EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Modelo predictivo de los 16 factores de personalidad para el análisis de la terminación de un estudiante de maestría

Presentada por

Ing. Ernesto Echeverria Ignacio

como requisito para la obtención del grado de
Maestro en Ciencias de la Computación

Director de tesis
Dra. Alicia Martínez Rebollar

Codirector de tesis:

Dr. Juan Francisco Mosiño

cenidet®
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Morelos, México. Septiembre de 2024

Centro Nacional de Investigación
y Desarrollo Tecnológico
Departamento de Ciencias Computacionales

Cuernavaca, Mor., 26/Agosto/2024
OFICIO No. DCC/141/2024

Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de **ERNESTO ECHEVERRÍA IGNACIO** con número de control M21CE010, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"Modelo predictivo de los 16 Factores de personalidad para el análisis de la terminación de un estudiante de maestría"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

Alicia Martínez Rebollar
Directora de tesis

Juan Francisco Mosiño

Juan Francisco Mosiño
Codirector de tesis

Hugo Estrada Esquivel
Revisor 1

José Jassón Flores Prieto
Revisor 2

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
29 AGO 2024
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SERVICIOS ESCOLARES
RECIBIDO

SEP TecNM CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
RECIBIDO
29 AGO 2024
SUBDIRECCIÓN ACADÉMICA

C.c.p. Depto. Servicios Escolares.
Expediente / Estudiante

Cuernavaca, Mor., 29/agosto/2024
No. De Oficio: SAC/254/2024
Asunto: Autorización de impresión de tesis

ERNESTO ECHEVERRÍA IGNACIO
CANDIDATO AL GRADO DE MAESTRO
EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Modelo predictivo de los 16 Factores de personalidad para el análisis de la terminación de un estudiante de maestría"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica
"Conocimiento y tecnología al servicio de México"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

C. c. p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz

Dedicatoria

"Siempre hay una primera vez para todo, incluso en una historia en la que me tocó un papel diferente". A veces no se trata de ser el mejor para llegar a la meta cuando no se trata de una competencia sino de estar motivado y consciente de que todo esfuerzo tiene su recompensa, por ello creo que el trabajar con psicología y computación me ha abierto la mente para poder discernir y comprender el hecho de que no todos trabajamos de la misma manera incluso cuando estamos enfermos, cuando estamos sanos, cuando estamos pasando dificultades emocionales, por eso creo que llegar al final de este proyecto ha sido uno de los más grandes logros que he tenido, porque me doy cuenta de lo mucho que he progresado.

Este trabajo de investigación va dedicado principalmente a las personas que me han motivado a seguir creciendo profesionalmente y una de ellas Socorro Linares, quien es secretaria del Instituto Nacional de Energías Eléctricas y Limpias por siempre motivarme a seguir preparándome, a mi mejor amigo Juan Alberto Núñez que siempre creyó que podía concluir con este proyecto y Nau Bueno Acuña que siempre me ha apoyado en todos mis proyectos de vida. A Dios, por darme la fuerza necesaria para culminar esta meta. A mis padres, por todo su amor y por motivarme a seguir hacia adelante. También a mis hermanos, por brindarme su apoyo moral en esas noches que tocaba investigar. Y, finalmente, a los que no creyeron en mí, con su actitud lograron que tomará más impulso.

Agradecimientos

Quiero expresar mi más profundo agradecimiento a mi directora de tesis, la Dra. Alicia Martínez, y a mi co-director de tesis, el Dr. Juan Francisco Mosiño. También extiendo mi gratitud a los miembros del comité tutorial, el Dr. Jassón Flores y el Dr. Hugo Estrada Esquivel, por su tiempo, dedicación, valiosos consejos y observaciones, los cuales han sido fundamentales para mejorar esta investigación.

De manera especial, agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca otorgada, la cual fue esencial para la realización de mis estudios de maestría. Su apoyo ha sido un pilar en mi formación académica y profesional, permitiéndome concentrarme en mi investigación y avanzar en mi carrera.

Asimismo, agradezco al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), que forma parte del TecNM, por brindarme la oportunidad de pertenecer a su comunidad estudiantil y por proporcionarme un entorno propicio para el aprendizaje y desarrollo.

A mi familia, gracias por respetar siempre mis decisiones y por ser una fuente constante de apoyo e impulso en todos mis proyectos. A mis padres, les agradezco por inculcarme valores, darme la educación necesaria para ser una persona íntegra, y enseñarme que todo se logra con trabajo. A mis hermanos, gracias por creer en mí y por recordarme siempre que soy un ejemplo de superación personal y profesional.

Resumen

La personalidad desempeña un papel fundamental en la eficiencia terminal de los programas de posgrado, y su análisis predictivo puede mejorar significativamente el proceso de selección de estudiantes. Este estudio propone un modelo predictivo basado en la técnica K-nn que utiliza el cuestionario de personalidad 16PF para anticipar la eficiencia terminal de los estudiantes de maestría en el Centro de Investigación y Desarrollo Tecnológico (CENIDET). La metodología comprende tres etapas: recolección y análisis de datos, pre-procesamiento de datos y modelado. Se desarrollaron experimentos utilizando seis algoritmos de minería de datos y se evaluó su desempeño en métricas como precisión, puntuación F1 y Recall. El algoritmo Random Forest demostró el mejor rendimiento, logrando una precisión del 82.35% en la clasificación de eficiencia terminal. Este modelo predictivo tiene el potencial de apoyar a futuros estudiantes en la toma de decisiones informadas sobre sus programas académicos, al tiempo que aumenta la motivación académica.

Palabras claves

16 PF, Personalidad, Eficiencia terminal, Modelos predictivos, Random Forest

Abstract

Personality plays a fundamental role in the student graduation rate in postgraduate programs, and its predictive analysis can significantly enhance the student selection process. This study proposes a predictive model based on the Random Forest (RF) technique that uses the 16PF personality questionnaire to anticipate the graduation rate of master's students at the Center for Research and Technological Development (CENIDET). The methodology comprises three stages: data collection and analysis, data preprocessing, and modeling. Experiments were conducted using six data mining algorithms, and their performance was evaluated in metrics such as accuracy, F1 score, and recall. The Random Forest algorithm demonstrated the best performance, achieving an accuracy of 82.35% in terminal efficiency classification. This predictive model has the potential to support prospective students in making informed decisions about their academic programs while enhancing academic motivation.

Keywords:

16PF, Personality, Terminal Efficiency, Predictive Models, Random Forest

Contenido

Capítulo 1 Introducción

1.1 Introducción/ Motivación

Actualmente dentro de los principales problemas que tiene la educación en México, se cuentan los altos índices de reprobación de materias, la deserción de alumnos y la baja eficiencia terminal de los egresados(Rochin Berumen, 2021). Resulta fácilmente discernible concebir entonces los problemas de deserción, rezago y baja Eficiencia Terminal como manifestaciones de una falta de calidad del proceso educativo. En el caso de la Eficiencia Terminal (ET), esta demuestra claramente los estragos de la reprobación y deserción (rendimiento escolar), además de que cuando se combina la información que la ET arroja con otros indicadores como la duración promedio de los estudios de los egresados y desertores con indicadores de gasto educativo es posible obtener una visión de los montos adicionales o pérdidas del sistema.

La Secretaría de Educación Pública (SEP)define la eficiencia terminal como la proporción entre el número de alumnos que ingresan y los que egresan de una misma generación considerando el año de ingreso y el año de egreso según la duración del plan de estudios. Por otro lado, la eficiencia terminal en el contexto educativo se define como "la proporción de estudiantes que concluyen un programa en determinado momento, frente al total de los que lo iniciaron un cierto número de años antes" (Cuéllar Saavedra & Espinoza, 2006a). Esta métrica es crucial para evaluar la efectividad del sistema educativo en retener a los estudiantes y guiarlos hacia la graduación o finalización exitosa de sus estudios. En el caso del sistema educativo mexicano, mejorar la eficiencia terminal es un objetivo prioritario para promover el acceso equitativo y la calidad educativa en todos los niveles. Estos problemas se atribuyen a varias causas entre las que figuran la rigidez y especialización excesiva de los planes de estudio, los métodos obsoletos de enseñanza y evaluación(*Gallardo et al 2019 Eficiencia Terminal*, n.d.) .

El estudio de variables psicológicas de personalidad ha cobrado relevancia ya que su investigación ha ayudado a determinar patrones normales o anormales de conducta de un individuo a comportarse en diferentes situaciones (Zambrano Cruz, n.d.). La personalidad como un patrón de características estables de una persona a comportarse de una manera determinada en diferentes situaciones, que nos hacen únicos e irrepetibles(Vinet, 2006).

El cuestionario de 16 factores de personalidad es un instrumento útil para predecir las conductas de las personas en diferentes situaciones y actividades de entre ellas la orientación personal, la orientación escolar y la orientación profesional como la selección de personal en el ámbito

laboral (Boyle et al., 2008). El Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) utiliza varios instrumentos de selección para identificar a los estudiantes con el perfil regular para hacer un posgrado como I) Exámenes, II) Entrevista con la coordinación, III) Curso propedéutico (TecNM | Tecnológico Nacional de México Campus CENIDET, n.d.) El utilizar estos instrumentos ayudan a mejorar el proceso de selección de los estudiantes, sin embargo, no todos llegan a titularse en el tiempo que se les define para la maestría que es de 2 años. Esto pone en riesgo la certificación de calidad del CENIDET ya que es evaluado por el Programa Nacional de Posgrados de Calidad.

1.2 Planteamiento del problema

Desde hace algún tiempo está cobrando relevancia el estudio de variables psicológicas de personalidad asociadas a la selección de personas para integrarse en empresas organizacionales. Este tema es de gran importancia para la mejora de selección de empleados, así como estudiantes en las diferentes instituciones de educativas. Algunos centros de investigación toman diferentes medidas de selección, por ejemplo, examen de conocimientos generales, examen de idiomas, examen EXANI-III Ceneval, entrevistas con la coordinación, en algunos casos especiales toman en cuenta el promedio obtenido en la licenciatura y exámenes psicométricos (BIOTECNLOGIA, s.f., párrafo quinto).

El Centro Nacional de Investigación y Desarrollo Tecnológico(CENIDET) también aplica varios instrumentos de selección para identificar a los estudiantes con el perfil idóneo para hacer un posgrado, sin embargo, es difícil que todos lleguen a titularse en tiempo destinado de dos años, por ello, uno de los instrumentos es la aplicación del cuestionario 16 de factores de personalidad, que mide las aptitudes y actitudes que deben tener los aspirantes para realizar dicho nivel académico, ya que es evaluado por el Programa Nacional de Posgrados de Calidad (PNPC).

El Sistema Nacional de Posgrados (SNP) forma parte de la política pública de fomento a la calidad del posgrado nacional que el Consejo Nacional de Ciencia y Tecnología y la Subsecretaría de Educación Superior de la Secretaría de Educación Pública han impulsado de manera ininterrumpidamente desde 1991 (CONACYT, 2019, párrafo primero).

Los exámenes psicométricos son instrumentos de medida de personalidad. La personalidad es un patrón persistente de actitudes, pensamientos, sentimientos y repertorio de conductas que caracterizan a una persona y que tienen una cierta estabilidad a lo largo de su vida de tal modo que las manifestaciones de ese patrón en las diferentes situaciones poseen algún grado de predictibilidad. Actualmente el CENIDET cuenta con información de generaciones anteriores. Esta información, es útil para predecir si un aspirante es candidato idóneo para desarrollar las tareas propias de la investigación. Tomar en cuenta esta información ayudaría al Centro de Investigación a tener más de un 60% de alumnos titulados a tiempo.

1.3 Estado del arte

1.3.1 Introducción

La eficiencia terminal en programas de maestría, entendida como la culminación exitosa y oportuna de los estudios, es un indicador crucial del rendimiento académico y de la efectividad de las instituciones educativas. Diversos factores pueden influir en esta eficiencia, incluyendo las características personales de los estudiantes. Este estado del arte explora cómo la personalidad, evaluada mediante la encuesta de 16 factores de personalidad (16PF), puede predecir la eficiencia terminal en programas de maestría.

1.3.2 Eficiencia Terminal en la Educación Superior

La eficiencia terminal ha sido ampliamente estudiada en el contexto de la educación superior. Investigaciones previas han identificado factores académicos, socioeconómicos y personales como determinantes clave. Los factores como el apoyo institucional, la motivación intrínseca y el manejo del tiempo son críticos para la culminación exitosa de los estudios de posgrado (Soria-Barreto & Zúñiga-Jara, 2014).

1.3.3 Personalidad y Rendimiento Académico

La relación entre personalidad y rendimiento académico ha sido un foco de investigación durante décadas. De acuerdo con el modelo de los Cinco Grandes (Big Five), ciertos rasgos de personalidad, como la responsabilidad y la apertura a nuevas experiencias, se han correlacionado positivamente con el éxito académico (O. P. John et al., n.d.). Sin embargo, el modelo 16PF, desarrollado por Raymond Cattell, ofrece una evaluación más detallada y matizada de la personalidad, con 16 factores específicos que pueden proporcionar una visión más granular.

1.3.4. La Encuesta de 16 Factores de Personalidad (16PF)

El 16PF es un instrumento psicológico robusto que evalúa 16 dimensiones de la personalidad, incluyendo factores como la estabilidad emocional, la dominancia, la sociabilidad y la sensibilidad (Boyle et al., 2008a). Estudios como el de Kline (2000) han validado su uso en diversas aplicaciones, desde la selección de personal hasta la orientación educativa. En el contexto educativo, investigaciones han demostrado que ciertos factores del 16PF pueden predecir comportamientos académicos, como la persistencia y la adaptabilidad al entorno de estudio (Rushton et al., 1983).

