	53	67	62.3	75.6

Como anteriormente ocurrió durante la validación cruzada usando el corpus A, los resultados cuando se asigna un peso uniforme a los ejemplos fueron menores que cuando el peso es la distancia en la ecuación Euclídea.

Se compararon los resultados de la Tabla 28 cuando el peso de los atributos es la distancia usando el corpus B vs los obtenidos con el corpus A con el mismo hiperparámetro (Tabla 25) bajo los mismos estatutos de los hiperparámetros en el algoritmo. Se tiene que hubo un aumento en la exactitud de clasificación en la iteración 1 (7% más), en tanto la iteración 2 disminuye sus resultados y la 3 permanece igual. Lo que nos orienta a la misma conclusión obtenida durante el comparativo de la validación cruzada con el corpus A entre el peso uniforme y la distancia, es decir, los valores contenidos en los ejemplos de la iteración 1

favorecen la clasificación cuando es usada la distancia Euclídea, suceso que permanece al modificar el corpus agregando los HSP de las moléculas (en el corpus B).

Para evaluar la efectividad de los atributos presentes en ambos corpus, se experimentó con un corpus B donde se adicionan los HSP de las moléculas, pero se respeta la permanencia de los atributos correspondientes a la cantidad de carbonos en la parte éter y éster de estas, denominándose corpus C. Los resultados se encuentran en la tabla Tabla 29.

Tabla 29. %CA de clase con validación cruzada en kNN con corpus C. Peso: distancia, métrica: Euclídea.

Valor de k	**%CA de clase (G, S o P)**			**Promedio de 3 iteraciones (%)**
	Iteración 1	**Iteración 2**	**Iteración 3**	
3	80	73	67	73.3
5	87	80	67	78.0
10	87	80	67	78.0

Para conocer con mayor claridad la aportación de los atributos correspondientes a los HSP de los oligómeros, comparamos los resultados de esta experimentación con los obtenidos con la validación cruzada con el corpus A (Tabla 25) mismo que contiene en sus atributos los atributos éter y éster. Los resultados son iguales, salvo en la iteración 1 cuando el valor de k=5, que sube un 7% con el corpus C. Esta mínima diferencia puede ser el reflejo de la contribución de los HSP de los oligómeros con la configuración manejada en esta experimentación. Son requeridas pruebas que fundamenten la selección de los atributos necesarios para optimizar la clasificación sin tener un over fitting o un under fitting (sobre o bajo entrenamiento).

7.1.5 Corpus C, validación simple.

Se llevó a cabo la validación simple para este nuevo corpus, se tomó como referencia las experimentaciones anteriores para asignar la distancia como peso para los atributos. Los resultados se muestran en la tabla Tabla 30.

Tabla 30. %CA de clase con validación simple en kNN con corpus C. Peso: distancia, métrica: Euclídea.

Valor de k	%CA de clase (G, S o P)			Promedio de 3 iteraciones (%)
	Iteración 1	Iteración 2	Iteración 3	
3	100	100	100	100
5	100	100	93	97.7
10	100	100	93	97.7

Los resultados obtenidos presentan similitud a los producidos durante la validación simple con el corpus A (Tabla 26). Se encontraron dos excepciones durante este comparativo: en la validación simple con el corpus C, la iteración 3 con k=3 y k=5, ambos producen un 93% de exactitud de clase, cuando anteriormente se produjo el 100% con el corpus A, por lo que existe una disminución del 3%. Esto puede deberse a que el corpus posee una sobreinformación, y que los valores agregados estén causando ruido e impidan la correcta clasificación al aumentar el radio de ejemplos similares a la de prueba cuando aumentamos la cantidad de vecinos en los ejemplos de la iteración 3.

A continuación, se mencionan las conclusiones a las que se llegó a partir de los resultados obtenidos y la evaluación comparativa de criterios. Se presentan en la siguiente sección los resultados del análisis por componentes de los conjuntos de entrenamiento y prueba por estructura de cada corpus.

7.2 Análisis de la composición de los conjuntos de entrenamiento y prueba de los corpus A y B correspondientes a las 3 iteraciones de la validación cruzada.

Para la formación de los conjuntos utilizados en los corpus A y B, inicialmente los 45 ejemplos de la base de datos total se mezclaron de manera homogénea manualmente, y se procedió a dividirlas en 3 diferentes subconjuntos de entrenamiento, y 3 diferentes subconjuntos de prueba, y con estos desarrollar una validación cruzada. Cada uno de los

subconjuntos de entrenamiento de cada iteración se compone de un total de 30 ejemplos. En tanto cada subconjunto de prueba contiene 15 ejemplos en total.

Las Tabla 31Tabla 32 yTabla **33** muestran la distribución por cantidad de ejemplos para cada atributo, en cada conjunto en las iteraciones 1, 2 y 3 respectivamente. Esto con la finalidad de saber la homogeneidad en la información, y relacionarlo con los resultados obtenidos con cada subconjunto durante la predicción. Los nombres correspondientes a la simbología representando los solventes, se presenta en la sección de nomenclatura.

Tabla 31. Distribución de los ejemplos en los conjuntos de entrenamiento y prueba de la iteración 1.

	Clase			Éster (Carbonos)			Solvente														
	G	P	S	1	3	4	HE	TO	ADE	CCH	THF	ACN	ACE	DMF	ET	DMS	MET	HEX	PEN	CL	ISOP
Iteración 1	Conjunto de entrenamiento																				
	14	1	15	8	10	12	2	2	2	1	2	3	2	2	3	2	2	1	2	2	2
	Conjunto de prueba																				
	6	1	8	7	5	3	1	1	1	2	1	0	1	1	0	1	1	1	1	1	1

Tabla 32. Distribución de los ejemplos en los conjuntos de entrenamiento y prueba de la iteración 2.

	Clase			Éster (Carbonos)			Solvente														
	G	P	S	1	3	4	HE	TO	ADE	CCH	THF	ACN	ACE	DMF	ET	DMS	MET	HEX	PEN	CL	ISOP
Iteración 2	Conjunto de entrenamiento																				
	12	2	16	13	7	10	1	2	2	2	2	2	2	2	1	2	2	2	2	3	3
	Conjunto de prueba																				
	8	0	7	2	8	5	2	1	1	1	1	1	1	1	2	1	1	1	1	0	0

Tabla 33. Distribución de los ejemplos en los conjuntos de entrenamiento y prueba de la iteración 3.

	Clase			Éster (Carbonos)			Solvente														
	G	P	S	1	3	4	HE	TO	ADE	CCH	THF	ACN	ACE	DMF	ET	DMS	MET	HEX	PEN	CL	ISOP
Iteración 3	Conjunto de entrenamiento																				
	14	1	15	9	13	8	3	2	2	3	2	1	2	2	2	2	2	3	2	1	1
	Conjunto de prueba																				
	6	1	8	6	2	7	0	1	1	0	1	2	1	1	1	1	1	0	1	2	2

Se analizó cada uno de los componentes que forman los corpus. Se contabilizó la presencia de cada una de las clases en los conjuntos para saber si existe homogeneidad o alguna de ellas tiene mayor presencia. La Figura 21 muestra la distribución de los ejemplos según cada clase posible (G, P o S) en los conjuntos de prueba y de entrenamiento.

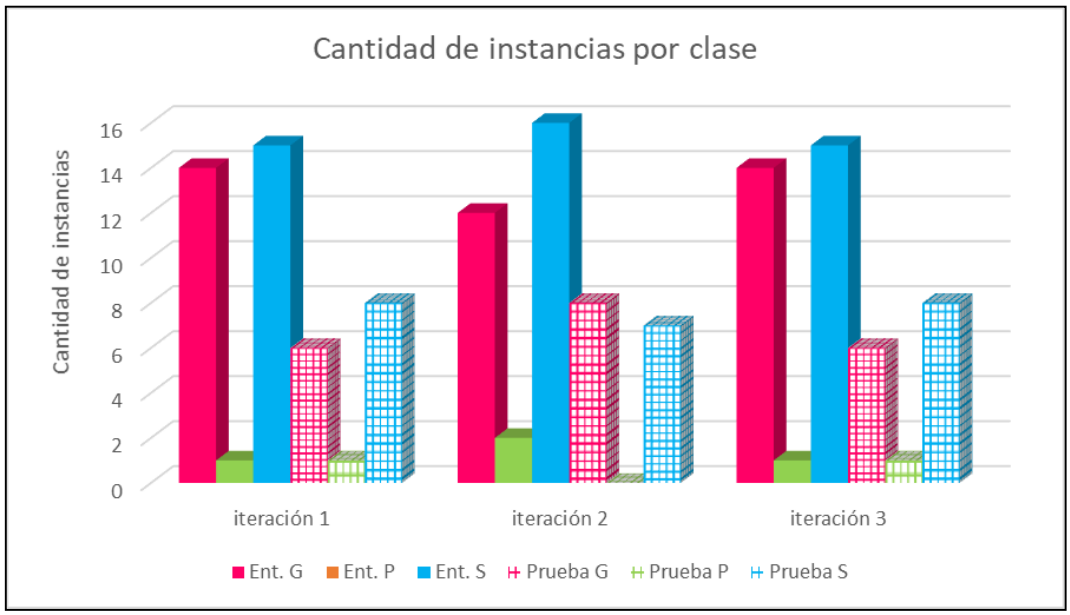


Figura 21. Distribución de cantidad de ejemplos por clase (G, P y S).