1.3.5. Predicción de Eficiencia Terminal mediante el 16PF

Varios estudios han intentado vincular la personalidad con la eficiencia terminal en la educación superior. Martínez y Sánchez (2017) encontraron que estudiantes con altos puntajes en los factores de responsabilidad y estabilidad emocional del 16PF tenían mayores probabilidades de completar sus programas de maestría en el tiempo estipulado. Asimismo, García y Rodríguez (2020) destacaron que la combinación de altos niveles de autoconfianza y baja ansiedad, también medidos por el 16PF, se asocia con una mejor gestión del estrés académico y, por ende, con una mayor eficiencia terminal.

1.3.6. Vacíos y Oportunidades de Investigación

A pesar de la evidencia existente, aún hay vacíos significativos en la literatura. La mayoría de los estudios se han centrado en programas de pregrado, dejando un área poco explorada en el contexto de programas de maestría. Además, hay una necesidad de investigaciones longitudinales que puedan establecer relaciones causales entre la personalidad y la eficiencia terminal. También es crucial explorar la interacción entre los factores de personalidad y otros determinantes como el apoyo institucional y las estrategias de aprendizaje.

1.3.7 Conclusiones

El estado del arte revela que la personalidad, medida a través del 16PF, tiene el potencial de ser un predictor significativo de la eficiencia terminal en programas de maestría. Sin embargo, se necesita más investigación para comprender plenamente estas relaciones y para desarrollar intervenciones efectivas que puedan mejorar la eficiencia terminal basándose en las características personales de los estudiantes.

1.4 Objetivos

1.4.1 Objetivo general

Desarrollar un modelo predictivo que utilice la encuesta de los 16PF (factores de personalidad) para predecir si un aspirante terminará sus estudios de maestría utilizando algoritmos de Inteligencia Artificial.

1.4.2 Objetivos específicos

Formar un datasets con la información de los 16 factores de personalidad, así como información de la generación que ingresó el estudiante, y los meses que le llevó estudiar la maestría para utilizarlo como dataset de entrenamiento del modelo predictivo.

Realizar el preprocesamiento al dataset obtenido para mejorar la calidad, utilidad y coherencia de los datos utilizando técnicas para la limpieza, normalización de datos, balanceo de clases…

Diseñar un marco experimental que permita comparar diversos modelos predictivos utilizando algoritmos de aprendizaje máquina.

Capítulo 2 Presenta el Marco teórico

El marco teórico de la personalidad abarca diversas teorías y enfoques que buscan comprender y explicar los patrones consistentes de pensamientos, sentimientos y comportamientos que caracterizan a un individuo a lo largo del tiempo y en diferentes situaciones. Aquí hay un resumen de algunas de las principales teorías de la personalidad:

2.1 Personalidad

La personalidad se refiere a los patrones consistentes de pensamientos, emociones y comportamientos que caracterizan a un individuo a lo largo del tiempo y en diversas situaciones. Es una construcción psicológica compleja que se desarrolla a partir de la interacción de factores genéticos, biológicos y ambientales. La personalidad influye en cómo percibimos el mundo, cómo nos relacionamos con los demás y cómo nos comportamos en diferentes contextos. Se estudia a través de diversas teorías y enfoques, que buscan comprender sus dimensiones, estructuras y procesos subyacentes. La personalidad tiene una gran importancia en nuestra vida cotidiana, ya que afecta nuestra forma de pensar, sentir y actuar, y puede influir en nuestras relaciones interpersonales, logros académicos y éxito en el trabajo (Roberts & Mroczek, 2008).

2.1.1La prueba de 16 factores de personalidad

La prueba de 16 factores de personalidad, también conocida como el Inventario de Personalidad de 16 Factores (16PF), es una herramienta psicométrica utilizada para evaluar y medir diferentes aspectos de la personalidad de un individuo (Boyle et al., 2008b). A continuación, se presenta un marco conceptual que abarca los principales aspectos de esta prueba:

2.1.2 Teoría de la personalidad

La prueba de 16 factores de personalidad está fundamentada en la teoría de la personalidad propuesta por Raymond Cattell. Según esta teoría, la personalidad puede entenderse en términos de múltiples dimensiones o factores subyacentes, en lugar de conceptos unitarios (Schermer et al., 2020).

2.1.3 Factores de personalidad

La prueba de 16 factores evalúa 16 dimensiones o factores de la personalidad, que representan diferentes aspectos del comportamiento y el funcionamiento psicológico del individuo. Estos factores son el resultado de análisis estadísticos de los rasgos de personalidad observados en una amplia muestra de individuos (Boyle & Lennon, 1994).

2.1.4 Clasificación de los factores

Los 16 factores de personalidad se clasifican en cuatro grandes dimensiones, que a su vez se subdividen en factores más específicos:

Factores Primarios:

1 Inteligencia
2 Estabilidad emocional
3 Dominancia
4 Impulsividad

Factores Secundarios:

5 Calidez
6 Razón emocional
7 Dominancia
8 Sensibilidad al cambio
9 Vigilancia
10 Privacidad
11 Apertura
12 Autocontrol
13 Tensión
14 Vivacidad
15 Abstracción
16 Reserva

2.1.5 Evaluación

La prueba de 16 factores de personalidad se administra mediante un cuestionario que consta de una serie de afirmaciones o preguntas sobre el comportamiento, las actitudes y las preferencias del individuo. El individuo responde a estas afirmaciones en función de su grado de acuerdo o desacuerdo con cada una de ellas (Pietrzak et al., 2015).

2.1.6 Interpretación

Los resultados de la prueba se interpretan en función de la puntuación obtenida en cada uno de los 16 factores de personalidad. Estas puntuaciones proporcionan información sobre las

características y tendencias del individuo en relación con cada uno de los factores evaluados (Boyle, n.d.).

2.1.7 Aplicaciones

La prueba de 16 factores de personalidad se utiliza en una variedad de contextos, como la selección de personal, la orientación vocacional, la psicoterapia, la investigación psicológica y el desarrollo personal. Proporciona información útil para comprender el funcionamiento psicológico del individuo y puede utilizarse para identificar áreas de fortaleza y áreas de desarrollo potencial(Boyle et al., 2008b).

En resumen, la prueba de 16 factores de personalidad es una herramienta ampliamente utilizada para evaluar y medir diferentes aspectos de la personalidad de un individuo, basada en una teoría psicológica sólida y fundamentada en la investigación empírica.

2.1.8 Eficiencia terminal

El concepto de "eficiencia terminal" se refiere a la capacidad de un sistema educativo para lograr que los estudiantes completen exitosamente su nivel de educación sin retrasos ni abandonos (Cuéllar Saavedra & Espinoza, 2006b) (*Revista146_S5A1ES*, n.d.). En el contexto de México, el marco conceptual de la eficiencia terminal abarca varios aspectos:

2.1.9 Acceso y cobertura

La eficiencia terminal está estrechamente relacionada con el acceso y la cobertura de la educación. Se refiere a asegurar que todos los niños y jóvenes tengan la oportunidad de ingresar y completar sus estudios en los diferentes niveles educativos, desde preescolar hasta la educación superior (Mariscal et al., n.d.).

2.2 Aprendizaje automático

El aprendizaje automático es una rama de la inteligencia artificial que se enfoca en desarrollar algoritmos y modelos que permitan a las computadoras aprender a partir de datos y realizar tareas específicas sin necesidad de ser programadas explícitamente para cada tarea (Horvitz & Mulligan, 2015). Aquí está el marco conceptual del aprendizaje automático:

2.2.1 Algoritmos de aprendizaje

Los algoritmos de aprendizaje automático son los métodos computacionales que permiten a las máquinas aprender a partir de los datos (Goodfellow et al., n.d.). Estos algoritmos se dividen en

varias categorías, incluyendo aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado y aprendizaje por refuerzo.

2.2.2 Aprendizaje supervisado

En el aprendizaje supervisado, los algoritmos se entrenan utilizando datos etiquetados, es decir, datos que tienen asociadas etiquetas o respuestas conocidas. El objetivo es aprender una función que mapee las entradas a las salidas deseadas, lo que permite al modelo hacer predicciones sobre datos no vistos (Sarker, 2021).

2.2.3 Aprendizaje no supervisado

En el aprendizaje no supervisado, los algoritmos se entrenan utilizando datos no etiquetados, es decir, datos que no tienen asociadas etiquetas o respuestas conocidas. El objetivo es descubrir patrones, estructuras o relaciones intrínsecas en los datos, como agrupamientos o asociaciones entre variables (Shrifan et al., 2022).

2. 2.4 Aprendizaje semisupervisado

El aprendizaje semisupervisado combina elementos del aprendizaje supervisado y no supervisado, utilizando tanto datos etiquetados como no etiquetados para entrenar modelos (*Semi-Supervised_Learning_Chapelle_O._et_al._Eds._2006_Book_reviews*, n.d.). Esto puede ser útil cuando hay una gran cantidad de datos no etiquetados disponibles y etiquetar manualmente todos los datos sería costoso o impracticable (García-Martínez & Ventura, 2020).

2.2.5 Aprendizaje por refuerzo

En el aprendizaje por refuerzo, los algoritmos aprenden a través de la interacción con un entorno, recibiendo retroalimentación en forma de recompensas o castigos en función de las acciones que toman. El objetivo es aprender una política óptima que maximice la recompensa acumulada a lo largo del tiempo.

En resumen, el aprendizaje automático es un campo interdisciplinario que combina elementos de la estadística, la informática y la inteligencia artificial para desarrollar algoritmos y modelos que permitan a las computadoras aprender a partir de datos y realizar tareas específicas de forma autónoma.

2.3 Análisis predictivo

El análisis predictivo es un enfoque analítico que utiliza técnicas estadísticas y de aprendizaje automático para predecir eventos futuros o comportamientos basándose en datos históricos y

patrones identificados (Bellazzi & Zupan, 2008) (Gu et al., 2022). Aquí hay un marco conceptual para comprender mejor el análisis predictivo:

2.3.1 Datos

El análisis predictivo se basa en la recopilación y el análisis de datos. Estos datos pueden ser de diferentes tipos, como datos estructurados (por ejemplo, tablas de bases de datos) o datos no estructurados (por ejemplo, texto, imágenes o videos). La calidad y la relevancia de los datos son fundamentales para el éxito del análisis predictivo (Gupta & Rani, 2019).

2.3.2 Preprocesamiento de datos

Antes de realizar cualquier análisis predictivo, es necesario preprocesar los datos para limpiarlos, transformarlos y prepararlos para su análisis. Esto puede implicar la eliminación de valores atípicos, la imputación de datos faltantes, la normalización de variables y otras técnicas de preparación de datos(Chu et al., 2016).

2.3.3 Selección de características

En el análisis predictivo, es importante identificar las características o variables que son más relevantes para el problema en cuestión. La selección de características ayuda a reducir la dimensionalidad de los datos y a mejorar la precisión de los modelos predictivos (Chandrashekar & Sahin, 2014).

2.3.4 Modelado predictivo

El modelado predictivo implica la construcción y evaluación de modelos estadísticos o de aprendizaje automático que pueden predecir eventos futuros o comportamientos. Algunos de los algoritmos comunes utilizados en el modelado predictivo incluyen regresión lineal, regresión logística, árboles de decisión, máquinas de vectores de soporte (SVM), redes neuronales y métodos de ensamblaje, como bosques aleatorios y gradient boosting (*CS-GY 6923 Machine Learning - Kannan*, n.d.).

2.3.5 Validación del modelo

Una vez que se ha construido un modelo predictivo, es importante validar su rendimiento utilizando técnicas como la validación cruzada, la división de datos de entrenamiento y prueba, o el uso de conjuntos de validación independientes. La validación del modelo ayuda a garantizar que el modelo sea generalizable y capaz de realizar predicciones precisas en nuevos datos no vistos (Yadav & Shukla, 2016).

2.3.6 Implementación y despliegue

Una vez que se ha validado un modelo predictivo, se puede implementar en un entorno operativo para hacer predicciones en tiempo real. Esto puede implicar la integración del modelo en sistemas existentes, la automatización de procesos de predicción y la monitorización continua del rendimiento del modelo (Sculley et al., n.d.).

2.3.7 Evaluación y refinamiento

El análisis predictivo es un proceso iterativo en el que se debe evaluar continuamente el rendimiento del modelo y refinarse según sea necesario. Esto puede implicar el reentrenamiento del modelo con datos más recientes, la incorporación de nuevas características o la exploración de diferentes algoritmos y enfoques de modelado (Nie et al., n.d.).

En resumen, el análisis predictivo es un proceso complejo que combina técnicas de análisis de datos, modelado estadístico y aprendizaje automático para hacer predicciones sobre eventos futuros o comportamientos. Un marco conceptual claro y bien definido puede ayudar a guiar el proceso y garantizar resultados precisos y útiles.

Capítulo 3 Modelo predictivo

En esta sección se presenta la metodología propuesta para desarrollar el modelo predictivo que nos permitió identificar los rasgos de personalidad que influyen en la culminación de los estudios de posgrado por parte de los estudiantes.

La metodología se compone de tres fases principales: (1) Recolección de datos, (2) Preprocesamiento de datos, y (3) Construcción del modelo predictivo. La Figura 1 resume estas tres fases de manera detallada. En la fase de recolección de datos, se utilizaron encuestas y entrevistas para obtener información relevante sobre los estudiantes. El preprocesamiento de datos incluyó la limpieza y normalización de la información para asegurar su calidad y consistencia. Finalmente, la construcción del modelo predictivo se realizó aplicando técnicas de machine learning, con el objetivo de maximizar la precisión y la capacidad de generalización del modelo.

La base de datos se elaboró digitalizando la información de 24 generaciones (generación 2005 a la generación 2020), que corresponde a 272 estudiantes de maestría en el CENIDET. Las líneas de investigación de estos estudiantes son: Ingeniería de software, Cómputo Inteligente, Sistemas Distribuidos, Inteligencia Artificial y Sistemas Híbridos Inteligentes.

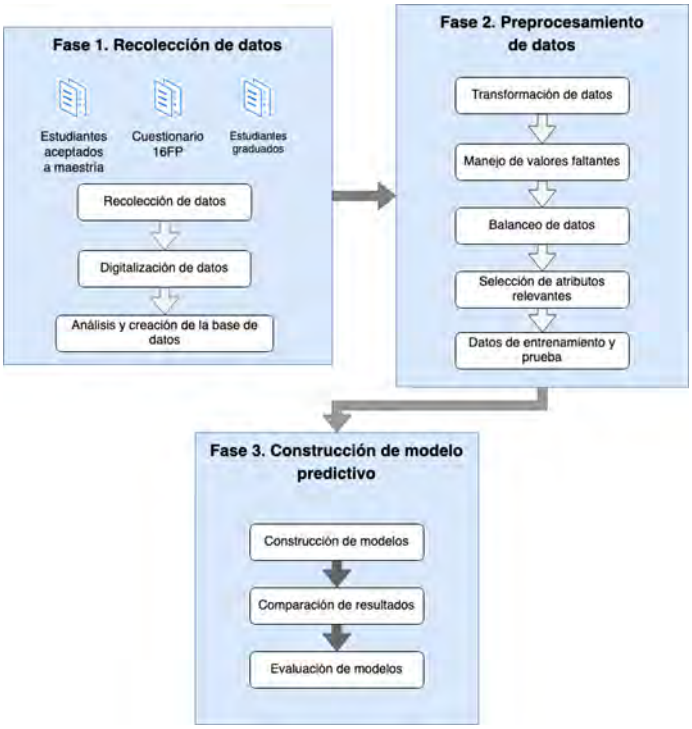


Figura 1 Fases de la Metodología de solución

3.1 Fase 1 Recolección de los datos

3.1.2 Recolección de los datos

La recolección de los datos es el núcleo de la primera fase de la metodología de solución propuesta. En esta fase se consideran tres fuentes de los datos:

1. la lista de estudiantes aceptados a la maestría por generación: en esta fuente de datos se tiene la información de cada generación, así como la lista de estudiantes aceptados en dicha generación, la fecha de inicio de la generación. Además de la línea de investigación en la que se encuentran los estudiantes.

2. el cuestionario de 16 factores de personalidad: El cuestionario aplicado a cada uno de los aspirantes tiene la siguiente información: Nombre del aspirante, fecha de aplicación del cuestionario, evaluación del Factor 1: expresividad emocional, evaluación del Factor 2: Inteligencia, Factor 3: Fuerza del yo, Factor 4: Dominancia, Factor 5: Impulsividad, Factor 6: Lealtad grupal, Factor 7: Aptitud situacional, Factor 8: Emotividad, Factor 9: Credibilidad, Factor 10: Actitud cognitiva, Factor 11: Sutileza, Factor 12: Conciencia, Factor 13: Posición social, Factor 14: Certeza individual, Factor 15: Autoestima, Factor 16: Estado de ansiedad. Las figuras 2 y 3 presentan el cuestionario de los 16 factores de personalidad. Su aplicación permite obtener un perfil detallado del individuo, facilitando intervenciones más efectivas.

REPORTE FINAL		
Fecha	24 de mayo 2018	Género:
Nombre:		**MUJER**
Especialic	Maestría en Ciencias de la Computación	

FACTOR/RANGO RECOMENDAD		CONCORDANCIA DEL PERFIL
A(+)	(3-7)	NO
B(+)	(5-10)	NO
C(+)	(4-8)	NO
E(-)	(3-7)	NO
F(-)	(3-7)	SI
G(-)	(2-7)	SI
H(+)	(4-8)	SI
I(+)	(3-7)	SI
L(-)	(2-6)	NO
M(-)	(3-7)	SI
N(-)	(2-6)	SI
O(-)	(1-6)	SI
Q1(-)	(1-5)	SI
Q2(+)	(5-7)	SI
Q3(+)	(6-10)	SI
Q4(+)	(3-7)	SI

FACTORES COINCIDENTES N/16	PORCENTAJE DE SIMILITUD %	PRONÓSTICO
11	68.7 %	CANDIDATA REGULAR

Observaciones

Pensamiento abstracto ligeramente por debajo de la media. Sin dificultades de integración, con madurez emocional, actitud de servicio y buena disposición hacia el trabajo. Flexible frente a las adversidades, poco precavida, respetuosa y conservadora. Muestra también seguridad en sí misma y persistencia en sus objetivos.

Figura 2 Cuestionario de 16 factores de personalidad

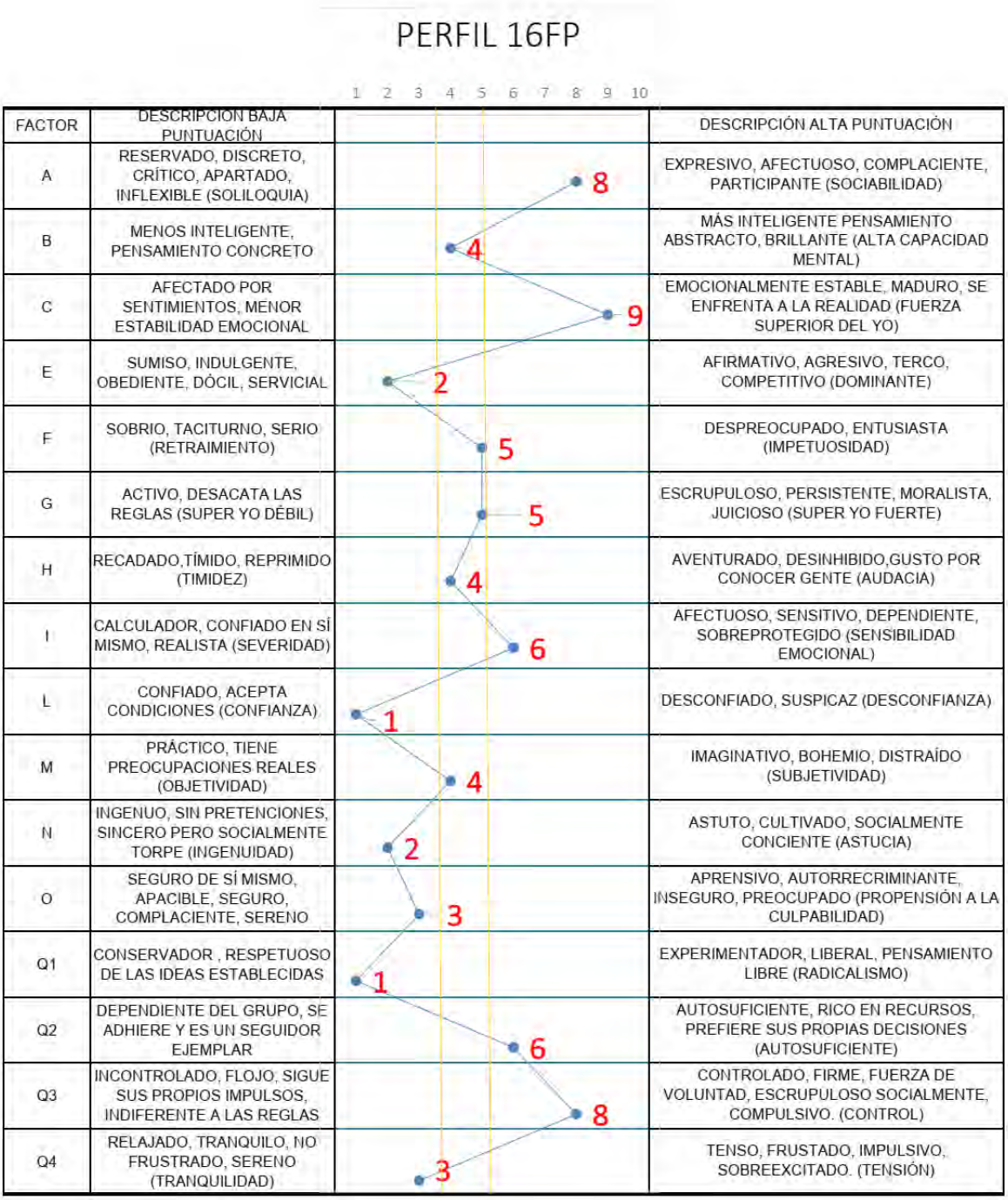
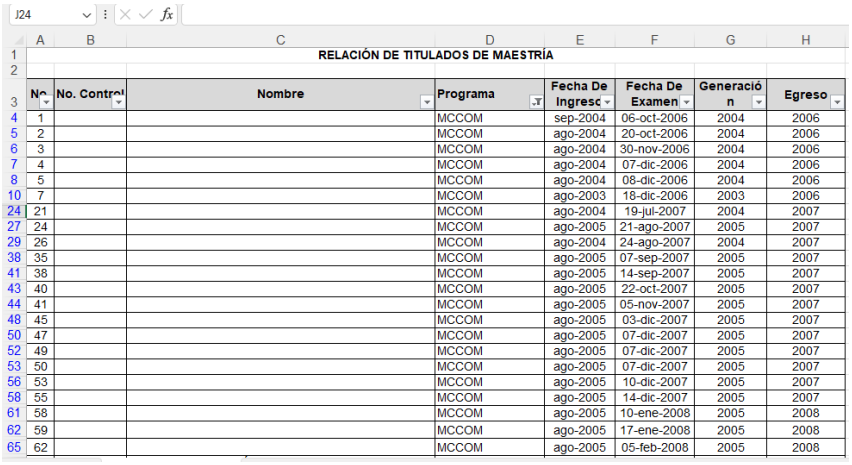


Figura 3 Perfiles del cuestionario de 16 Factores de personalidad

Es importante resaltar que esta fuente de datos corresponde a XX generaciones del departamento de Ciencias Computacionales, quienes nos proporcionaron esta información.

3. la lista de estudiantes graduados. Los graduados son aquellos que han completado el programa de maestría. Esta información contiene el nombre del estudiante, el tipo de programa (maestría), la fecha de ingreso, la fecha del examen, la generación, y la fecha de egreso. Esta lista es una herramienta esencial para verificar la trayectoria educativa y profesional de los graduados. La información recopilada también es utilizada para fines estadísticos y de acreditación institucional lista de estudiantes graduados. La figura 4 presenta los datos de estos estudiantes.



RELACIÓN DE TITULADOS DE MAESTRÍA

No	No. Control	Nombre	Programa	Fecha De Ingreso	Fecha De Examen	Generación	Egreso
1			MCCOM	sep-2004	06-oct-2006	2004	2006
2			MCCOM	ago-2004	20-oct-2006	2004	2006
3			MCCOM	ago-2004	30-nov-2006	2004	2006
4			MCCOM	ago-2004	07-dic-2006	2004	2006
5			MCCOM	ago-2004	08-dic-2006	2004	2006
7			MCCOM	ago-2003	18-dic-2006	2003	2006
21			MCCOM	ago-2004	19-jul-2007	2004	2007
24			MCCOM	ago-2005	21-ago-2007	2005	2007
26			MCCOM	ago-2004	24-ago-2007	2004	2007
35			MCCOM	ago-2005	07-sep-2007	2005	2007
38			MCCOM	ago-2005	14-sep-2007	2005	2007
40			MCCOM	ago-2005	22-oct-2007	2005	2007
41			MCCOM	ago-2005	05-nov-2007	2005	2007
45			MCCOM	ago-2005	03-dic-2007	2005	2007
47			MCCOM	ago-2005	07-dic-2007	2005	2007
49			MCCOM	ago-2005	07-dic-2007	2005	2007
50			MCCOM	ago-2005	07-dic-2007	2005	2007
53			MCCOM	ago-2005	10-dic-2007	2005	2007
55			MCCOM	ago-2005	14-dic-2007	2005	2007
58			MCCOM	ago-2005	10-ene-2008	2005	2008
59			MCCOM	ago-2005	17-ene-2008	2005	2008
62			MCCOM	ago-2005	05-feb-2008	2005	2008

Figura 4 Relación de estudiantes titulados de maestría.

En la Figura 5 se presenta un resumen detallado de las variables recolectadas de diversas fuentes de información, proporcionando una visión integral de los datos recopilados. Estas variables han sido seleccionadas cuidadosamente para asegurar una cobertura amplia y representativa de los aspectos relevantes del estudio. Además, se ha llevado a cabo un riguroso proceso de verificación para garantizar la precisión y la fiabilidad de la información. Este análisis exhaustivo permite una comprensión profunda de las tendencias y patrones observados.

Fuente de información	Variables recolectadas
Cuestionario 16FP	• Nombre del aspirante • Fecha de aplicación del cuestionario • Cuestionario 16FP o Expresividad emocional o Inteligencia o Fuerza del yo o Dominancia o Impulsividad o Lealtad grupal o Aptitud situacional o Emotividad o Credibilidad o Actitud cognitiva o Sutileza o Conciencia o Posición social o Certeza individual o Autoestima o Estado de ansiedad
Alumnos aceptados por generación	• Nombre del estudiante • Especialidad • Posgrado • Generación • Fecha de ingreso al posgrado
Alumnos graduados	• Nombre del estudiante • Especialidad • Posgrado • Fecha de graduación

Figura 5Fuentes de información recolectadas para el análisis de los datos.

3.1.2 Digitalización de los datos

La digitalización de datos es el proceso mediante el cual la información física o analógica se transforma en formato digital, permitiendo su almacenamiento, procesamiento y transmisión a través de medios electrónicos. Este proceso involucra la conversión de documentos en papel, imágenes, audio y video a formatos digitales que pueden ser manipulados y gestionados por sistemas informáticos. La base de datos se elaboró digitalizando la información de 24 generaciones pasadas en las cuales tuvimos los datos completos de cada estudiante como se muestra en la Figura 5. En total obtuvimos 272 registros de estas tres fuentes de datos. La digitalización facilita la accesibilidad, búsqueda y análisis de datos de manera eficiente, contribuyendo a la organización y preservación de la información.