Se observa un predominio de la clase S en el conjunto de entrenamiento a comparación de las clases G y P, siendo esta última la de menor presencia en las iteraciones 1 y 3, y nula en la 2. Existe similitud en la distribución de clases de los ejemplos en las iteraciones 1 y 3, sin embargo, los resultados de la exactitud de clasificación son más altos en la iteración 1.

Parte de este análisis gira en torno al alcance en la exactitud de la clasificación que se logra con los diferentes corpus. En la sección continua se explica la búsqueda de un modelo predictivo con una mayor precisión clasificatoria.

7.3 Relevancia de la exactitud de clasificación.

Como se observó, la caracterización de los solventes y los oligómeros mediante los parámetros de solubilidad de Hansen aportan información al algoritmo kNN que le permite una clasificación mayor que cuando se caracterizan estos componentes con valores asignados para describirlos cualitativamente, pero que no se relacionan con su comportamiento fisicoquímico, como se hizo en el apartado de experimentación con corpus cualitativo.

Cabe resaltar, que es importante la optimización de los componentes del modelo predictivo. Es decir, tanto el corpus de información, que contenga la información puntual, sin exceso que aporte ruido ni una caracterización incompleta que impida que el algoritmo haga el proceso de clasificación correctamente, como la selección del algoritmo que haga un "match" adecuado con el corpus.

Hasta esta experimentación, se ha logrado el 100% en la clasificación cuando sometemos al algoritmo a ejemplos de prueba ya conocidas por él, al ser parte del conjunto de entrenamiento, sin embargo, la clasificación que se logra durante la validación cruzada es lo más similar a lo que será una predicción con una instancia donde le preguntamos al algoritmo que producto se podrá obtener con un nuevo solvente propuesto o una nueva molécula diseñada. De aquí la relevancia del ajuste de las variables involucradas en el modelo.

Como se comentó previamente en la sección 4.3.2, Uno de los factores fundamentales en la evaluación de un modelo de predicción es probar con un conjunto de elementos nunca antes visto por el algoritmo, a modo de corroborar su precisión [5], por lo que, en la siguiente sección se incluye la experimentación definitiva en la cual se someten dos moléculas no probadas antes a clasificación con las distintas configuraciones previamente evaluadas de los modelos, para así, determinar cuál de ellas es la que produce una mayor exactitud al clasificar, y comprobar qué hiperparámetros tienen mayor relevancia en su diseño.

7.4 Prueba de Clasificación de las moléculas 1_{14} y 2_{14}.

Se formaron dos conjuntos de prueba base a partir de la caracterización con atributos de dos moléculas previamente diseñadas y probadas durante experimentación, parte de la misma familia de OABs, con una variación en la longitud de sus cadenas éter y éster, tal como se describe en la sección 5.3.2.3. A continuación, se detallan los resultados de cada una de las configuraciones sometidas a Prueba mediante la clasificación de estos OABs novel.

7.4.1 Prueba con conjuntos de entrenamiento con clases No distribuidas.

Para observar los resultados, se presentan dos tablas con las clases obtenidas durante las pruebas de clasificación, una para la molécula 1_{14} (Tabla 34) y otra para la 2_{14} (Tabla 35). Estas tablas a su vez se dividen en dos secciones, una dónde los ejemplos del conjunto de entrenamiento los OAB se presentan a concentración constante, y otra en la que se manejan a concentración variable.

Nota: En todas las tablas con resultados, se encuentran marcadas con gris las celdas que pertenecen a los ejemplos clasificados erróneamente

Tabla 34. Resultados de clasificación para conjunto de Prueba con OAB 1_{14}, conjuntos de entrenamiento clases No distribuidas. Configuración de kNN: métrica Euclídea, k=5, peso uniforme.

Solvente	1_{14} Conc. (%v/v)	Conjunto de entrenamiento de Concentración VARIABLE			Conjunto de entrenamiento de Concentración CONSTANTE		
		A	B	C	A	B	C
Hexano	10	YES	NO	YES	NO	NO	YES
Ciclohexano		YES	NO	YES	NO	NO	YES
Heptano		YES	NO	YES	NO	NO	YES
Acetato de etilo		NO	NO	NO	NO	NO	NO
Acetona		NO	NO	NO	NO	NO	NO
THF		NO	NO	NO	NO	NO	NO
Tolueno		NO	NO	NO	NO	NO	NO
DMS		YES	YES	YES	NO	YES	YES
Cloroformo		NO	NO	NO	NO	NO	NO
Isopropanol		YES	NO	YES	NO	NO	YES
DMF		NO	NO	YES	NO	NO	YES
Metanol		YES	YES	YES	NO	YES	YES
Pentano		YES	NO	YES	NO	NO	YES
Acetonitrilo		NO	NO	NO	NO	NO	NO
Etanol		YES	NO	YES	NO	NO	YES
% CA para ejemplos de concentración constante		**73**	**60**	**80**	**46**	**60**	**80**

Tabla 35. Resultados Prueba con OAB 2_{14}, Conjuntos de entrenamiento clases No distribuidas. Configuración de kNN: métrica Euclídea, k=5, peso uniforme.

Solvente	2_{14} Conc. (%v/v)	Conjunto de entrenamiento de Concentración VARIABLE			Conjunto de entrenamiento de Concentración CONSTANTE		
		A	B	C	A	B	C
Hexano	10	NO	NO	NO	NO	NO	NO
Ciclohexano		NO	NO	NO	NO	NO	NO
Heptano		NO	NO	NO	NO	NO	NO
Acetato de etilo		NO	NO	NO	NO	NO	NO
Acetona		NO	NO	NO	NO	NO	NO
THF		NO	NO	NO	NO	NO	NO
Tolueno		NO	NO	NO	NO	NO	NO
DMS		YES	YES	YES	YES	YES	YES
Cloroformo		NO	NO	NO	NO	NO	NO
Isopropanol		YES	NO	NO	YES	NO	NO
DMF		YES	NO	YES	YES	NO	YES
Metanol		YES	YES	YES	YES	YES	YES
Pentano		NO	NO	NO	NO	NO	NO
Acetonitrilo		NO	NO	NO	NO	NO	NO
Etanol		YES	NO	NO	YES	NO	NO
% CA para ejemplos de concentración		**93**	**73**	**80**	**93**	**73**	**80**

7.4.2 Prueba con conjuntos de entrenamiento con clases distribuidas.

Las siguientes tablas contienen los resultados de la clasificación de las moléculas 1_{14} (Tabla 36) y 2_{14} (Tabla 37). De manera similar a la sección previa, las tablas están subdivididas de acuerdo con los conjuntos de entrenamiento que contienen ejemplos a concentraciones constante y variables de OAB.

Tabla 36. Resultados de clasificación para conjuntos de Prueba con OAB 1_{14}, Conjunto de entrenamiento con clases distribuidas. Configuración de kNN: métrica Euclídea, k=5, peso uniforme.

Solvente	1_{14} Conc. (%v/v)	Conjunto de entrenamiento de Concentración VARIABLE			Conjunto de entrenamiento de Concentración CONSTANTE		
		A	B	C	A	B	C
Hexano	10	YES	NO	YES	YES	NO	YES
Ciclohexano		YES	NO	YES	YES	NO	YES
Heptano		YES	NO	YES	YES	NO	YES
Acetato de etilo		NO	NO	NO	NO	NO	NO
Acetona		NO	NO	NO	NO	NO	NO
THF		NO	NO	NO	NO	NO	NO
Tolueno		NO	NO	NO	NO	NO	NO
DMS		YES	YES	YES	YES	YES	YES
Cloroformo		NO	NO	NO	NO	NO	NO
Isopropanol		YES	YES	YES	YES	NO	YES
DMF		YES	NO	YES	YES	NO	YES
Metanol		YES	YES	YES	YES	YES	YES
Pentano		YES	NO	YES	YES	NO	YES
Acetonitrilo		YES	YES	YES	YES	YES	YES
Etanol		YES	YES	YES	YES	YES	YES
% CA para ejemplos de concentración constante		**80**	**66**	**86**	**87**	**73**	**87**

Tabla 37. Resultados Prueba con OAB 2_{14}, Conjuntos de entrenamiento con clases distribuidas. Configuración de kNN: métrica Euclídea, k=5, peso uniforme.

Solvente	2_{14} Conc. (%v/v	Concentración VARIABLE			Concentración CONSTANTE		
		A	B	C	A	B	C
Hexano	10	YES	NO	NO	NO	NO	NO
Ciclohexano		NO	NO	NO	NO	NO	NO
Heptano		YES	NO	NO	YES	NO	NO
Acetato de etilo		NO	NO	NO	NO	NO	NO
Acetona		NO	NO	NO	NO	NO	NO
THF		NO	NO	NO	NO	NO	NO
Tolueno		NO	NO	NO	NO	NO	NO
DMS		YES	YES	YES	YES	YES	YES
Cloroformo		NO	NO	NO	NO	NO	NO
Isopropanol		YES	YES	YES	YES	YES	YES
DMF		YES	NO	YES	YES	NO	YES
Metanol		YES	YES	YES	YES	YES	YES
Pentano		YES	NO	NO	YES	NO	NO
Acetonitrilo		YES	YES	YES	YES	YES	YES
Etanol		YES	YES	YES	YES	YES	YES
% CA para ejemplos de concentración constante		**80**	**93**	**100**	**87**	**93**	**100**

De manera general, se observa en las Tablas 36 y 37, que en comparación con los resultados de clasificación con los conjuntos de entrenamiento con clases no distribuidas (Tablas 34 y

35) se tiene un aumento en la exactitud de la clasificación. La mejora en el desempeño se adjudica en este caso a la distribución uniforme de los ejemplos por su clase en los conjuntos de entrenamiento, lo que contribuye a una disminución del error en la selección de los vecinos más cercanos que pudieran pertenecer a una clase no correspondiente por ser predominante en el conjunto.