3.1.3 Análisis y creación de la base de datos

En la investigación psicológica, la selección cuidadosa de atributos en una base de datos es crucial para garantizar análisis significativos y válidos. Esta tarea implica identificar variables que sean teóricamente relevantes para el fenómeno bajo estudio y que proporcionen información confiable y válida. Los atributos seleccionados deben estar alineados con los objetivos específicos del estudio y considerar la disponibilidad y calidad de los datos disponibles. La tabla 1 muestra como se ha desarrollado una base de datos que incluye 20 atributos clave, tales como

variables demográficas, medidas psicológicas fundamentales y otras variables pertinentes, diseñada para explorar relaciones complejas y abordar preguntas específicas dentro del campo de la psicología.

1)Especialidad	6)Dominancia	11)Credibilidad	16)Certeza_individual
2)Sexo	7)Impulsividad	12)Actitud_cognitiva	17)Autoestima
3)Expresividad_emocional	8)Lealtad_grupal	13)Sutileza	18)Estado_de_ansiedad
4)Inteligencia	9)Aptitud_situacional	14)Conciencia	19)Factores_coincidentes
5)Fuerza_del_yo	10)Emotividad	15)Posicion_social	20)Pronostico

Tabla 1 Atributos seleccionados

3.1.4 Clasificación de las clases

Simplicidad y claridad: Al dividir los datos en dos categorías bien definidas, como "aprobado" o "rechazado", se facilita su interpretación y análisis. Este enfoque binario simplifica la toma de decisiones al ofrecer una distinción clara entre las opciones disponibles. Además, proporciona una visualización más comprensible de los resultados, lo que facilita el análisis para los especialistas y responsables de la toma de decisiones. En situaciones que requieren una clasificación rápida y precisa, esta metodología es particularmente eficaz. La división en dos clases también reduce la complejidad en el modelado al centrarse en dos resultados específicos y claramente diferenciables(Leevy et al., 2023).

Este método es ampliamente utilizado en escenarios donde la categorización de los datos en dos posibles estados es fundamental para la evaluación y el control. Por ejemplo, en la detección de fraudes o en diagnósticos médicos, clasificar los resultados en categorías binarias permite una respuesta rápida y eficaz. Además, este enfoque puede ayudar a identificar patrones y tendencias que podrían ser menos evidentes en una clasificación más compleja. La claridad y la simplicidad del análisis binario también facilitan la comunicación de resultados a personas sin formación técnica, haciendo que los hallazgos sean más accesibles y comprensibles.

De manera similar, en la base de datos se clasificaron las clases en dos categorías: "TERMINA" y "NO TERMINA". Esta clasificación se realizó con el objetivo de garantizar resultados más precisos. La separación en estas dos clases permite un análisis más claro y específico de los datos, facilitando la identificación de patrones y tendencias relevantes. Además, esta categorización ayuda a mejorar la precisión de los modelos predictivos al proporcionar una distinción clara entre los diferentes estados de los datos. La implementación de estas categorías contribuye a una

evaluación más efectiva y a la obtención de insights más fiables. Cuando se habla de insights, se refiere a la comprensión profunda y clara de algo, obtenida a partir del análisis de datos, información o experiencia. En el contexto del análisis de datos y el machine learning, los insights son las conclusiones o hallazgos significativos que se extraen al examinar los datos. Estos insights pueden revelar patrones, tendencias o relaciones ocultas que son útiles para tomar decisiones informadas, desarrollar estrategias o mejorar procesos(Mach-Król & Hadasik, 2021). La tabla 2 muestra cómo se clasificaron las categorías en nuestra base de datos. La tabla 2 ilustra el proceso de categorización y la distribución resultante. Este enfoque permite una mejor organización y análisis de la información almacenada.

Clase	Numero de Instancias
TERMINA	213
NOTERMINA	59
Total	272

Tabla 2 Clasificación de clases de la base de datos.

3.2 Fase 2 Preprocesamiento de los datos

El tratamiento de datos es la conversión de datos en una forma utilizable y deseada llamada conjunto de datos. Los datos sin procesar son muy susceptibles al ruido, los valores faltantes y la incoherencia. La calidad de los datos afecta los resultados de la minería de datos. Para mejorar la calidad de los datos y, en consecuencia, de los resultados de la minería, los datos sin procesar se procesan previamente para mejorar la eficiencia y la facilidad del proceso de minería. En este trabajo de investigación, se mejoró la calidad de los datos sin procesar mediante el uso de un procedimiento de tratamiento que incluye:

3.2.1 Transformación de datos:

Esta tarea consistió en transformar los atributos de tipo carácter a atributos de tipo numérico, excluyendo la clase objetivo. La transformación de datos es un proceso fundamental en el análisis de datos que implica convertir la información de su formato original a un formato adecuado para su procesamiento y análisis. Esta tarea consiste en transformar los atributos de tipo carácter a atributos de tipo numérico, excluyendo la clase objetivo. Como resultado, se genera un conjunto de datos transformados. Este proceso incluye una variedad de técnicas, como la limpieza de datos, la normalización y estandarización, la codificación de variables categóricas y la generación de nuevas variables. Los beneficios de la transformación de datos son numerosos: mejora la calidad de los datos al asegurar su precisión y consistencia, facilita el análisis al adecuar los datos a técnicas estadísticas y de modelado, y mejora el rendimiento de los modelos de aprendizaje automático. Además, estandarizar los datos hace que sean más accesibles y comparables, permitiendo una integración más efectiva de diferentes fuentes de datos. En resumen, la transformación de datos es esencial para obtener resultados precisos y significativos en cualquier proyecto de análisis de datos. Como resultado, se genera un conjunto de datos transformados.

Porcentaje de similitud: El porcentaje de similitud del Cuestionario de 16 Factores de Personalidad (16PF) se refiere a una medida utilizada para evaluar la consistencia de las respuestas proporcionadas por un individuo a lo largo del cuestionario. El 16PF, desarrollado por Raymond B. Cattell, evalúa 16 dimensiones básicas de la personalidad y se usa ampliamente en la psicología y recursos humanos para comprender mejor la personalidad de una persona.

El porcentaje de similitud generalmente se calcula para detectar patrones de respuesta que podrían indicar falta de atención, deshonestidad, o respuestas al azar por parte del respondiente. Este porcentaje compara las respuestas a preguntas similares o relacionadas dentro del

cuestionario. En esta actividad se le asigno un número del 1 al 4, cono se muestra en la tabla No 3 dependiendo cuantos atributos son similares al que requiere para hacer una maestría se le asigna una de las cuatro categorías; 1: Excelente candidato, 2: Buen candidato, 3: Candidato regular y 4: no cubre el perfil.

Porcentaje de similitud
Excelente candidato = 1
Buen candidato = 2
Candidato regular = 3
No cubre el perfil = 4;

Tabla 3 Asignación de números a al atributo porcentaje de similitud

Especialidad: La especialidad hace referencia al área al que pertenece cada estudiante dentro del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), que está dividida en cinco especialidades y cada diciplina se enfoca a desarrollar un conocimiento profundo de habilidades avanzadas. Este enfoque especializado permite al investigador contribuir de manera significativa a su campo mediante la generación de nuevos conocimientos, teorías y aplicaciones prácticas.

Las especialidades se clasificaron en cinco categorías, numeradas del 1 al 5, de la siguiente manera: 1) Ingeniería de Software, 2) Sistemas Distribuidos, 3) Inteligencia Artificial, 4) Cómputo Inteligente y 5) Sistemas Híbridos, así como se muestra en la tabla 4. Cada categoría representa un área distinta de investigación y desarrollo dentro del campo de la computación. Esta clasificación permite organizar y enfocar los recursos y esfuerzos de investigación de manera más efectiva.

Especialidad
IS =1
SD=2
IA=3

CI=4
SHI=5

Tabla 4 Clasificación de especialidades

Sexo: El sexo, como categoría biológica que clasifica a los individuos en masculino y femenino, desempeña un papel crucial en la evaluación de la personalidad mediante el Cuestionario de 16 Factores (16PF). Esta herramienta psicométrica, desarrollada por Raymond B. Cattell, se utiliza ampliamente para explorar y medir diversas características de la personalidad. Cada uno de estos factores puede manifestarse de manera distinta según el género del individuo. La inclusión del sexo como variable en el análisis de los resultados del 16PF no solo permite capturar estas variaciones, sino también contextualizar las interpretaciones de los rasgos de personalidad en función de las diferencias biológicas y socioculturales entre hombres y mujeres. Por lo tanto, se clasificó como 1 para femenino y 0 para masculino. Esta codificación binaria facilita el análisis estadístico y la interpretación de datos en estudios que consideran el sexo como variable clave. Además, permite realizar comparaciones significativas entre grupos de género y entender mejor cómo las diferencias biológicas pueden influir en los resultados del Cuestionario de 16 Factores de Personalidad. Esta metodología estándar asegura consistencia y precisión en la investigación psicométrica y aplicada. La tabla 5 muestra la clasificación del sexo.

Sexo
Femenino =1
Masculino=0

Tabla 5 Clasificación del sexo

3.2.2 Tratamiento de valores faltantes:

La gestión de datos faltantes es una etapa crítica en cualquier análisis de datos, ya que la ausencia de información puede afectar la validez y la fiabilidad de los resultados. Los datos faltantes pueden surgir por diversas razones, como errores en la recolección de datos, respuestas omitidas en encuestas, o problemas técnicos durante la transferencia o almacenamiento de datos. En cualquier caso, es fundamental abordar estos datos de manera adecuada para evitar sesgos y garantizar la precisión de las conclusiones derivadas del análisis.

El tratamiento de datos faltantes implica estrategias y técnicas diseñadas para manejar este tipo de situaciones de manera efectiva. Esto puede incluir desde métodos sencillos como la eliminación de observaciones con datos faltantes hasta enfoques más avanzados como la imputación de valores basada en modelos estadísticos. Cada método tiene sus ventajas y limitaciones, y la elección del enfoque adecuado dependerá del contexto específico del estudio, la naturaleza de los datos y el impacto potencial en los resultados. Normalmente, los datos no están limpios esto hace referencia a la presencia de problemas como valores corruptos u omisos, inconsistencias, errores de formato, duplicados o cualquier otra anomalía que pueda afectar la integridad y la calidad de los datos. Para el tratamiento de los valores faltantes se utilizó la imputación simple para completar un dato faltante. Una de las técnicas que se utilizó fue imputación por la mediana, que implica utilizar la mediana de los datos para completar un dato faltando.

Se optó por utilizar la técnica de imputación por la mediana, la cual consiste en completar un dato faltante utilizando el valor mediano de los datos. En términos de porcentajes, tenemos que el 97.4% de las observaciones están completas, mientras que 2.57% de las observaciones les faltan valores para una o dos variables esto con base en el artículo de (Papageorgiou et al., 2018)se destaca que un pequeño porcentaje de datos faltantes (menos del 5%) es generalmente considerado como manejable y no suele impactar significativamente los resultados del análisis. Asimismo, menciona que la proporción de datos faltantes debe ser cuidadosamente monitoreada, ya que incluso pequeños porcentajes pueden influir en la precisión del análisis dependiendo del contexto del estudio. La figura 6 muestra los atributos con datos faltantes.

Esta elección se fundamenta en que otras técnicas como la media, el promedio y la regresión arrojaban valores decimales, mientras que la imputación por mediana proporciona un valor entero. Esta característica es especialmente relevante para asegurar la consistencia y la interpretabilidad de los datos, evitando así la introducción de precisiones innecesarias o distorsiones en los resultados finales del análisis.

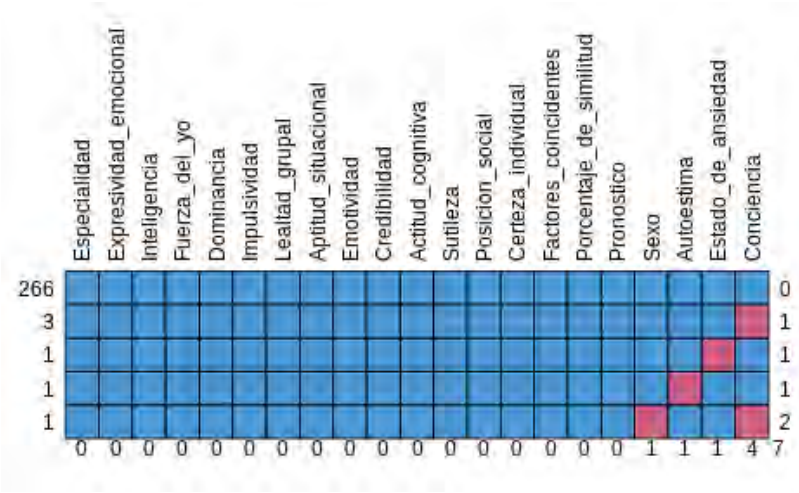


Figura 6 Atributos con datos faltantes

3.2.3 Balanceo de clases

En el análisis de datos y el aprendizaje automático, el balanceo de clases es fundamental cuando se trabaja con conjuntos de datos donde algunas clases están subrepresentadas. Aplicar técnicas de balanceo de clases mejora el rendimiento del modelo, ya que los algoritmos tienden a sesgarse hacia las clases mayoritarias, resultando en una baja precisión para las minoritarias. Al balancear las clases, se asegura que el modelo reconozca y clasifique correctamente todas las clases, aumentando así su equidad y precisión. Esto es especialmente importante en aplicaciones críticas como la detección de fraudes y el diagnóstico médico. Además, permite una evaluación más completa del rendimiento del modelo utilizando métricas como la precisión, el recall y la F1-score, en lugar de depender únicamente de la precisión global, que puede ser engañosa en datos desequilibrados(Singh Rawat & Kumar Mishra, n.d.) (Werner de Vargas et al., 2023). Este paso se realiza cuando alguna de las clases del conjunto de datos tiene una cantidad mucho mayor que el resto de las clases, lo cual las desbalancea. Este trabajo considera el caso de conjuntos de datos que solamente tiene cuatro clases y una de ellas cuenta con una mayor cantidad de ejemplos que las otras, específicamente cuando están altamente desproporcionados. Para resolver el problema de desequilibrio de clases se utilizó la técnica SMOTE (Synthetic Minority Over- Sampling Technique). Este algoritmo para cada ejemplo de la clase minoritaria introduce ejemplos sintéticos en la línea que une al elemento con su k vecinos más cercanos. Los nuevos objetos se generan por medio de diferencias entre el objeto y su vector de características considerando sus vecinos más cercanos. Cada diferencia se multiplica por cero o uno y los vectores de características diferentes de cero son considerados como nuevos objetos sintéticos(Chawla et al., 2002). La figura 7 muestra un ejemplo gráfico de la implementación de SMOTE donde muestra

cómo SMOTE equilibra las clases desbalanceadas. Visualización del impacto de SMOTE en el conjunto de datos.