Ahora bien, centrándonos en los conjuntos de entrenamiento que producen un desempeño más alto, los resultados observados en la clasificación de las moléculas 1_{14} y 2_{14}, se presentan con una clara diferencia en ambas moléculas (Tablas 36 y 37), observándose la más alta clasificación al etiquetar a la molécula 2_{14}, logrando hasta el 100% de CA.

Debido a que los conjuntos de entrenamiento se vieron robustecidos con ejemplos, se evaluó la clasificación asignando un número de vecinos *k* más reducido para eliminar la posibilidad de una selección errónea de vecinos debido a ruido. Se usaron los mismos conjuntos con clases distribuidas y se aplicó la misma configuración a kNN variando sólo el valor de *k=2*. Los resultados se muestran en las **¡Error! No se encuentra el origen de la referencia.**38, p ara la molécula 1_{14}, y 39 para la 2_{14}.

Tabla 38. Resultados de clasificación para conjuntos de Prueba con OAB 1_{14}, Conjunto de entrenamiento con clases distribuidas. Configuración de kNN: métrica Euclídea, k=2, peso uniforme.

Solvente	1_{14} Conc. (%v/v	Conjunto de entrenamiento de Concentración VARIABLE			Conjunto de entrenamiento de Concentración CONSTANTE		
		A	B	C	A	B	C
Hexano	10	YES	NO	YES	YES	NO	YES
Ciclohexano		NO	NO	NO	NO	NO	NO
Heptano		YES	NO	YES	YES	NO	YES
Acetato de etilo		NO	NO	NO	NO	NO	NO
Acetona		NO	NO	NO	NO	NO	NO
THF		NO	NO	NO	NO	NO	NO
Tolueno		NO	NO	NO	NO	NO	NO
DMS		YES	YES	YES	YES	YES	YES
Cloroformo		NO	NO	NO	NO	NO	NO
Isopropanol		YES	YES	YES	YES	YES	YES
DMF		YES	NO	YES	YES	NO	YES
Metanol		YES	YES	YES	YES	YES	YES
Pentano		YES	NO	YES	YES	NO	YES
Acetonitrilo		YES	YES	YES	YES	YES	YES
Etanol		YES	YES	YES	YES	YES	YES

% CA para ejemplos de concentración constante	93	67	93	93	67	93

Tabla 39. Resultados Prueba con OAB 2_{14}, Conjuntos de entrenamiento con clases distribuidas. Configuración de kNN: métrica Euclídea, k=2, peso uniforme.

Solvente	2_{14} Conc. (%v/v	Concentración VARIABLE			Concentración CONSTANTE		
		A	B	C	A	B	C
Hexano	10	YES	NO	NO	NO	NO	NO
Ciclohexano		NO	NO	NO	NO	NO	NO
Heptano		YES	NO	NO	NO	NO	NO
Acetato de etilo		NO	NO	NO	NO	NO	NO
Acetona		NO	NO	NO	NO	NO	NO
THF		NO	NO	NO	NO	NO	NO
Tolueno		NO	NO	NO	NO	NO	NO
DMS		YES	YES	YES	YES	YES	YES
Cloroformo		NO	NO	NO	NO	NO	NO
Isopropanol		YES	NO	YES	YES	NO	YES
DMF		YES	NO	NO	NO	NO	NO
Metanol		YES	YES	YES	YES	YES	YES
Pentano		YES	NO	NO	NO	NO	NO
Acetonitrilo		YES	NO	YES	YES	NO	YES
Etanol		YES	NO	YES	YES	NO	YES
% CA para ejemplos de concentración constante		**80**	**73**	**93**	**93**	**73**	**93**

De acuerdo con los valores de clasificación mostrados en las tablas 38 y 39, se tiene una ligera disminución en los resultados de la molécula 2_{14} cuando se aplica la estructura de corpus B y C. Por otra parte, la clasificación de la molécula 1_{14} se elevó con los tres tipos de estructuras A, B y C. Las siguientes Tablas 40 y 41 muestran un concentrado de los resultados de clasificación para ambas moléculas.

Tabla 40. Resultados de %CA para las moléculas 1_{14} y 2_{14} bajo 3 distintas composiciones de conjuntos y 5 tipos de estructuras de corpus, a concentración constante.

	%CA					
Corpus	Molécula 1_{14}			Molécula 2_{14}		
	Conjuntos clases No distribuidas k=5	Conjuntos clases distribuidas k=5	Conjuntos clases distribuidas k=2	Conjuntos clases No distribuidas k=5	Conjuntos clases distribuidas k=5	Conjuntos clases distribuidas k=2
A	46	87	93	93	87	93
B	60	73	67	73	93	73
C	80	87	93	80	100	93

Tabla 41. Resultados de %CA para las moléculas 1_{14} y 2_{14} bajo 3 distintas composiciones de conjuntos y 5 tipos de estructuras de corpus, a concentración variable.

	%CA					
Corpus	**Molécula 1_{14}**			**Molécula 2_{14}**		
	Conjuntos clases No distribuidas k=5	**Conjuntos clases distribuidas k=5**	**Conjuntos clases distribuidas k=2**	**Conjuntos clases No distribuidas k=5**	**Conjuntos clases distribuidas k=5**	**Conjuntos clases distribuidas k=2**
A	73	80	93	93	80	80
B	60	66	67	73	93	73
C	80	86	93	80	100	93

Hasta este punto, se puede concluir qué, la molécula 2_{14} logra un porcentaje de clasificación correcto por encima del logrado con la molécula 1_{14}. El conjunto de entrenamiento con mejor desempeño clasificatorio es cuando se tiene un conjunto con clases de ejemplos distribuidas a concentración constante con k=5, con la estructura de corpus C. Sin embargo, esta estructura no es logra la mayor exactitud en los dos OAB´s clasificados, por lo que, se sabe entonces que existe una codependencia entre el conjunto de entrenamiento y el de prueba para lograr una clasificación exacta.

En cuanto al análisis del contenido de los conjuntos de acuerdo con los atributos de caracterización, se tiene que, los Corpus A y C cuentan con los atributos Éter y Éster, combinandose siempre juntos. Sin embargo, no se conoce experimentalmente si es la presencia de uno de los dos o ambos lo que eleva la clasificación, por lo que, se corrobora desarrollándose una prueba. Se formaron conjuntos por separado con el tipo de estructura B, y conteniendo cada uno de forma individual el atributo Éster y luego el Éter. Los resultados de esta experimentación se muestran en la Tabla 42.

Tabla 42. Resultados de %CA para las moléculas 1_{14} y 2_{14} bajo 3 distintas composiciones de conjuntos y 5 tipos de estructuras de corpus, a concentración variable.

	%CA					
	Molécula 1_{14}			Molécula 2_{14}		
Corpus B	**Conjuntos clases No distribuidas k=5**	**Conjuntos clases distribuidas k=5**	**Conjuntos clases distribuidas k=2**	**Conjuntos clases No distribuidas k=5**	**Conjuntos clases distribuidas k=5**	**Conjuntos clases distribuidas k=2**
B	60	73	67	73	93	73
+Éter	60	73	67	80	100	73
+Éster	80	87	93	73	100	93

Como se observa en la Tabla 42, en 5 casos la clasificación aumentó cuando se agregó el atributo Éster, y solamente dos cuando se añadió el atributo Éter. Por lo tanto, se puede concluir que es el atributo Éster el que contribuye al aumento en la clasificación exacta por parte de un modelo.

Para un análisis más profundo con respecto a la influencia de cada atributo al momento de clasificar, se presentan los análisis de varianza (ANOVA) aplicados a cada estructura de corpus evaluada.

7.4.3 ANOVAS de conjuntos de entrenamiento.

De acuerdo con cada una de las diferentes combinaciones de corpus, según la concentración y la distribución de las clases, las siguientes tablas contienen los resultados de los análisis de varianza practicados a cada configuración, y por cada estructura de corpus, con el fin de saber la influencia de los atributos con relación a cada configuración.

Tabla 43. Resultados de análisis de varianza para cada estructura de corpus, con valores críticos F.

Distribución de clases: No distribuidas									
Concentración: Constante									
Estructura									
A	**F**	**B**	**F**	**C**	**F**	**B+éter**	**F**	**B+éster**	**F**
Éter	18.60	S. H	16.55	Éter	18.60	Éter	18.60	S. H	16.55
S. H	16.55	S. polares	15.49	S. H	16.55	S. H	16.55	S. polares	15.49
S. polares	15.49	M. dispersivas	2.76	S. polares	15.49	S. polares	15.49	Éster	9.23
Éster	9.23	M. polares	2.63	Éster	9.23	M. dispersivas	2.76	M. dispersivas	2.76
S. dispersivas	0.10	M. H	2.48	M. dispersivas	2.76	M. polares	2.63	M. polares	2.63
		S. dispersivas	0.10	M. Polares	2.63	M. H	2.48	M. H	2.48
				M. H	2.48	S. dispersivas	0.10	S. dispersivas	0.10
				S. Dispersivas	0.10				

Tabla 44. Resultados de análisis de varianza para cada estructura de corpus, con valores críticos F.

Distribución de clases: No distribuidas									
Concentración: Variable									
Estructura									
A	**F**	**B**	**F**	**C**	**F**	**B+éter**	**F**	**B+éster**	**F**
Éter	39.62	S. H	29.17	Éter	39.62	Éter	39.62	S. H	29.17
S. H	29.17	S. polares	24.72	S. H	29.17	S. H	29.17	S. polares	24.72
S. polares	24.72	M. dispersivas	4.18	S. polares	24.72	S. polares	24.72	Éster	23.61
Éster	23.61	M. polares	4.55	Éster	23.61	M. dispersivas	4.81	M. dispersivas	4.81
Concentración	0.23	M. H	4.26	M. dispersivas	4.18	M. polares	4.55	M. polares	4.55
S. dispersivas	0.001	Concentración	0.23	M. polares	4.55	M. H	4.26	M. H	4.26
		S. dispersivas	0.001	M. H	4.26	Concentración	0.23	Concentración	0.23
				Concentración	0.23	S. dispersivas	0.001	S. dispersivas	0.001
				S. dispersivas	0.001				

Tabla 45. Resultados de análisis de varianza para cada estructura de corpus, con valores críticos F.

Distribución de clases: Distribuidas									
Concentración: Constante									
Estructura									
A	**F**	**B**	**F**	**C**	**F**	**B+éter**	**F**	**B+éster**	**F**
Éter	47.82	S. H	31.45	Éter	47.82	Éter	47.82	S. H	31.45
S. H	31.45	S. polares	31.20	S. H	31.45	S. H	31.45	S. polares	31.20
S. polares	31.20	M. dispersivas	7.10	S. polares	31.20	S. polares	31.20	Éster	21.54
Éster	21.54	M. polares	6.76	Éster	21.54	M. dispersivas	7.10	M. dispersivas	7.10
S. dispersivas	0.13	M. H	6.31	M. dispersivas	7.10	M. polares	6.76	M. polares	6.76
		S. dispersivas	0.13	M. polares	6.76	M. H	6.31	M. H	6.31
				M. H	6.31	S. dispersivas	0.13	S. dispersivas	0.13
				S. dispersivas	0.13				

Tabla 46. Resultados de análisis de varianza para cada estructura de corpus, con valores críticos F.

Distribución de clases: Distribuidas									
Concentración Variable									
Estructura									
A	**F**	**B**	**F**	**C**	**F**	**B+éter**	**F**	**B+éster**	**F**
Éter	102.90	S. H	59.33	Éter	102.90	Éter	102.90	S. H	59.33
S. H	59.33	S. polares	52.19	S. H	59.33	S. H	59.33	Éster	55.88
Éster	55.88	M. dispersivas	12.07	Éster	55.88	S. polares	52.19	S. polares	52.19
S. polares	52.19	M. polares	11.33	S. polares	52.19	M. dispersivas	12.07	M. dispersivas	12.07
Concentración	0.64	M. H	10.52	M. dispersivas	12.07	M. polares	11.33	M. polares	11.33
S. dispersivas	0.001	Concentración	0.64	M. polares	11.33	M. H	10.52	M. H	10.52
		S. dispersivas	0.001	M. H	10.52	Concentración	0.64	Concentración	0.64
				Concentración	0.64	S. dispersivas	0.001	S. dispersivas	0.001
				S. dispersivas	0.001				

En los ANOVA desarrollados para cada estructura de corpus en cada conjunto, el valor *F* de cada atributo es constante, esto debido a que la columna de datos en ese conjunto es la misma. Por esto, se presenta en la Tabla 47 los atributos en orden descendiente de acuerdo con su valor *F* de varianza en cada uno de los casos de las composiciones de los conjuntos de ejemplos.

Tabla 47. Atributos ordenados de acuerdo con su valor crítico *F* en cada configuración de conjuntos de entrenamiento

Atributo	**Configuración**			
	Clases No distribuidas / Concentración Constante 135 ejemplos	Clases No distribuidas / Concentración Variable 270 ejemplos	Clases distribuidas / Concentración Constante 186 ejemplos	Clases distribuidas / Concentración Variable 366 ejemplos
	Valores de F			
Éter	18.60	39.62	47.82	102.90
S. H	16.55	29.17	31.45	59.33
S. polares	15.49	24.72	31.20	52.19
Éster	9.23	23.61	21.54	55.88
M. dispersivas	2.76	4.18	7.10	12.07
M. polares	2.63	4.55	6.76	11.33
M. H	2.48	4.26	6.31	10.52
Concentración	-	0.23	-	0.64
S. dispersivas	0.10	0.001	0.13	0.001
%CA máxima alcanzada	**93**	**93**	**100**	**100**

Se presentan de manera visual a través de un gráfico, cada valor crítico F, para cada atributo en la Figura 22.

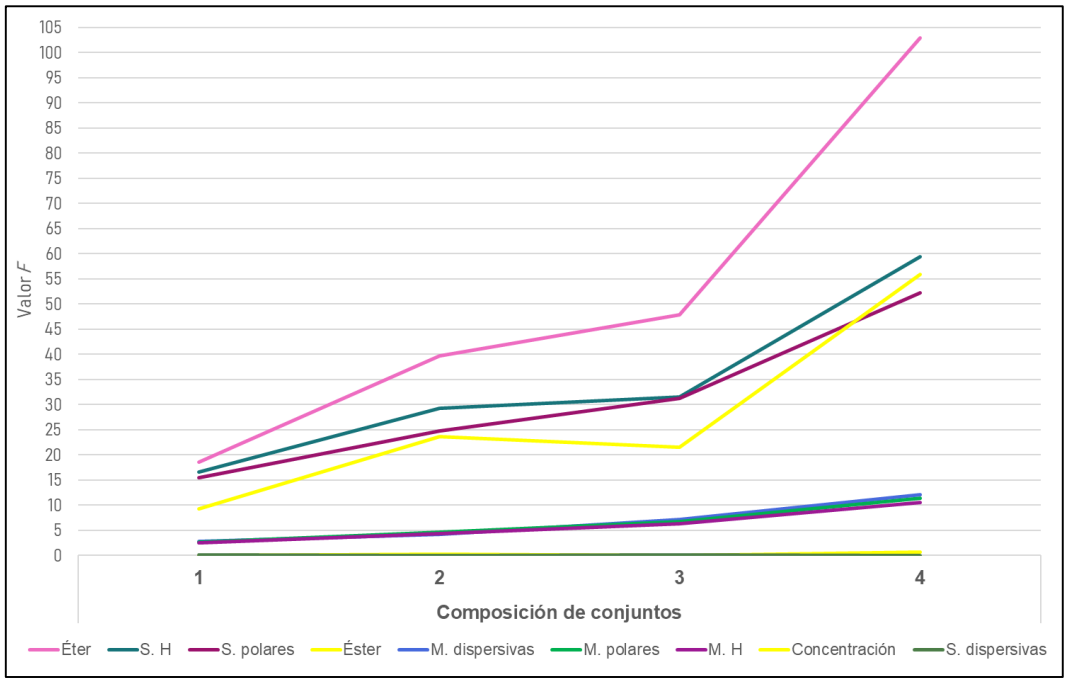


Figura 22. Valores críticos F para cada atributo por cada composición de conjuntos de entrenamiento.

En la Figura 22 se puede observar como los valores de F varían en función de la cantidad de ejemplos en los conjuntos, debido a que es la totalidad de valores con los que se calcula la varianza. Si se observa la variación en los valores *F* por atributo en orden de la composición de los conjuntos, se tiene una dependencia con respecto a la distribución y la cantidad de los ejemplos. La varianza es mayor cuando las clases de los ejemplos están uniformemente distribuidas, y cuando se adicionan los ejemplos con una segunda concentración (concentración variable). Esto denota que, la distribución de los valores que representan a cada atributo es mayor, de manera proporcional a la distribución uniforme de los ejemplos de acuerdo con su clase y cantidad.

La relación que existe entre la varianza y la contribución de un atributo a la clasificación, se infiere por la forma en que la diferencia entre los valores de un ejemplo u otro, lo cual distingue a un ejemplo de otro al momento de la selección de la clase a la que pertenece una instancia desconocida. Más que la cantidad de ejemplos en un conjunto de entrenamiento, es la distribución uniforme de los ejemplos de acuerdo con las clases lo que contribuye a un mejor desempeño clasificatorio del modelo.

Durante el análisis basado en los resultados de clasificación presentes en las tablas 40 a la 42, fue notoria una tendencia en ciertos solventes a la clasificación adecuada de sus productos con las moléculas estudiadas. De modo similar, algunos de los solventes del conjunto probado mostraron reincidencia en su clasificación errónea, como se observa en la Tabla 48.

Tabla 48. Porcentaje promedio de exactitud en la clasificación producida de acuerdo con cada uno de los 15 solventes experimentados para todos los OABs.

Solvente	%CA
Hexano	75.0
Ciclohexano	80.6
Heptano	72.2
Acetato de etilo	100.0
Acetona	100.0
THF	100.0
Tolueno	100.0
DMS	97.2
Cloroformo	100.0
Isopropanol	44.4
DMF	52.7
Metanol	97.2
Pentano	72.2
Acetonitrilo	61.0
Etanol	75.0

Con el fin de observar y localizar las causas de las más bajas clasificaciones, se realizó un análisis de los solventes con exactitud de clasificación menor. En la Tabla 49 se presentan los resultados de clasificación de los solventes con mayor cantidad de errores predictivos con cada una de las dos moléculas probadas (1_{14} y 2_{14}). A manera de comparación, se tienen las clases reales de 1_{14} y 2_{14} con cada solvente, y las etiquetas arrojadas por el modelo de

clasificación a partir de tres configuraciones distintas, es así que se pueden apreciar cada uno de los errores y su porcentaje total.