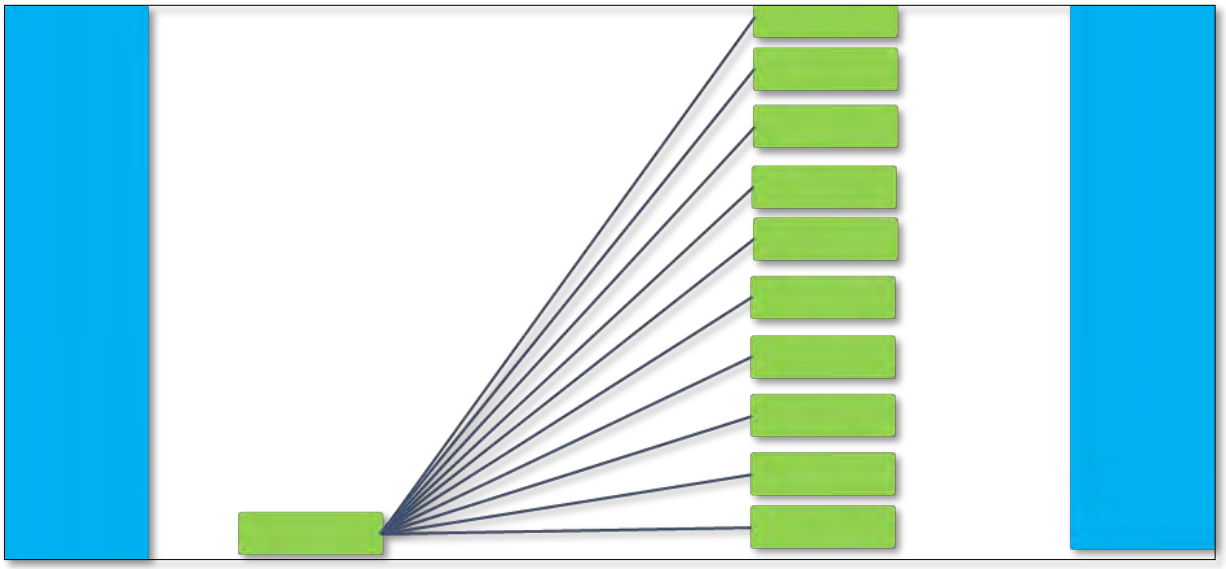


Figura 7 Ejemplo Gráfico de la Implementación de SMOTE.

En la Tabla 3, la primera columna presenta las clases del conjunto de datos. La segunda columna indica el número de instancias de cada clase. La tercera columna muestra el número de instancias obtenidas tras aplicar la técnica SMOTE para el balanceo de clases. Esta técnica es fundamental para abordar el problema del desbalanceo de clases en conjuntos de datos, asegurando una mejor representación de cada clase. El aumento de instancias minoritarias ayuda a mejorar la precisión y la generalización de los modelos de aprendizaje automático. Se observa un incremento significativo en el número de instancias en las clases previamente subrepresentadas. Además, el balanceo de clases contribuye a una distribución más equitativa de los datos, lo que es crucial para el entrenamiento eficaz de los modelos predictivos. Finalmente, la aplicación de SMOTE mejora el rendimiento y la estabilidad del modelo, reduciendo el sesgo hacia las clases mayoritarias.

Clase	*Núm. De instancias*	*Número instancias después de aplicar SMOTE*
TERMINA	*202*	*202*
NOTERMINA	*50*	*202*
TOTAL	*252*	*404*

Tabla 6 Numero de instancias aplicando SMOTE.

3.2.4 Selección de atributos relevantes

La selección de atributos permitió identificar un subconjunto relevante que se utilizó en la construcción del modelo predictivo. Esta etapa es crucial, ya que elimina características redundantes o irrelevantes que podrían afectar negativamente el rendimiento del modelo(Tang et al., n.d.). Al enfocarse únicamente en los atributos más significativos, se mejora la eficiencia del modelo, reduciendo el tiempo de procesamiento y la complejidad computacional. Además, al trabajar con un conjunto de datos más depurado, se incrementa la interpretabilidad del modelo, facilitando la comprensión de cómo cada atributo influye en las predicciones. Esta técnica también ayuda a prevenir el sobreajuste, asegurando que el modelo generalice bien con datos nuevos. En consecuencia, el proceso de selección de atributos contribuye a crear modelos más robustos y precisos. La metodología aplicada para seleccionar estos atributos incluye análisis estadísticos y técnicas de minería de datos avanzadas, garantizando que solo las características más impactantes sean consideradas(Balogun et al., 2020; Werner de Vargas et al., 2023). Este enfoque metodológico asegura que el modelo predictivo tenga un rendimiento óptimo en diversas condiciones y escenarios.

En este trabajo de investigación se utilizaron tres métodos para llevar a cabo la selección de atributos.

3.2.5 CorrelationAttributeEval

El método CorrelationAttributeEval evalúa la correlación (Pearson) entre cada atributo y la clase objetivo. Se consideran relevantes aquellos atributos con una correlación positiva o negativa moderada a alta (aproximadamente cercana a -1 o 1), utilizando 0.2 como punto de corte para determinar la relevancia(Li et al., 2017). Los atributos que no muestran correlación significativa no se consideran relevantes para el modelo. Este enfoque permite identificar características que potencialmente influyen en las predicciones del modelo, proporcionando una selección inicial de atributos basada en su relación con la variable objetivo(Yu & Liu, n.d.).

Además, al establecer un umbral claro de correlación, se simplifica la tarea de identificar qué atributos contribuyen de manera significativa a la capacidad predictiva del modelo. Esta técnica es especialmente útil en conjuntos de datos complejos donde es crucial filtrar atributos redundantes o irrelevantes para mejorar la precisión y la interpretación del modelo(Whare W-ananga et al., 1999). Al enfocarse en atributos con correlaciones más fuertes, se optimiza el

proceso de construcción del modelo, reduciendo el ruido y aumentando la robustez frente a datos nuevos o variaciones en el conjunto de datos original. Sin embargo, no muestra atributos relevantes porque no hay una correlación entre ellos.

3.2.6 Classifiersubsetevaluator

El método ClassifierSubsetEval se utiliza para evaluar subconjuntos de características en el contexto de clasificación. Este enfoque implica seleccionar y evaluar la efectividad de diferentes subconjuntos de atributos utilizando un clasificador determinado. Los subconjuntos se consideran relevantes si logran mejorar la precisión del clasificador, reducir el error y aumentar la capacidad de generalización del modelo. Durante este proceso, se analizan múltiples combinaciones de atributos para identificar el conjunto óptimo que maximiza el rendimiento del clasificador. Este método es especialmente útil en conjuntos de datos complejos, donde la inclusión de características redundantes o irrelevantes puede afectar negativamente el rendimiento del modelo(Čehovin & Bosnić, 2010). Los estudios empíricos han demostrado que ClassifierSubsetEval puede mejorar significativamente la precisión y robustez de los modelos clasificadores, enfocándose en los atributos más informativos y descartando aquellos que no aportan de manera significativa a la predicción.

3.2.7 Componentes principales o PCA

Se utilizó una aproximación basada en componentes principales para reducir la dimensionalidad del conjunto de características, resultando en la selección de 11 atributos relevantes de un total inicial de 22. Esta técnica permitió optimizar la eficiencia del modelo al eliminar atributos redundantes y centrarse en aquellos que más impactan en las predicciones. La selección cuidadosa de estos atributos también contribuyó significativamente a mejorar la interpretabilidad del modelo, facilitando la comprensión de cómo cada atributo afecta las decisiones predictivas(Jollife & Cadima, 2016)(Yao et al., 2024).

Además, el uso de componentes principales aseguró que los atributos seleccionados capturaran la mayor varianza en los datos originales, preservando al mismo tiempo la información crítica para las predicciones. Esta estrategia no solo redujo la complejidad del modelo, sino que también ayudó a mitigar posibles problemas de sobreajuste al trabajar con un conjunto de datos más depurado. El proceso de selección de atributos se llevó a cabo mediante análisis detallados de las cargas de componentes y la varianza explicada, garantizando así la representación óptima de las características más relevantes.

Como resultado, la implementación de la aproximación por componentes principales no solo mejoró el rendimiento del modelo predictivo, sino que también proporcionó una base sólida para futuros análisis y refinamientos en la aplicación de técnicas de aprendizaje automático.

En este modelo evaluó 11 atributos y se tomó la decisión de que este método nos favorecía, ya que los datos que tomó fueron 11 de los 16FP. Ver Tabla 7.

Atributos relevantes
1. Expresividad emocional
2. Fuerza del yo
3. Impulsividad
4. Aptitud situacional
5. Emotividad
6. Sutileza
7. Posición social
8. Autoestima
9. Estado de ansiedad
10. Factores coincidentes
11. Pronóstico

Tabla 7 Atributos relevantes.

3.2.8 Datos de entrenamiento y prueba

Para evitar modelos sesgados y estimaciones demasiado optimistas, la división de los datos de entrenamiento y prueba se realizó después del equilibrio de clases y la extracción de características. En este estudio, el conjunto de datos se dividió en un 80% para entrenamiento y un 20% para pruebas. Este enfoque asegura que el modelo no solo aprenda adecuadamente durante la fase de entrenamiento, sino que también se evalúe de manera justa y precisa durante la fase de prueba. La correcta división y preparación de los datos es fundamental para la construcción de modelos robustos y fiables, como se destaca en estudios sistemáticos sobre la evaluación de modelos de aprendizaje automático ((Benedick et al., 2021) (Sardinha et al., n.d.).

Esto asegura que el seleccionador o clasificador de características no tenga acceso a los datos de prueba. Se implementó una validación cruzada de 5 veces, utilizando una proporción de 1/5 para evaluar los datos. Este método permite que cada subconjunto del conjunto de datos sea

usado tanto para entrenamiento como para prueba en diferentes iteraciones, garantizando una evaluación exhaustiva del modelo. Además, la validación cruzada ayuda a mitigar el riesgo de sobreajuste al proporcionar una medida más robusta del rendimiento del modelo en datos no vistos. Esta técnica es especialmente útil para obtener estimaciones confiables y generalizables del rendimiento del modelo, asegurando que los resultados sean consistentes y precisos(Kohavi, n.d.) (Kohavi, 1995).

3.3 Fase 3 Construcción del modelo predictivo

El aprendizaje computacional se utilizó inicialmente para crear técnicas que permitieran a las computadoras aprender. Actualmente, debido a que incorpora una serie de métodos estadísticos avanzados para la regresión y la clasificación, se aplica en una amplia gama de campos. Estos incluyen diagnósticos médicos, detección de fraudes con tarjetas de crédito, reconocimiento facial y de voz, y análisis del mercado de valores (Mitchell, 2006).

El modelo para predecir la eficiencia terminal de un estudiante de maestría en CENIDET se desarrolló mediante la reducción de características y el equilibrio de clases. Para este propósito, se utilizó la técnica de Análisis de Componentes Principales (PCA), la cual describe un conjunto de datos en términos de nuevas variables («componentes») no correlacionadas. Estos componentes se ordenan según la cantidad de varianza original que explican, lo que hace a PCA útil para reducir la dimensionalidad de un conjunto de datos. Esta técnica permitió seleccionar 11 de los 20 atributos iniciales. Una vez seleccionadas las variables, se procedió a entrenar el modelo utilizando distintos clasificadores.

Se evaluaron varios algoritmos de clasificación, incluyendo árboles de decisión, máquinas de vectores de soporte, Naive Bayes, AdaBoost, y k-NN (k-Nearest Neighbors), para determinar cuál ofrecía el mejor rendimiento (Witten et al., n.d.). La validación cruzada se utilizó para asegurar la robustez del modelo, y se aplicaron técnicas de sobremuestreo y submuestreo para manejar el desequilibrio en las clases. Los resultados mostraron que el clasificador basado k-NN (k-Nearest Neighbors) proporcionó la mayor precisión. Finalmente, el modelo fue evaluado en un entorno real, confirmando su efectividad para predecir la eficiencia terminal de los estudiantes.

Los experimentos se realizaron con seis de diez algoritmos de minería de datos principales identificados por la Conferencia internacional de minería de datos (Witten et al., n.d.). Una vez que se seleccionaron las variables se procedió a hacer el entrenamiento con los diferentes algoritmos de clasificación.

3.3.1 Tres.j48

El algoritmo J48, implementado en WEKA, es una versión del C4.5 utilizada para clasificación. Este algoritmo selecciona el atributo que mejor divide los datos usando la ganancia de información y los divide recursivamente en subconjuntos hasta que se logran clases puras o se

cumple un criterio de parada. Construye un árbol de decisión donde cada nodo representa un atributo y cada hoja una clase. Para evitar el sobreajuste, realiza una poda eliminando ramas insignificantes. J48 es comprensible, eficiente y maneja datos faltantes, aunque puede sobreajustarse y sesgarse hacia atributos con muchos valores. En WEKA, el usuario carga el conjunto de datos, selecciona J48, configura sus parámetros, lo ejecuta y evalúa los resultados (Witten et al., n.d.) (Ross Quinlan et al., 1994).