Tabla 49. Resultados de clasificación de acuerdo con los solventes con mayor cantidad de errores durante la predicción con los OABs 1_{14} y 2_{14}.

Solvente	%CA	OAB	*Composición de Conjuntos	Concentración Variable			Concentración Constante			Clase Real	% de Respuesta de clasificación	
				Corpus							YES	NO
				A	B	C	A	B	C			
Isopropanol	44.4	1_{14}	1	YES	NO	YES	NO	NO	YES	**NO**	14	4
			2	YES	YES	YES	YES	NO	YES			
			3	YES	YES	YES	YES	YES	YES			
		2_{14}	1	YES	NO	NO	YES	NO	NO	**YES**	12	6
			2	YES	YES	YES	YES	YES	YES			
			3	YES	NO	YES	YES	NO	YES			
DMF	52.7	1_{14}	1	NO	NO	YES	NO	NO	YES	**YES**	10	8
			2	YES	NO	YES	YES	NO	YES			
			3	YES	NO	YES	YES	NO	YES			
		2_{14}	1	YES	NO	YES	YES	NO	YES	**YES**	9	9
			2	YES	NO	YES	YES	NO	YES			
			3	YES	NO	NO	NO	NO	NO			
Acetonitrilo	61.0	1_{14}	1	NO	NO	NO	NO	NO	NO	**YES**	12	6
			2	YES	YES	YES	YES	YES	YES			
			3	YES	YES	YES	YES	YES	YES			
		2_{14}	1	NO	NO	NO	NO	NO	NO	**YES**	10	8
			2	YES	YES	YES	YES	YES	YES			
			3	YES	NO	YES	YES	NO	YES			

*Composición de los conjuntos:

1-Clases no distribuidas en conjuntos de entrenamiento y k=5.

2-Clases distribuidas en conjuntos de entrenamiento y k=5.

3-Clases distribuidas en conjuntos de entrenamiento y k=2.

En el caso del isopropanol, tomando en cuenta la experimentación con las configuraciones probadas y para cada tipo de corpus, se presentan más errores con la molécula 1_{14}, siendo 14 ejemplos erróneamente clasificados de un total de 18 (23% de aciertos, Tabla 49).

En la Tabla 50 se presentan las clases experimentales de cada OAB con cada uno de los solventes con menor clasificación correcta. Las casillas marcadas de color rosa son los ejemplos de cada uno de los tres solventes con los OABs de cadena éter más larga, por lo tanto, mayor similitud con los OABs de prueba (1_{14} y 2_{14}).

Tabla 50. Clases experimentales de cada OAB en los conjuntos de entrenamiento producida de acuerdo con los tres solventes de exactitud predictiva más baja.

Solvente	1_8	1_{10}	1_{12}	3_8	3_{12}	4_8	4_{10}	4_{12}
Isopropanol	YES	YES	YES	NO	YES	NO	NO	NO
DMF	NO	NO	YES	NO	NO	NO	NO	NO
Acetonitrilo	NO	NO	YES	NO	YES	NO	NO	NO

A partir del análisis de la Tabla 50, se infiere que es posible que haya mayor influencia de los ejemplos pertenecientes a los OABs de cadena éster más corta (metil) en los conjuntos de entrenamiento por la similitud con la cadena corta de los ejemplos de prueba (metil en 1_{14}), por lo tanto, la tendencia a etiquetar como YES, a pesar de ser errónea. Basados en esta observación, se puede concluir que, la similitud entre los ejemplos de los conjuntos de prueba y los de entrenamiento contribuye a la elevación de la exactitud de la clasificación, como es el caso visto en los atributos estructurales representados por la longitud de las cadenas éter y éster.

En la siguiente sección, se detallan las conclusiones a partir de cada experimentación realizada.

Capítulo 8

8 Conclusiones y recomendaciones

Este capítulo se divide en dos secciones, la primera explica las primeras conclusiones que ayudaron al diseño de un modelo de clasificación capaz de producir resultados exactos al etiquetar moléculas OAB nuevas. La segunda parte, explica cómo se logró un modelo con capacidad de clasificación alta, y las configuraciones a través de las cuales se consigue.

8.1 Influencia de la clasificación a partir de la validación y prueba con corpus cualitativos.

La manera en que un algoritmo procesa la información que brindan los atributos descriptivos, es diferente dependiendo de cómo son representados. Así pues, la clase de atributo, véase "cualitativo ordinal o carácter" ya sea en una escala establecida o descriptivo, tendrá como resultado dentro de un modelo una clasificación diferente a comparación del uso de atributos de clase "continua", es decir numéricos cuantitativos.

Fue posible observar que el tipo de atributo y lo que refleja acerca del proceso de gelificación y sus componentes, tiene mayor relevancia que la cantidad de atributos al momento de entrenar al algoritmo acerca de lo que se desea predecir. El manejo de las variables cualitativo-numéricas ordinales permite tener una clasificación más alta a comparación de la obtenida con atributos cualitativos de tipo alfanuméricos.

Tras haber probado una serie diferente de configuraciones y distintos valores de atributos categóricos, se puede conocer si las hipótesis inicialmente planteadas en la sección 5.4 página 74 se pueden aceptar o refutar.

1. La experimentación con atributos cualitativos (datos categóricos) pertenecientes a los solventes y los OAB´s, sí permite guiar a una configuración óptima de los hiperparámetros que otorguen mayor calidad de clasificación del modelo predictivo, por lo cual esta hipótesis se acepta.
2. La cantidad de atributos descriptivos no es directamente proporcional al aumento en la eficacia en el modelo kNN. Por lo que, esta hipótesis es rechazada. Tiene mayor relevancia la calidad de los atributos que su cantidad en un modelo, por lo que se acepta la hipótesis que habla sobre esta verdad.

En particular, para el contexto de los OABs y solventes analizados, se logró hasta un 100% de clasificación aplicando un valor de vecinos de k=3, a partir de atributos representados por valores numéricos de rangos amplios. Uno de los hiperparámetros fundamentales al momento de influenciar al desempeño de un modelo, es la selección adecuada de los atributos y sus valores. Lo que contribuye a concluir que existe un impacto en los valores numéricos de cada atributo, y la escala que se da entre ellos para establecer sus diferencias y similitudes. Esto se ve reflejado en la manera en que kNN aprende de esta información y el predictor selecciona una respuesta.

A partir de los resultados obtenidos a lo largo de la experimentación, se tiene que kNN es un algoritmo adecuado para el diseño de un modelo para el tipo de datos propuesto. En la configuración de este algoritmo, los hiperparámetros que se deben estudiar para conocer sus valores óptimos de acuerdo con los corpus evaluados, son: el valor de k, la métrica y el peso asignado a cada atributo.

Para conocer la influencia de los atributos alimentados en cada conjunto al no presentarse una diferencia en los resultados cuando el peso de los atributos se declara uniforme o cuando se asigna a la distancia entre ejemplos, se tiene la necesidad de analizar con un método

externo si existe alguna influencia de algún atributo con respecto a otro que guíe a la optimización en la clasificación, o en su defecto que cause ruido o sobre equipamiento en los algoritmos.

8.2 Influencia de la clasificación a partir de la validación y prueba con corpus fisicoquímicos.

En esta sección se presentan algunas de las contribuciones que se lograron a partir de la observación y análisis de los resultados obtenidos con el uso de información fisicoquímica en los atributos de dos corpus para un modelo predictivo. Cabe señalar que inicialmente se trabajó con una tercera parte de la base de datos disponible, que corresponde a la familia dodecil éter. Esto para agilizar el análisis, y posteriormente evaluar la aplicabilidad de las conclusiones al resto de los datos correspondientes a las familias de moléculas geladores.

La aplicación de los parámetros de Hansen de los componentes de un sistema de gelificación a un corpus fisicoquímico produce una exactitud de clasificación mayor que la obtenida con un corpus de valores asignados para describir cualitativamente solventes y moléculas.

Tabla 51. Comparativo de resultados más altos obtenidos con un corpus cualitativo vs un corpus fisicoquímico en kNN.

Valor de k	**%CA de clase (G, S o P)**	
	Corpus fisicoquímico	**Corpus cualitativo numérico ordinal**
3	100	100.00
5	100	91.93
10	100	88.70

La Tabla 51 muestra un comparativo entre los resultados obtenidos durante la validación simple de un corpus cualitativo alfanumérico de 15 solventes y 16 oligómeros compuesto de atributos que describen los oligómeros por el número de éter y éster en ellos, y a los solventes por su homogeneidad elemental, polaridad baja, media o alta, por su estructura cíclica o heterocíclica, grupo funcional y saturación de enlace; y los resultados de la validación simple

con un corpus fisicoquímico de 15 solventes y 3 oligómeros cuyos atributos de los solventes son los HSP y de los oligómeros el número de éter y de éster. Ambas experimentaciones se realizaron para valores de k=3, 5 y 10, usando la distancia Euclídea y la distancia asignada como peso para los atributos.

Se logró obtener una clasificación del 100% para todos los valores de vecinos a través de un corpus fisicoquímico con menor cantidad de atributos (5 en total) y de ejemplos (45 en total). Con esto se comprueba la factibilidad del uso de valores numéricos que caracterizan el comportamiento fisicoquímico de los solventes y los oligómeros como una alternativa que simplifica y aumenta la exactitud de la clasificación. Es necesaria la evaluación de un corpus con el resto de las moléculas (familias éter butil, hexil, octil y decil) para comprobar el funcionamiento de esta estructura de corpus como óptimo para la clasificación predictiva.

Se comprueba que la selección de los HSP para formar un corpus fisicoquímico permite una precisión del 100% en la clasificación cuando se valida con un conjunto de ejemplos de prueba conocido por el algoritmo.

8.3 Influencia de hiperparámetros en un corpus de clasificación, y en la configuración del algoritmo.

A partir de los comparativos de resultados probando diferentes configuraciones de los hiperparámetros aplicados a los modelos de validación en kNN, se presentan las siguientes conclusiones.

- Atributos.

Se tiene que el uso de los HSP de los solventes y el éter y éster de los oligómeros funcionan como atributos al obtener una clasificación mayor que la obtenida en pruebas homólogas con el uso de otros atributos como los cualitativos. El atributo Éster tiene mayor relevancia por encima del atributo Éter durante la clasificación, esto por las diferencias en los valores de este atributo para cada ejemplo, lo que ayuda a crear una diferencia al momento de la selección de la etiqueta por parte del modelo.

La distribución de los valores que representan a cada atributo es mayor, de manera proporcional a la distribución uniforme de los ejemplos de acuerdo con su clase y cantidad. Además de que existe una simbiosis entre la composición del conjunto utilizado para entrenar al modelo, y la del conjunto de ejemplos sometidos a prueba.

La relación que existe entre la varianza y la contribución de un atributo a la clasificación, se infiere por la forma en que la diferencia entre los valores de un ejemplo u otro, lo cual distingue a un ejemplo de otro al momento de la selección de la clase a la que pertenece una instancia desconocida. Más que la cantidad de ejemplos en un conjunto de entrenamiento, es la distribución uniforme de los ejemplos de acuerdo con las clases lo que contribuye a un mejor desempeño clasificatorio del modelo.

- Peso asignado a los atributos.

La asignación de un peso uniforme a los atributos contenidos en los ejemplos produce una exactitud en la clasificación notablemente menor que el uso de la distancia en la métrica Euclídea.

- Composición de los conjuntos.

De acuerdo con las pruebas realizadas variando la cantidad de ejemplos en los conjuntos de entrenamiento, duplicando estos, se pudo concluir que esto no aporta un aumento significativo al desempeño clasificatorio. Sin embargo, cuando se triplicaron los ejemplos con la etiqueta YES correspondiente a los geles, se formaron así conjuntos con clases de ejemplos distribuidos y a concentración constante, es decir, sin agregar esta variable como un atributo, lo cual sí aportó un claro aumento en las clases de prueba correctamente etiquetadas.

Como se menciona en el apartado de análisis por atributos, la cantidad de carbonos en la parte éter y éster contribuye a elevar la exactitud de clasificación. Esto se confirma cuando se observan los resultados del análisis de la composición de los conjuntos.

En conclusión, y con relación a las hipótesis previamente planteadas, se tiene lo siguiente.

1. El aumento en la cantidad de ejemplos y la distribución de estos mismos de acuerdo con su clase, para que se encuentren en cantidades equitativas en el corpus de datos y en los conjuntos de entrenamiento pueden optimizar el desempeño de la clasificación por parte del modelo. Esta hipótesis es aceptada.
2. Para el algoritmo kNN, el valor óptimo en la cantidad de vecinos más cercanos se encuentra entre los valores de k= 5 y 10, y para su selección se debe tomar en cuenta la robustez de los corpus de acuerdo con la cantidad de ejemplos. Esta hipótesis es aceptada.

8.4 Recomendaciones para el diseño de un modelo de clasificación.

Como parte de los productos obtenidos mediante este trabajo de investigación, se recomienda el siguiente diseño de modelo de clasificación para moléculas tipo OAB, descrito en las Figura 23 y Figura 24.

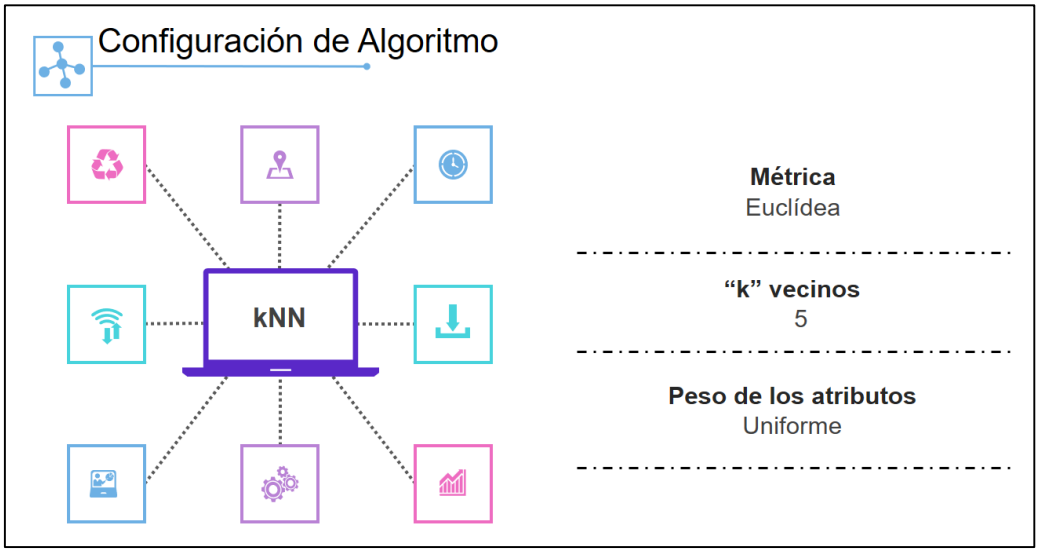


Figura 23. Configuración con mayor exactitud clasificatoria para kNN.

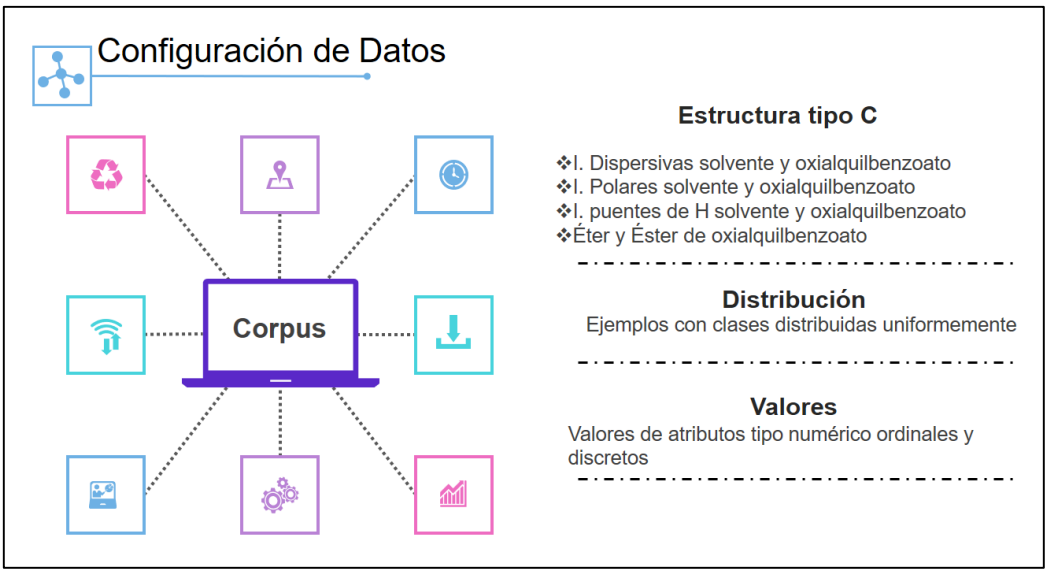


Figura 24. Configuración con mayor exactitud clasificatoria para un corpus de datos basado en moléculas tipo OAB.

En cuanto a los hiperparámetros que son relevantes para ser tomados en cuenta al diseñar un modelo de clasificación para moléculas tipo OAB, se mencionan los siguientes puntos.

1. Valores de k.
2. Configuración.
3. Estructura de los corpus.

Estos son los tres puntos que al conjugarse de manera precisa elevan el desempeño de los modelos durante la clasificación.

En cuanto a la composición de los conjuntos de entrenamiento, los parámetros que tienen impacto sobre la clasificación debido a la modificación del contenido de los ejemplos son los siguientes.

1. La distribución de las clases.
2. El tipo de variable (valor).
3. Los atributos descriptivos.

Glosario

Aprótico:
Solventes que no contienen en su estructura enlaces O-H ni N-H por lo que son incapaces de formar puentes de hidrógeno o donar protones. Casos típicos son la acetona, el hexano, el dimetilsulfoxido.

Anisótropo:
Propiedad general de la materia en la que las propiedades tales como elasticidad, temperatura, conductividad, velocidad de propagación de la luz varían de acuerdo con la dirección en que son examinadas. Un material anisótropo podría presentar diferentes características al tener una estructura molecular o atómica irregular, y ser analizado en diferentes direcciones.

Atributos continuos:
Son las variables que pueden tomar cualquier valor, generalmente dentro de un rango dado. Los valores que asumen son números reales y son representados por un número finito de dígitos.

Atributos discretos:
Variables que sólo pueden tomar como valor uno de los provistos en una lista predefinida de valores. Se representan a menudo por números enteros.

Boosting:
Meta algoritmo de aprendizaje supervisado que reduce el sesgo.

Dímero:
Especie química que consiste en dos subunidades estructuralmente similares denominadas monómeros unidas por enlaces que pueden ser fuertes o débiles.

GCMC:
Grand Canonical Monte Carlo simulation.

Enlace peptídico:
Enlace covalente formado entre el grupo amino (-NH2) de un aminoácido, y el grupo carboxilo (-COOH) de otro aminoácido.