3.3.2 Random Forest

El algoritmo Random Forest, concebido por Leo Breiman y Adele Cutler, se emplea ampliamente para tareas de clasificación y regresión. Construye múltiples árboles de decisión utilizando subconjuntos aleatorios del conjunto de datos de entrenamiento, aplicando el método de bagging para reducir la varianza y evitar el sobreajuste. Durante la construcción de cada árbol, selecciona aleatoriamente un subconjunto de atributos en cada división para aumentar la diversidad entre los árboles. La predicción final se determina por mayoría de votos (clasificación) o promedio de predicciones (regresión) de todos los árboles. Random Forest es conocido por su precisión y capacidad para manejar datos faltantes, aunque puede ser más complejo de interpretar que un solo árbol de decisión debido a su naturaleza combinada de múltiples modelos (Breiman, 2001a).

3.3.3 Naive Bayes

Naive Bayes es un algoritmo de clasificación que se basa en el teorema de Bayes con una suposición simplificada de independencia entre los atributos dados los valores de clase. Es conocido por su eficiencia y facilidad de implementación, especialmente en conjuntos de datos grandes. Este método asume que los atributos son independientes entre sí después de condicionar en la clase, lo que puede no ser realista, pero permite una implementación rápida y resultados aceptables en muchas aplicaciones prácticas. Naive Bayes tiene variantes como Gaussian Naive Bayes para atributos numéricos y Multinomial Naive Bayes para datos discretos como conteos de palabras. Aunque sensible a la suposición de independencia, Naive Bayes generalmente ofrece buena generalización y puede manejar datos con valores faltantes (G. H. John, n.d.).

3.3.4 SMO (Sequential Minimal Optimization)

SMO (Sequential Minimal Optimization) es un algoritmo diseñado para entrenar clasificadores SVM (Support Vector Machine), centrado en la optimización eficiente de problemas cuadráticos

durante el proceso de entrenamiento. Desarrollado por John Platt, SMO aborda estos problemas dividiéndolos en subproblemas más pequeños y resolviéndolos de manera secuencial y analítica. Durante cada iteración, selecciona pares de variables duales para optimizar, siguiendo reglas heurísticas para maximizar el progreso hacia la solución óptima. Esto permite obtener modelos SVM con alta precisión en la clasificación de datos, especialmente en problemas de clasificación binaria lineal y no lineal, y es particularmente eficaz en grandes conjuntos de datos debido a su enfoque eficiente de optimización secuencial (Aregbesola & Griva, 2022).

3.3.5 AdaBoost

El algoritmo AdaBoost ha evolucionado significativamente en el campo del aprendizaje computacional, inicialmente desarrollado para mejorar la precisión de los modelos de clasificación mediante la combinación de múltiples clasificadores débiles. Hoy en día, AdaBoost encuentra aplicaciones en una variedad de campos, desde diagnósticos médicos hasta análisis de mercado, gracias a su capacidad para manejar conjuntos de datos desequilibrados y mejorar la generalización del modelo (Ayaz, Shaukat, & Anjum, 2019), (Nissar et al., 2019).

3.3.69 KNN

KNN (K-Nearest Neighbors) es un algoritmo de aprendizaje supervisado utilizado para clasificación y regresión. Funciona identificando las k instancias más cercanas a un nuevo punto de datos según una medida de distancia, como la euclidiana. La predicción se realiza basándose en la mayoría de las clases (clasificación) o promedio de valores (regresión) de estas instancias cercanas. Es simple de entender e implementar, no hace suposiciones sobre la distribución de los datos, y el valor de k es crucial para su rendimiento. Sin embargo, puede ser computacionalmente costoso con grandes conjuntos de datos y es sensible a la escala de los datos y características irrelevantes (Jiang et al., 2012).

La Tabla 8 muestra la comparación de los algoritmos de clasificación, analizando la correcta e incorrecta clasificación de las instancias, el estadístico Kappa, el error medio absoluto medio, el error cuadrático medio, error relativo porcentual y el error relativo cuadrático medio.

	trees.J48	RandomForest	Naive Bayes	SMO	k-nn	AdaBoost
Correctly Classified Instances	72.28%	88.12%	61.63%	60.89%	80.69%	**62.13%**
Incorrectly Classified Instances	27.72%	11.88%	38.37%	39.11%	19.31%	**37.87%**
Kappa statistic	0.4455	0.7624	0.2327	0.2178	0.6139	**0.2426**
Mean absolute error	0.3065	0.2891	0.4238	0.3911	0.1947	**0.4409**
Root mean squared error	0.4967	0.3352	0.4936	0.6254	0.4382	**0.4706**
Relative absolute error	61.30%	57.83%	84.77%	78.22%	38.95%	**88.18%**
Root relative squared error	99.34%	67.05%	98.71%	125.07%	87.64%	**94.12%**
Total Number of Instances			**404**			

Tabla 8 Comparación de resultados de los modelos creados.

Para evaluar el desempeño de cada modelo creado se usaron las métricas de evaluación como se muestran en la Tabla 9.

Algoritmo	Correctly Classified Instances	Incorrectly Classified Instances	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Clase
J48	0.909	0.556	0.667	0.909	0.769	0.406	0.662	0.65	TERMINA
	0.44	0.091	0.8	0.444	0.571	0.406	0.662	0.693	NOTERMINA
	0.7	0.346	0.727	0.7	0.68	0.406	0.662	0.67	Weighted Avg.
Random Forest	0.906	0.144	0.863	0.906	0.884	0.763	0.941	0.922	TERMINA
	0.856	0.094	0.901	0.856	0.878	0.763	0.941	0.949	NOTERMINA
	0.881	0.119	0.882	0.881	0.881	0.763	0.941	0.935	Weighted Avg.
Naive Bayes	0.54	0.307	0.637	0.54	0.584	0.235	0.677	0.645	TERMINA
	0.693	0.46	0.601	0.693	0.644	0.235	0.677	0.694	NOTERMINA
	0.616	0.384	0.619	0.616	0.614	0.235	0.677	0.669	Weighted Avg.
SMO	0.569	0.351	0.618	0.569	0.593	0.219	0.609	0.567	TERMINA
	0.649	0.431	0.601	0.649	0.624	0.219	0.609	0.565	NOTERMINA
	0.609	0.391	0.61	0.609	0.608	0.219	0.609	0.566	Weighted Avg.
K-nn	0.639	0.025	0.963	0.639	0.768	0.652	0.808	0.814	TERMINA
	0.975	0.361	0.73	0.975	0.835	0.652	0.808	0.727	NOTERMINA
	0.807	0.193	0.846	0.807	0.801	0.652	0.808	0.771	Weighted Avg.
AdaBost	0.505	0.262	0.658	0.505	0.571	0.249	0.695	0.704	TERMINA
	0.738	0.495	0.598	0.738	0.661	0.249	0.695	0.678	NOTERMINA
	0.621	0.379	0.628	0.621	0.616	0.249	0.695	0.691	Weighted Avg.

Tabla 9 Evaluación de modelos predictivos creados.

La experimentación permitió determinar que el mejor algoritmo clasificador fue Random Forest con una precisión de 88% y un F1 de 88.1%, sin embargo, es posible obtener un modelo más eficiente si se realizan pruebas con la totalidad de los registros 16 FP.

Capítulo 4 Evaluación

4.1 Introducción

Actualmente, uno de los principales desafíos que enfrenta la educación en México son los altos índices de reprobación, la deserción estudiantil y la baja eficiencia terminal de los egresados. Estos problemas se presentan como síntomas de una posible deficiencia en la calidad del proceso educativo. En particular, la Eficiencia Terminal (ET) permite evidenciar los efectos negativos de la reprobación y la deserción, los cuales se reflejan en un bajo rendimiento escolar. Cuando la ET se analiza en conjunto con otros indicadores, como la duración promedio de los estudios y el gasto educativo, se puede obtener una comprensión más profunda de los costos adicionales y las pérdidas para el sistema educativo.

La Secretaría de Educación Pública (SEP) define la eficiencia terminal como la proporción entre el número de alumnos que ingresan y los que egresan de una misma generación, considerando el año de ingreso y el año de egreso según la duración del plan de estudios. De manera complementaria, Savedra et al. (2005) describen la eficiencia terminal como "la proporción de estudiantes que concluyen un programa en un tiempo determinado, frente al total de los que lo iniciaron un cierto número de años antes". Los problemas relacionados con la baja eficiencia terminal se atribuyen a múltiples factores, entre los que destacan la rigidez y especialización excesiva de los planes de estudio, así como la persistencia de métodos obsoletos de enseñanza y evaluación (Gallardo et al., 2019).

En este contexto, el estudio de los factores de personalidad ha cobrado relevancia, dado su potencial para predecir acciones futuras, lo que se refleja en la importancia de comprender cómo estos factores pueden influir en el desempeño académico y la retención de los estudiantes en programas de posgrado. Yuchengzhang (2017) demostró que los factores de personalidad ofrecen un valor predictivo significativo en la anticipación de conductas en diversas situaciones. La investigación en variables psicológicas de personalidad ha sido crucial para identificar patrones de comportamiento tanto normales como anormales en distintos contextos (Zambrano-Cruz, 2011). Vinet y Forns (2006) definen la personalidad como un conjunto de características estables que determinan el comportamiento de una persona en diferentes situaciones, lo que la hace única e irrepetible.

Dado que la personalidad influye significativamente en cómo los estudiantes enfrentan los retos académicos, el uso de herramientas que permitan medir y analizar estos rasgos se vuelve esencial en el contexto educativo. En particular, el cuestionario de 16 factores de personalidad se ha convertido en una herramienta clave para predecir el comportamiento de las personas en diversas situaciones. Este instrumento ha demostrado su utilidad no solo en la orientación personal, escolar y profesional, sino también en la selección de personal en el ámbito laboral (Boyle et al., 2008). En el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), se emplean diversos instrumentos de selección para identificar a los estudiantes con el perfil adecuado para cursar un posgrado. Estos instrumentos incluyen I) exámenes, II) entrevistas con la coordinación, y III) cursos propedéuticos (TecNM | Tecnológico Nacional de México Campus CENIDET, n.d.).

Sin embargo, a pesar de la implementación de estos filtros, no todos los estudiantes logran titularse dentro del tiempo estipulado de dos años para la maestría, lo cual compromete la certificación de calidad del CENIDET, evaluada por el Programa Nacional de Posgrados de Calidad. Ante este desafío, se propone el desarrollo de un modelo predictivo basado en el cuestionario de 16 factores de personalidad, con el objetivo de anticipar la eficiencia terminal de los estudiantes de maestría en el CENIDET. Este enfoque permitirá no solo mejorar la selección y seguimiento de los estudiantes, sino también incrementar la eficiencia terminal y fortalecer la calidad educativa y la acreditación del CENIDET.

En este capítulo se presentan las actividades realizadas durante el periodo escolar de junio a octubre de 2022, correspondiente al cuarto semestre. Estas actividades se llevaron a cabo en las instalaciones del CENIDET. Durante este periodo, se dividió el conjunto de datos original en datos de entrenamiento y datos de prueba, lo que permitió la obtención de un modelo para predecir la eficiencia terminal de un estudiante de maestría en el CENIDET. Posteriormente, se evaluó el modelo obtenido.

Para lograr estos resultados, se llevaron a cabo tres fases principales: I) Recolección y análisis de los datos, II) Preprocesamiento de los datos, y III) Modelado predictivo. Cada una de estas fases fue esencial para garantizar la precisión y validez del modelo predictivo. Además, se consideraron diversas métricas de rendimiento para evaluar la efectividad del modelo, asegurando que este pueda ser aplicado de manera práctica en futuros ciclos de selección y seguimiento de estudiantes.

4.2 Modelo propuesto

El modelo propuesto se compone de cuatro fases principales: (I) recolección y análisis de datos, (II) preprocesamiento de la información, (III) modelado predictivo, y (IV) Experimentación. Cada una de estas fases es crucial para asegurar la integridad y precisión del modelo. En la primera fase, se recopilan y examinan los datos relevantes, mientras que en la segunda fase, se limpian y preparan para el modelado. Finalmente, la tercera fase se enfoca en la construcción del modelo, y la cuarta en verificar su efectividad mediante diversas métricas de evaluación.

Fase 1. Recolección y análisis de datos

La comprensión de los datos constituye el núcleo de la primera fase de la metodología de solución propuesta. En esta fase, se integran tres fuentes de información clave: la lista de estudiantes aceptados a la maestría por generación, el cuestionario de 16 factores de personalidad, y la lista de estudiantes graduados. Para construir la base de datos, se digitalizó la información de 24 generaciones pasadas, asegurando que se contara con datos completos para cada uno de los estudiantes. Una vez consolidada la base de datos, se procedió a analizar la distribución de los datos mediante técnicas de descripción estadística y representación gráfica, lo que permitió identificar patrones y tendencias fundamentales para el desarrollo del modelo predictivo.

Fase 2. Pre-procesamiento de datos

El preprocesamiento de datos es el proceso mediante el cual se convierten los datos en una forma utilizable y optimizada, conocida como conjunto de datos. Los datos sin procesar son propensos a problemas como ruido, valores faltantes e incoherencias, lo que puede afectar negativamente la calidad de los resultados obtenidos en la minería de datos. Dado que la calidad de los datos influye directamente en la precisión y efectividad del análisis, es esencial mejorar los datos antes de su procesamiento. En esta investigación, se mejoró la calidad de los datos sin procesar mediante un procedimiento de preprocesamiento que incluyó: detección y manejo de datos faltantes, análisis de datos atípicos, reducción de la dimensionalidad y balanceo de clases. Estas etapas resultaron cruciales para optimizar la eficiencia y precisión del modelo predictivo desarrollado.

Fase 3. Modelado predictivo

La selección del algoritmo para el modelo es una actividad crucial en la tercera fase del proceso. Por ello, se llevaron a cabo varios experimentos para desarrollar el modelo predictivo de eficiencia terminal a partir del cuestionario de 16 factores de personalidad propuesto en esta investigación. En esta fase, se emplearon seis de los diez algoritmos principales de minería de datos identificados por la Conferencia Internacional de Minería de Datos (Witten, 2011). La elección de estos algoritmos tuvo como objetivo evaluar su rendimiento y determinar cuál de ellos proporcionaba los mejores resultados en la predicción de la eficiencia terminal. Los experimentos incluyeron una comparación exhaustiva de los algoritmos en términos de precisión, robustez y capacidad de generalización. Este análisis permitió seleccionar el algoritmo más adecuado para mejorar la exactitud del modelo predictivo y asegurar su aplicabilidad en escenarios reales.

Fase 4. Experimentación

En la cuarta fase, se pone en marcha el modelo y se prueba con datos nuevos sin etiquetar. Esta etapa es crucial para evaluar la capacidad del modelo predictivo para generalizar a nuevos conjuntos de datos. Los resultados obtenidos se analizan para determinar la robustez y precisión del modelo. La evaluación incluye la comparación de las predicciones del modelo con los resultados esperados y la identificación de cualquier ajuste necesario. Esta fase permite confirmar la eficacia del modelo y garantizar que sea capaz de ofrecer predicciones confiables en escenarios reales, contribuyendo así a su aplicabilidad práctica y a la mejora continua del proceso predictivo.

Metodología de eficiencia terminal se muestra en la figura 8.

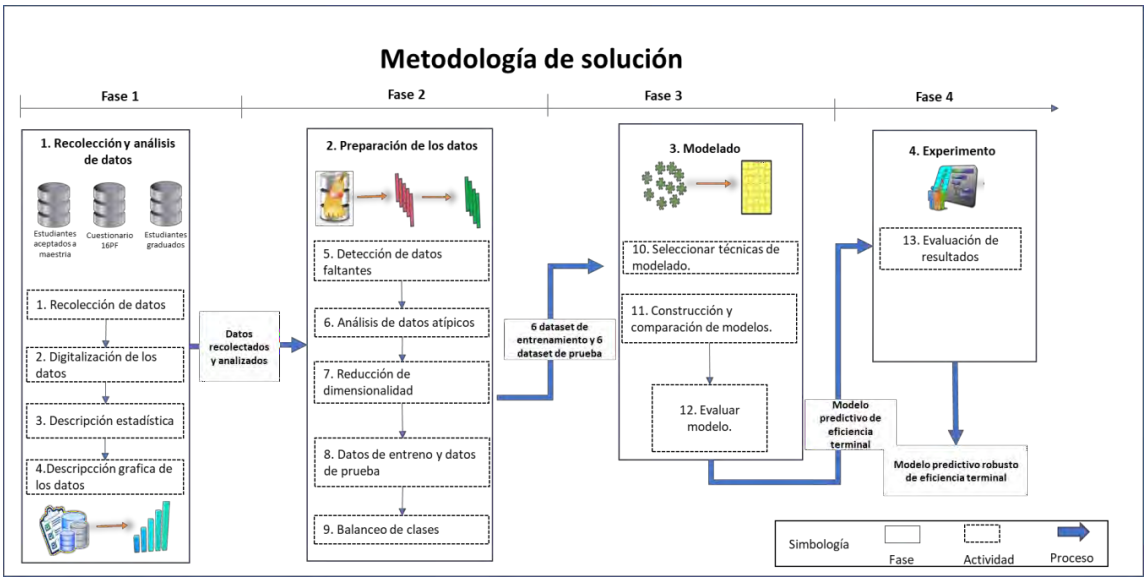


Figura 8 Mapa conceptual de la metodología de solución.

Como resultado de la metodología aplicada, se desarrolló un modelo para predecir la eficiencia terminal de los estudiantes de posgrado del CENIDET, utilizando los siguientes atributos:

1. Expresividad emocional
2. Fuerza del yo
3. Impulsividad
4. Aptitud situacional
5. Emotividad
6. Sutileza
7. Posición social
8. Autoestima
9. Estado de ansiedad
10. Factores coincidentes
11. Pronóstico

El modelo empleado para esta predicción es el algoritmo K Nearest Neighbor, el cual ha demostrado un nivel de precisión del 80.6931% y una medida F (f-measure) del 80.1%. Estos resultados evidencian la eficacia del modelo en la predicción de la eficiencia terminal, reflejando su capacidad para identificar con precisión a los estudiantes que tienen mayor probabilidad de completar sus estudios de posgrado en el CENIDET.

4.3 Modelado predictivo

Tras identificar las variables más relevantes para nuestro estudio mediante técnicas de aprendizaje computacional, se identifican patrones de comportamiento que permiten crear un modelo predictivo como se muestra en la figura 9.

Figura 9 Representación del proceso de modelado predictivo de eficiencia terminal de estudiantes de posgrado.

En el contexto del aprendizaje automático, un modelo es una representación de un patrón extraído mediante el procesamiento de un conjunto de datos. Los modelos se entrenan y ajustan utilizando algoritmos de aprendizaje automático aplicados a esos datos. Entre las representaciones de modelos más comunes se encuentran los árboles de decisiones y las redes neuronales. Un modelo predictivo define un mapeo (o función) que asocia un conjunto de atributos de entrada con un valor para un atributo de destino. Una vez creado, el modelo puede aplicarse a nuevas instancias dentro del mismo dominio.

Un modelo predictivo es un mecanismo que anticipa el comportamiento de un individuo, utilizando las características del mismo como entrada y proporcionando una calificación predictiva como salida. Cuanto más elevada sea la calificación, mayor será la probabilidad de que el individuo exhiba el comportamiento previsto.

El modelo predictivo de eficiencia terminal para estudiantes de posgrado se utilizará para predecir si un estudiante terminará o no su maestría en el tiempo previsto, en función de los datos proporcionados por el cuestionario de 16 factores de personalidad. Una vez que se introduzcan los datos del estudiante y se aplique el modelo predictivo, se obtendrá un resultado que indicará si el estudiante completará o no la maestría en el tiempo esperado.

Además, este modelo puede ser refinado continuamente mediante la incorporación de nuevos datos, mejorando así su precisión a lo largo del tiempo. La posibilidad de ajustar el modelo a contextos específicos permite adaptarlo a diferentes instituciones educativas o programas

académicos. Asimismo, el análisis de los resultados generados por el modelo puede proporcionar información valiosa para la creación de estrategias de intervención temprana. Con esto, se busca no solo predecir, sino también apoyar al estudiante en su proceso educativo. Por último, la aplicación de estos modelos fomenta una cultura de toma de decisiones basada en datos en el ámbito académico.

4.4 Seleccionar técnicas de modelado

El aprendizaje computacional se empleó originalmente para desarrollar técnicas que permitieran a las computadoras aprender. Hoy en día, ya que incluye una serie de métodos estadísticos avanzados para la regresión y la clasificación, tiene aplicación en una amplia variedad de campos, incluyendo diagnósticos médicos, detección de fraudes de tarjetas de crédito, reconocimiento de la cara y el habla y el análisis del mercado de valores. Los experimentos se realizaron con seis de diez algoritmos de minería de datos principales identificados por la Conferencia internacional de minería de datos (Witten, 2011). A continuación, se proporciona una breve descripción de algunos de estos métodos utilizados para el análisis predictivo.

4.4.1 Máquinas de vectores de soporte

Las máquinas de vectores de soporte (SVM) se usan para detectar y explotar patrones complejos de datos agrupando, ordenando y clasificando los datos. Son máquinas de aprendizaje que se utilizan para realizar clasificaciones binarias y estimaciones de regresión. Usualmente usan métodos basados en kernel para aplicar técnicas de clasificación lineal a problemas de clasificación no lineal. Hay una serie de tipos de SVM tales como lineal, polinomial, sigmoide, etc.

4.4.2 Naïve Bayes

El clasificador bayesiano ingenuo se basa en la regla de probabilidad condicional de Bayes, que se utiliza para la tarea de clasificación. El clasificador bayesiano asume que los predictores son estadísticamente independientes, lo que hace que sea una herramienta de clasificación eficaz que sea fácil de interpretar. Se emplea mejor cuando se enfrenta al problema de la "maldición de la dimensionalidad", es decir, cuando el número de predicciones es muy alto.

4.4.3 K-vecinos más cercanos

El algoritmo vecino más próximo k-NN (Nearest Neighbor) pertenece a la clase de métodos estadísticos de reconocimiento de patrones. El método no impone a priori ninguna suposición sobre la distribución de la que se extrae la muestra de modelado. Se trata de un conjunto de

entrenamiento con valores positivos y negativos. Una nueva muestra se clasifica calculando la distancia al vecino más cercano del conjunto de entrenamiento. El signo de ese punto determinará la clasificación de la muestra. En el clasificador k-vecino más cercano, se consideran los k puntos más cercanos y se utiliza el signo de la mayoría para clasificar la muestra. El rendimiento del algoritmo k-NN está influenciado por tres factores principales:

- la medida de distancia utilizada para localizar a los vecinos más cercanos
- la regla de decisión usada para derivar una clasificación de los k-vecinos más cercanos
- el número de vecinos utilizados para clasificar la nueva muestra.

Se puede demostrar que, a diferencia de otros métodos, este método es universal y asintóticamente convergente, es decir, a medida que el tamaño del conjunto de entrenamiento aumenta, si las observaciones son independientes e idénticamente distribuidas, independientemente de la distribución a partir de la cual se dibuja la muestra, la clase predicha convergerá a la asignación de clase que minimiza el error de clasificación errónea.

4.4.4 Arboles de decisión

Son una técnica de aprendizaje automático supervisado, resuelve problema complejo haciendo preguntas organizadas jerárquicamente. Divide el problema en una región más específica del espacio de decisión. Los nodos caen siempre en una categoría específica del campo objetivo. El método utiliza particionamiento recursivo para dividir los registros en segmentos minimizando la impureza en cada paso. La Impureza de los nodos se calcula mediante la Entropía de los datos en el nodo.

4.4.5 RandomForest

Es un algoritmo de aprendizaje de ensamble basado en árboles, el clasificador Random Forest es un conjunto de árboles de decisión de un subconjunto del conjunto de entrenamiento seleccionado al azar. Agrega los votos de los diferentes árboles de decisión para decidir la clase final del objeto de prueba.

4.4.6 AdaBoost

AdaBoost, abreviatura de Adaptive Boosting , es un meta-algoritmo de clasificación estadística, es una técnica de modelado de conjuntos que aborda problemas de clasificación binaria. Estos algoritmos mejoran el poder de predicción al convertir una cantidad de estudiantes débiles en estudiantes fuertes. Lo que hace este algoritmo es que construye un modelo y otorga pesos iguales a todos los puntos de datos. Luego asigna pesos más altos a los puntos que están mal clasificados. Ahora todos los puntos que tienen pesos más altos tienen más importancia en el próximo modelo. Seguirá entrenando modelos hasta que se reciba un error mínimo.

Estos nos son las únicas técnicas de aprendizaje computacional, existen otras como la función de base radial, el perceptrón multicapa o modelado predictivo geoespacial. Sin embargo, estas técnicas no son objeto de estudio del presente trabajo.

4.4.7 Construcción y comparación de modelos

El modelo para predecir eficiencia terminal de un estudiante de maestría de CENIDET se obtuvo a partir de la reducción de características y el equilibrio de clases. Este modelo predictivo se obtuvo también gracias a la técnica PCA (Análisis de componentes principales) es una técnica utilizada para describir un conjunto de datos en términos de nuevas variables («componentes») no correlacionadas. Los componentes se ordenan por la cantidad de varianza original que describen, por lo que la técnica es útil para reducir la dimensionalidad de un conjunto de datos. Esta técnica seleccionó 11 atributos de los 20. Una vez que se seleccionaron las variables se procedió a hacer los entrenos con distintos clasificadores.

	trees.J48	RandomForest	Naive Bayes	SMO	k-nn	AdaBoost
Correctly Classified Instances	72.28%	88.12%	61.63%	60.89%	80.69%	**62.13%**
Incorrectly Classified Instances	27.72%	11.88%	38.37%	39.11%	19.31%	**37.87%**
Kappa statistic	0.4455	0.7624	0.2327	0.2178	0.6139	**0.2426**
Mean absolute error	0.3065	0.2891	0.4238	0.3911	0.1947	**0.4409**
Root mean squared error	0.4967	0.3352	0.4936	0.6254	0.4382	**0.4706**
Relative absolute error	61.30%	57.83%	84.77%	78.22%	38.95%	**88.18%**
Root relative squared error	99.34%	67.05%	98.71%	125.07%	87.64%	**94.12%**
Total Number of Instances			**404**			

Tabla 10 Comparación de resultados de los modelos creados.

Para evaluar el desempeño de cada modelo creado se usaron las métricas de evaluación como se muestran en la tabla 11.

Clasificador	TPRate	FPRate	Precision	Recall	F-Measure	MCC	ROC Area	PRC area	Class
trees.J48	0.909	0.556	0.667	0.909	0.769	0.406	0.662	0.65	TERMINA
	0.444	0.091	0.8	0.444	0.571	0.406	0.662	0.693	NOTERMINA
	0.7	0.346	0.727	0.7	0.68	0.406	0.662	0.67	Weighted Avg.
RandomForest	0.906	0.144	0.863	0.906	0.884	0.763	0.941	0.922	TERMINA
	0.856	0.094	0.901	0.856	0.878	0.763	0.941	0.949	NOTERMINA
	0.881	0.119	0.882	0.881	0.881	0.763	0.941	0.935	Weighted Avg.
Naive Bayes	0.54	0.307	0.637	0.54	0.584	0.235	0.677	0.645	TERMINA
	0.693	0.46	0.601	0.693	0.644	0.235	0.677	0.694	NOTERMINA
	0.616	0.384	0.619	0.616	0.614	0.235	0.677	0.669	Weighted Avg.
SMO	0.569	0.351	0.618	0.569	0.593	0.219	0.609	0.567	TERMINA
	0.649	0.431	0.601	0.649	0.624	0.219	0.609	0.565	NOTERMINA
	0.609	0.391	0.61	0.609	0.608	0.219	0.609	0.566	Weighted Avg.
K-nn	0.639	0.025	0.963	0.639	0.768	0.652	0.808	0.814	TERMINA
	0.975	0.361	0.73	0.975	0.835	0.652	0.808	0.727	NOTERMINA
	0.807	0.193	0.846	0.807	0.801	0.652	0.808	0.771	Weighted Avg.
AdaBoost	0.505	0.262	0.658	0.505	0.571	0.249	0.695	0.704	TERMINA
	0.738	0.495	0.598	0.738	0.661	0.249	0.695	0.678	NOTERMINA
	0.621	0.379	0.628	0.621	0.616	0.249	0.695	0.691	Weighted Avg.