Enantiomorfas:
Los enantiómeros, también llamados isómeros ópticos, son una clase de estereoisómeros tales que en la pareja de compuestos uno es imagen especular del otro y no son superponibles, es decir, cada uno es imagen especular no superponible con la otra.

Entropía:
Es una medida del desorden. En la teoría de la información basada en la entropía, se calcula el número de bits (información, preguntas sobre atributos) que hace falta

suministrar para conocer la clase a la que pertenece un ejemplo.

Esméctico:

Es el estado mesomorfo de la materia, más próximo al cristalino que al líquido.

Hiperparámetros:

Son utilizados para organizar y estandarizar la información que se va a ingresar al modelo. Son herramientas que se utilizan para describir la configuración del modelo.

HSP: Parámetros de solubilidad de Hansen, son llamados así en honor a Charles M. Hansen, quién fue el primero en descubrir la vinculación entre estos parámetros fisicoquímicos para estimar la compatibilidad entre materiales

In silico:

Hecho por computadora o vía simulación virtual.

Ingeniería Cristalina:

La comprensión de las interacciones intermoleculares en referencia a los entramados cristalinos que preceden el diseño de sólidos supramoleculares con propiedades físicas y químicas particulares.

Ligante-ligando:

En química de coordinación, un ligando es un ión o molécula que se une a un átomo de metal central para formar un complejo de coordinación.

Liotrópico:

Fase en la que se encuentran los conocidos cristales líquidos, tipo especial de estado de agregación de la materia que posee propiedades de las fases líquida y sólida. Dependiendo del tipo de cristal líquido, es posible que las moléculas tengan libertad de movimiento en un plano, pero no entre planos, y que tengan libertad de rotación, pero no de traslación.

LMWG:

Geles de bajo peso molecular o Low Molecular Weight Gels.

Metátesis:

La denominada reacción de metátesis, término acuñado a partir de las palabras griegas meta (cambio) y tithemi (lugar), consiste básicamente en la ruptura y formación de dobles enlaces en un proceso catalizado por ciertos complejos metálicos.

Moietie:

En química orgánica un moety es una parte de una molécula a la que normalmente se le otorga un nombre para identificarla, debido a que puede encontrarse dentro de otros tipos de moléculas más grandes. Este término se reserva para describir partes de moléculas con características esenciales, no para los grupos funcionales. En ocasiones, pueden

estar formados por uno o varios grupos funcionales o moeties más pequeños.

ROC:

Receiver Operating Characteristic, es una gráfica que ilustra la evaluación de un sistema de clasificación binario. Esta curva es creada trazando la tasa positiva verdadera (TPR) contra la tasa de falsos positivos (FPR). La tasa positiva verdadera también se conoce como sensibilidad o probabilidad de detección en el aprendizaje automático. La tasa de falsos positivos también se conoce como probabilidad de falsa alarma.

Parámetro:

Son los datos que se consideran como imprescindibles y orientativos para lograr evaluar o valorar una determinada situación.

Parametrizar:

Organizar y estandarizar la información que se ingresa en un sistema a manera de poder realizar distintos tipos de consulta y obtener resultados fiables.

Péptido:

Moléculas formadas por la unión de varios aminoácidos mediante enlaces peptídicos.

Prótico:

Proveniente de la palabra protones, los solventes que no contienen enlaces O-H ni N-H en su estructura. En solución, son capaces de donar hidrógenos o formar puentes de hidrógeno, casos comunes son el agua y los alcoholes.

Quiralidad:

Aplicado a la química orgánica, una molécula es quiral cuando ella y su imagen especular no se pueden superponer. La quiralidad a menudo se asocia con la presencia de carbonos asimétricos. Un carbono es asimétrico cuando está unido a cuatro sustituyentes diferentes.

Racémico:

Una mezcla racémica o racemato es una mezcla en la que dos compuestos químicos con actividad óptica, que guardan estrecha relación de imágenes especulares entre sí, se encuentran en proporciones equivalentes.

Solvatocromismo:

Es el término usado para describir el fenómeno que se observa cuando el color particular de un soluto es diferente cuando ese soluto se disuelve en distintos solventes.

Sorción:

Retención de una sustancia por otra cuando están en contacto; incluye las operaciones de absorción, adsorción, intercambio iónico y diálisis.

Tixotrópico:

Propiedad de algunos fluidos no newtonianos y pseudoplásticos que muestran un cambio de su viscosidad en el tiempo; cuanto más se somete el fluido a esfuerzos de cizalla, más disminuye su viscosidad. Un fluido tixotrópico tarda un tiempo finito en alcanzar una viscosidad en equilibrio cuando hay un cambio instantáneo en el ritmo de cizalla. El término también se aplica a los fluidos pseudoplásticos que no tienen una relación viscosidad/tiempo. Los fluidos tixotrópicos muestran una disminución de la viscosidad a lo largo del tiempo a una velocidad de corte constante.

Validación cruzada:

Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar la precisión de un modelo que se llevará a cabo a la práctica. Es una técnica muy utilizada en proyectos de inteligencia artificial para validar modelos generados.

Bibliografía

[1] Biswas G., Moon H., Boratynski P., Jeong B. y Kwon Y., «Structural sensitivity of peptoid-based low molecular mass organogelator,» *Materials and design,* vol. 108, nº Elsevier, pp. 659-665, 2016.

[2] J. Sosa-Sevilla, S.B. Brachetti , J. F. Pérez Sánchez, J. I. Lozano-Navarro y N. P. Díaz-Zavala, «Alkoxybenzoate Derivatives: Design and Gelation Effect on Organic Solvents, Fuels, and Oils,» *Water, air, and soil pollution,* vol. 232, nº 239, 2021.

[3] W. Truong, L. Lewis y P. Thordarson, «Functional molecular gels,» *The royan society of chemistry,* vol. 1, pp. 157-194, 2014.

[4] D. Zurcher y A. McNeil, «Tools for Identifying Gelator Scaffolds and Solvents,» *Journal of Organic Chemistry,* vol. 80, p. 2473−2478, 2015.

[5] J. Gupta, N. Berry y D. Adams, «Will it gel? Successful computational prediction of peptide gelators using physicochemical properties and molecular fingerprint,» *Chemical Science,* vol. 7, p. 4713–4719, 2016.

[6] A. Katrisky, M. Kuanar, S. Slavov y D. Hall, «Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction,» *Chemical Reviews.,* vol. 110, p. 5714–5789, 2010.

[7] Li F., Han J., Cao T., Lam W., Fan B., Tang W., Chen S., Lam K. y Li L., «Design of self-assembly dipeptide hydrogels and machine learning via their chemical features,» *Biophysics andComputational Biology,* vol. 116, nº 23, p. 11259–11264, 2019.

[8] Cihan M., Vianney J. M. y Er S., «Pushing the limits of solubility prediction via quality-oriented data selection,» *iScience,* pp. 1-17, 2020.

[9] Haghi, M., Li, J., Heidar-Zadeh, F., Liu, Y., Guan, X. y Head-Gordon, T., «Interplay of Feature Representation, Data, and Machine Learning Methods,» *Chem,* vol. 6, nº 7, pp. 1527-1542, 2020.

[10] Raynal, M. y Bouteiller, L., «Organogel formation rationalized by Hansen solubility parameters,» *Chemical Communications,* vol. 47, pp. 8271-8273, 2011.

[11] K. Diehn, O. Hyuntaek y R. Hashemipour, «Insights into organogelation and its kinetics from Hansen solubility parameters. Toward a priori predictions of molecular gelation,» *Soft matter,* vol. 10, pp. 2632-2640, 2014.

[12] Delbecq, F., Adenier, G. y Ogue, Y., «Prediction of solvent gelation via machine learning using Hansen solubility parameters,» *Journal of Molecular Liquids,* vol. 303, p. 112587, 2018.

[13] Iqbal, S., Miravet, J. y Escuder, B., «Biomimetic Self-assembly of Tetrapeptides into Fibrillar Networks and Organogels,» *Chemistry europe,* vol. 2008, nº 27, pp. 4580-4590, 2008.

[14] Haldar, S. y Karmakar, K., «A systematic understanding of gelation selfassembly: solvophobically assisted supramolecular gelation via conformational reorientation across amide functionality on a hydrophobically modulated dipeptide based ambidextrous gelator, N-n-acyl-(L)Val-X(OBn),,» *RSC Advances,* vol. 5, p. 66339–66354, 2015.

[15] H.J. Jong, «Stabilization of an Asymmetric Bolaamphiphilic Sugar-Based Crown Ether Hydrogel by Hydrogen Bonding Interaction and its Sol-gel Transcription,» *Tetrahedron,* vol. 63, pp. 7449-7456, 2007.

[16] P. Terech y S. Friol, «Rheometry of an Androstanol Steroid Derivative Paramagnetic Organogel. Methodology for a Comparison with a Fatty Acid Organogel,» *Tetrahedron,* vol. 62, pp. 7366-7374, 2007.

[17] T. Hirakura, Y. Nomura, Y. Aoyama y K. Akiyoshi, «Photoresponsive Nano-gels Formed by the Self-Assembly of Spiropyrane-Bearing Pullulan That Act as Artificial Molecular Chaperones,» *Biomacromolecules,* vol. 5, pp. 1804-1809, 2004.

[18] C. Hansen, Hansen Solubility Parameters: A User's Handbook, Florida, EU: CRC Press, 2000.

[19] J. Israelachvili, Intermolecular and surface forces, London, England: Academic Press, 1992, pp. 341-435.

[20] M. Carranza, «Complejos de metales de transición con ligandos n-dadores. Análisis e influencia de las interacciones supramoleculares,» Ciudad del Real, España, 2009.

[21] H. Beibei, S. Wei, Y. Baixue , L. Heran, Z. Liuchenzi y L. Sanming, «Application of Solvent Parameters for Predicting Organogel Formation,» *American Association of Pharmaceutical Scientists,* vol. 12, pp. 1-13, 2018.

[22] Y. Lan, M. Corradini, R. Weiss, R. Raghavanc y M. Rogers, «To gel or not to gel: correlating molecular gelation with solvent parameters,» *Royal Society of Chemistry,* 2015.

[23] F. Acuña, «Química Orgánica,» 2006.

[24] R. Chang, Química, D.F., México: McGrawHill/Interamericana Editores, 2006.

[25] A. Liu, «Data Science and data Scientist,» *IBM Analytics,* p. 11, 2015.

[26] Stuart Rusell y Peter Norvig, Inteligencia Artificial, un enfoque moderno, segunda ed., Madrid: Pearson Prentice Hall, 2004.

[27] M. Kuhn y K. Johnson, «Ciencia de datos,» 16 agosto 2019. [En línea]. Available: http://topepo.github.io/caret/index.html. [Último acceso: 27 octubre 2019].

[28] B. Klein, «Python course,» Python, 2018. [En línea]. Available: https://www.python-course.eu/machine_learning.php..

[29] C. G. Cambronero y I. G. Moreno, «Algoritmos de aprendizaje: KNN & KMEANS,» *Inteligencia en Redes de Telecomunicación,* vol. 1, pp. 1-8, 2016.

[30] J. A. Rodrigo, «Machine Learning con R y caret,» www.cienciadedatos.ne, 2018.

[31] D. Nieves, «Quora,» 2019. [En línea]. Available: https://es.quora.com/Qué-es-el-escenario-de-entrenamiento-validación-y-prueba-de-conjuntos-de-datos-en-aprendizaje-automático. [Último acceso: 26 octubre 2019].

[32] A. Casis, «Inteligencia artificial,» 20 octubre 2015. [En línea]. Available: https://inteligenciaartificial101.wordpress.com/2015/10/20/aprendizaje-supervisado/.

[33] V. Roman, «medium,» 19 march 2019. [En línea]. Available: https://medium.com/datos-y-ciencia/machine-learning-c%C3%B3mo-desarrollar-un-modelo-desde-cero-cc17654f0d48. [Último acceso: march 2020].

[34] FH Joanneum, « Cross-Validation Explained,» Institute for Genomics and Bioinformatics, 2005.

[35] S. Rozada, «Hiperparámetros en Machine Learning,» 16 mayo 2018. [En línea]. Available: https://es.quora.com/Qué-son-los-hiperparametros-y-para-qué-sirven-en-machine-learning. [Último acceso: 28 octubre 2019].

[36] Delbecq, F., Masuda, Y., Ogue, Y. y Kawai, t., «Salt complexes of two-component N-acylamino acid diastereoisomers: self-assembly studies and modulation of gelation abilities,» *Tetrahedron Letters,* vol. 53, nº 48, pp. 6588-6593, 2012.

[37] A. Tropsha, «Best Practices for QSAR Model Development, Validation, and Exploitation,» *Molecular Informatics,* vol. 29, p. 476 – 488, 2010.

[38] J. Gao, S. Wu y M.A. Rogers, «Harnessing Hansen solubility parameters to predict organogel formation,» *Journal of materials Chemistry,* vol. 22, nº 25, pp. 12651-12658, 2012.

[39] M. Raynal y L. Bouteiller, «Organogel formation rationalized by Hansen solubility parameters,» *Chemical Communications*, vol. 47, pp. 8271-8273, 2011.

[40] J. Bonnet, G. Suissa, M. Raynal y L. Bouteiller, «Organogel formation rationalized by Hansen solubility parameters: dos and don´ts,» *Soft Matter,* vol. 10, pp. 3154-3160, 2014.

[41] J. Bonnet, G. Suissa, M. Raynal y L. Bouteiller, «Organogel formation rationalized by Hansen solubility parameters: influence of gelator structure,» *Soft Matter,* vol. 11, pp. 2308-2314, 2015.

[42] R. Fedors, «A Method for Estimating Both the Solubility Parameters and molar volumes of liquids,» *Polymer engineering and science,* vol. 14, nº 2, pp. 147-154, 1974.

[43] J. Brandrup, E. H. Immergut y E. A. Grulke, Polymer Handbook, USA: A Wiley-Interscience Publication, 1999.

[44] L. Jordon, Molecular Gels: Materials with Self-Assembled Fibrillar Networks, New York: The Chemical Catalog Company, 1926.

[45] N. Sangeetha y U. Maitra, «Supramolecular gels: Functions and uses,» *The Royal Society of Chemistry,* vol. 34, pp. 821-836, 2005.

[46] R. Weiss, «The Past, Present, and Future of Molecular Gels. What Is the Status of the Field, and Where Is It Going,» *Journal of the American Chemical Society,* vol. 136, pp. 7519-7530, 2014.

[47] T. Adalder y P. Dastidar, «Crystal Engineering Approach toward Selective Formation of an Asymmetric Supramolecular Synthon in Primary Ammonium Monocarboxylate (PAM) Salts and Their Gelation Studies,» *Cristal Growth and design,* vol. 14, p. 2254–2262, 2014.

[48] K. King y A. McNeil, «Streamlined approach to a new gelator: inspiration from solid-state interactions for a mercury-induced gelation,» *Chemical Communications,* vol. 20, pp. 3511-3513, 2010.

[49] J. Craig, *OpenSMILES specification,* Cambridge, UK, 2016.

[50] M. Corradini y M. Rogers, «Molecular gels: improving selection and design through computational methods,» *Current opinion in food science,* vol. 9, pp. 84-92, 2016.

[51] J. Kaszynska, A. Lapiski, M. Bielejewski, R. Luboradzki y J. Tritt-Goc, «On the relation between the solvent parameters and the physical properties of methyl-4,6-O-benzylidene-α-D-glucopyranoside organogels,» *Tetrahedron,* vol. 68, pp. 3803-3810, 2012.

[52] Z. Guangyu y J. Dordick, «Solvent Effect on Organogel Formation by Low Molecular Weight Molecules,» *Chemical Materials,* vol. 18, pp. 5988-5995, 2006.

[53] R. Taft, J. Abboud y M. Kamlet, «Linear Solvation Energy Relationships, an Analysis of Swain's Solvent "Acity" and "Basity" Scales,» *Journal of Organic Chemistry,* vol. 49, pp. 2001-2005, 1984.

[54] . M. Raynal y L. Bouteiller, «Organogel formation rationalized by Hansen solubility parameters,» *Chemical Communications,* vol. 29, pp. 8271-8273, 2011.

[55] C. Hansen y H. Yamamoto, «Hansen Solubility Parameters in Practice,» p. 2013.

[56] W. Fräßdorf, M. Fahrländer, K. Fuchs y C. Friedrich, «Thermorheological properties of self-assembled dibenzylidene sorbitol structures in various polymer matrices: Determination and prediction of characteristic temperatures,» *Journal of rheology,* vol. 47, pp. 1445-1454, 2003.

[57] R. Weiss y P. Terech, Molecular Gels: Materials with Self-Assembled Fibrillar Networks. Dordretch, Netherlands: Springer, 2006.

[58] R. Weiss y P. Terech, «Low Molecular Mass Gelators of Organic Liquids and the Properties of Their Gels,» *Chemical Review,* vol. 97, pp. 3133-3159, 1997.

[59] M. Bell, A. Aggeli, N. Boden, J. Keen, P. Knowles, T. McLeish, M. Pitkeathly y S. Radford, «Responsive gels formed by the spontaneous self-assembly of peptides into polymeric [beta]- sheet tapes,» *Nature,* vol. 386, pp. 259-262, 1997.

[60] R. Scott y J. Hildebrand, Regular Solutions, Englewood Cliffs, NJ: Prentice-Hall Inc., 1962.

[61] C. Johnson, M. Gordon y D. Boucher, «Rationalizing the self- assembly of poly-(3-hexylthiophene) using solubility and solvatochromic parameters,» *Journal of Polymers Science,* vol. 53, pp. 841-850, 2015.

[62] M. Nic, J. Jirat y B. Kosata, «IUPAC, Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"),» Oxford, 2006.

[63] A. Cerda, «Efectos de solvente en ambiente de líquidos iónicos en reacciones orgánicas,» Santiago, Chile, 2013.

[64] C. Reichardt, «Solvatochromic Dyes as Solvent Polarity Indicators,» *Chemical Reviews,* vol. 94, pp. 2319-2358, 1994.

[65] C. M. Hansen, Hansen Solubility Parameters a User´s Handbook, C. Press, Ed., Boca Raton, Florida, 2000, p. 222.

[66] P. Refaeilzadeh, L. Tang y H. Liu, «Cross-Validation,» de *Encyclopedia of Database Systems*, Ö. M. LIU L., Ed., Boston, MA: Springer, 2009.

[67] Unipython, «Desition trees,» www.unipython.com, 2015.

[68] «accelrys,» collaborative science, [En línea]. Available: http://accelerys.com/products/collaborative-science/biovia-pipeline-pilot/..

[69] M. Kuhn, J. Wing, S. Weston, A. Williams y C. Keefer, «Classification and regression Training. R package version,» [En línea]. Available: http://CRAN.R-project.org/package.. [Último acceso: 03 10 2015].

[70] R. Team, «R foundation for statical computing,» 2015.