Tabla 11 Evaluación de modelos predictivos creados.

Como se puede observar el mejor algoritmo clasificador fue RandomForest con una precisión de 88% y un F1 de 88.1% en comparación con los demás sin embargo al hacer las pruebas con los datos que anteriormente no entraron al entreno podemos tener un modelo más eficiente en la tabla 12 se muestra los resultados de nuestras pruebas con datos etiquetados.

Una vez creado el modelo, podemos visualizarlos con un árbol de decisiones como se muestra en la figura 10.

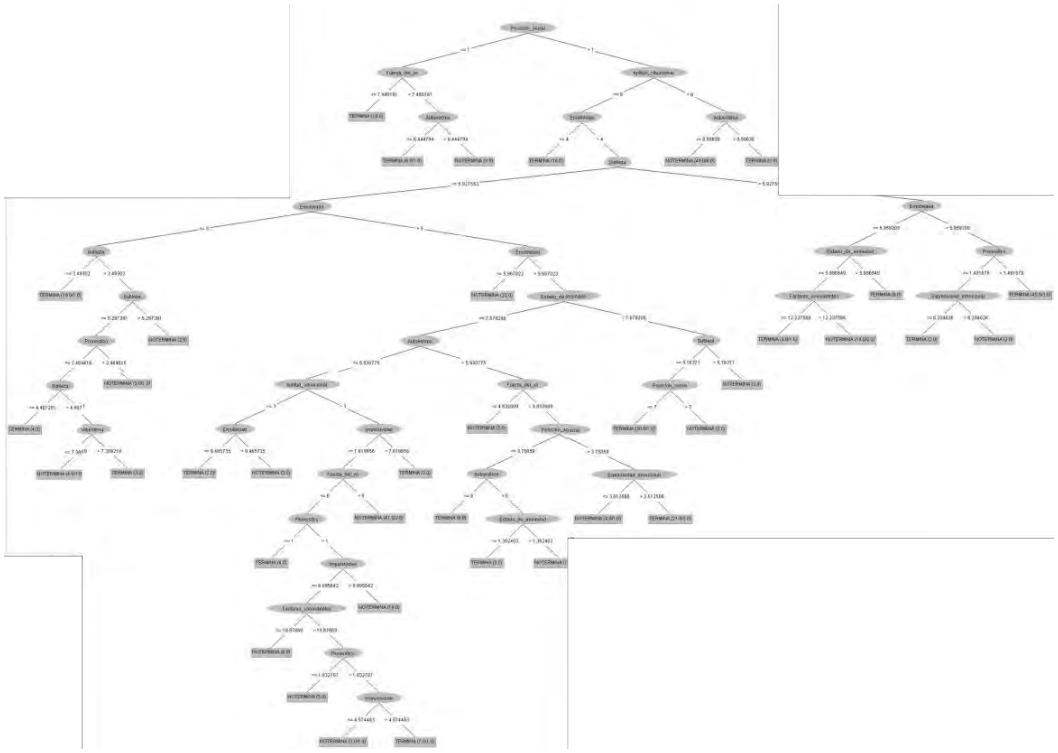


Figura 10 Visualización del modelo predictivo con un árbol de decisiones

4.5 Evaluación del modelo

Una vez se ha creado un modelo es necesario comprobar que el mismo funciona de manera correcta, este es el aspecto más importante de los modelos predictivos, su evaluación. Para la creación del modelo predictivo se utilizan unidades de muestra disponibles con atributos y un comportamiento conocido, a este conjunto de datos se le denomina conjunto de entrenamiento. Por otro lado, se utilizará unas series de unidades de otra muestra con atributos similares, pero de las cuales no se conoce su comportamiento, a este conjunto de datos se le denomina conjunto de prueba.

Para predecir el desempeño de un clasificador en nuevos datos, necesitamos evaluar la tasa de error en un conjunto de datos que no se usó en la construcción del clasificador la figura 3 muestra los pasos para evaluar nuestros modelos creados.

Figura 11 Pasos para evaluar nuestro modelo predictivo.

Esta actividad consistió en evaluar el modelo obtenido con los datos de prueba que se dividió en la actividad anterior. Esta evaluación consistió en ingresar los datos que en total son 20 registros, donde 11 están etiquetados como "TERMINA" y 9 como" NOTERMINA". En la tabla 3 se muestra la comparación de los resultados donde tenemos el modelo que mejor clasifico fue K-nn con 75% de instancias correctamente clasificadas y 25.0% instancias clasificadas

incorrectamente.

Summary	trees.J48	RandomForest	Naive Bayes	SMO	K-nn	AdaBoost
Correctly Classified Instances	70%	55%	60%	65%	75.00%	60%
Incorrectly Classified Instances	30%	45%	40%	35%	25.00%	40%
Kappa statistic	0.3684	0.0217	0.1919	0.3	0.4898	0.1919
Mean absolute error	0.3254	0.446	0.4061	0.35	0.2512	0.4443
Root mean squared error	0.5366	0.5301	0.4994	0.5916	0.4988	0.4696
Total Number of Instances			20			

Tabla 12 Comparación de resultados de las pruebas con 20 instancias etiquetadas.

También se obtuvo la matriz de confusión la que nos muestra dónde es que es que se está equivocando nuestro modelo, y nos dice que clasificó 9 personas etiquetadas con la clase "TERMINA" mientras que 2 las clasificó en la clase" NOTERMINA", de la clase NOTERMINA por otro lado clasificó 6 que NOTERMINA y 3 que TERMINA, la tabla 4 muestra la comparación de resultados.

Clasificador	a	b	classified as
	10	1	a = TERMINA
trees.J48	5	4	b = NOTERMINA
	10	1	a = TERMINA
RandomForest	9	1	b = NOTERMINA
	7	4	a = TERMINA
Naive Bayes	4	5	b = NOTERMINA
	7	4	a = TERMINA
SMO	3	6	b = NOTERMINA
	9	2	a = TERMINA
K-nn	3	6	b = NOTERMINA
	7	4	a = TERMINA
AdaBoost	4	5	b = NOTERMINA

Tabla 13 Matriz de confusión de las pruebas realizadas.

Un modelo predictivo nunca proporcionara el 100% de aciertos, incluso muchas veces se aleja bastante de esos resultados. Esto se debe a que por mucho que se haya repetido un patrón de comportamiento en el pasado, no tiene por qué repetirse. Sin embargo, siempre será mejor predecir con ayuda del modelo ya que nos da un porcentaje que simplemente adivinar.

4.6 Pruebas del método propuesto

4.6.1 Descripción del grupo experimental

El grupo de estudio está formado por estudiantes de maestría del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) ubicado en la ciudad de Cuernavaca del estado de Morelos. Este grupo de estudio estuvo formado por diecisiete estudiantes, quienes eran de las líneas de investigación Ingeniería de software, Inteligencia Artificial, Computo Inteligente, Sistemas distribuidos y

Sistemas híbridos inteligentes, de la generación 2017 a la generación 2019. En la Figura 4 se muestra las características demográficas de los diecisiete estudiantes del grupo experimental expresados en porcentajes.

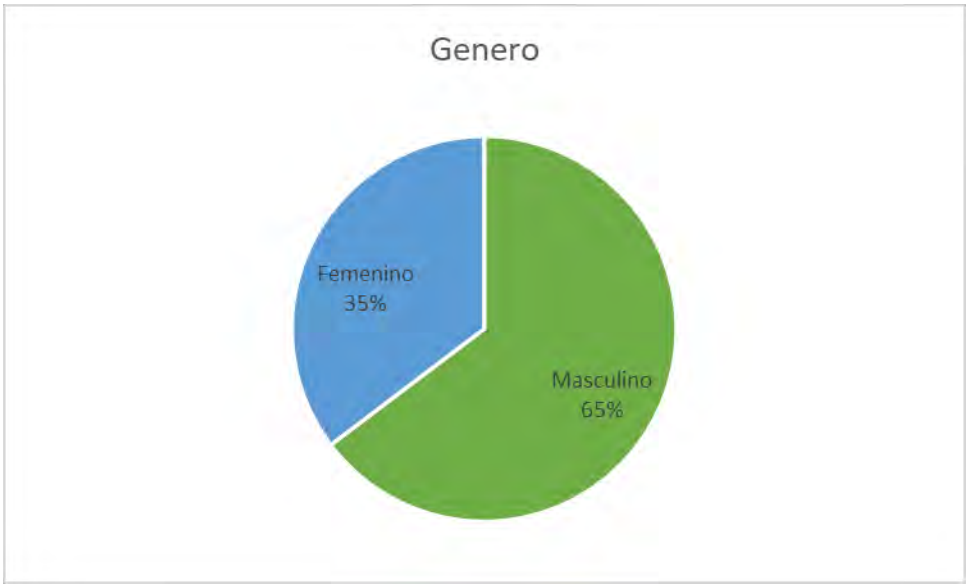


Figura 12 Grafica que muestra los porcentajes de genero del grupo de estudiantes.

4.6.2 Procedimiento de pruebas

El procedimiento de la prueba consistió en ingresar los once atributos del cuestionario de 16 factores de personalidad al modelo creado y este debe arrojar la predicción como se muestra en la figura 4.

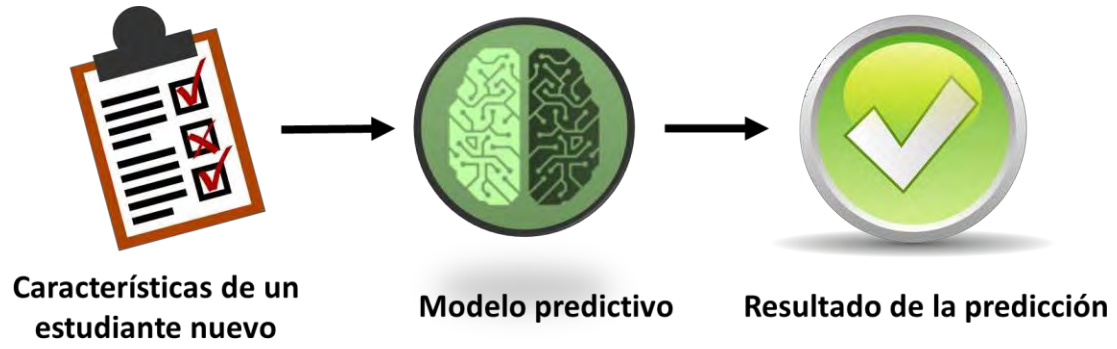


Figura 13 Proceso de predicción con datos sin etiquetar.

El modelo predictivo está conformado por once atributos pertenecientes al cuestionario de 16 factores de personalidad que son: Expresividad_emocional, Fuerza_del_yo, Impulsividad, Aptitud_situacional, Emotividad, Sutileza, Posicion_social, Autoestima, Estado_de_ansiedad, Factores_coincidentes y Pronóstico. El modelo de eficiencia terminal se evalúa tomando en cuenta diecisiete estudiantes con los once atributos que se generaron de la reducción de la dimensionalidad y generando como salida la clasificación de un estudiante si termina o no termina. La figura 6 representa este proceso.

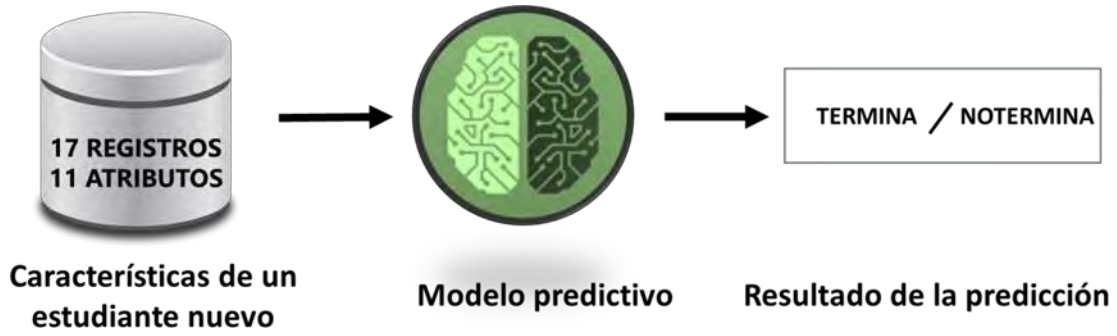


Figura 14 Representación del proceso del procedimiento de pruebas.

4.6.3 Evaluación de resultados obtenidos

En la tabla 5 se muestra la distribución de los registros del grupo de estudio donde se observa el número de cada factor de personalidad.

No	1) Expresividad_emocional	2) Fuerza_del_yo	3) Impulsividad	4) Aptitud_situacional	5) Emotividad	6) Sutileza	7) Posicion_social	8) Autoestima	9) Estado_de_ansiedad	10) Factores_coincidentes	11) Pronostico
ESTUDIANTE1	4	6	5	4	7	6	4	8	5	15	1
ESTUDIANTE 2	8	9	5	4	6	2	1	8	3	11	3
ESTUDIANTE3	8	8	3	5	8	3	7	7	1	11	3
ESTUDIANTE4	9	10	4	7	5	9	2	8	1	12	2
ESTUDIANTE5	8	8	10	7	8	3	4	6	4	13	2
ESTUDIANTE6	6	7	4	5	7	6	3	9	1	12	2
ESTUDIANTE7	8	4	6	7	5	2	2	6	8	13	2
ESTUDIANTE8	5	6	6	7	5	3	3	6	3	13	2
ESTUDIANTE9	6	2	6	5	7	4	5	7	7	15	1
ESTUDIANTE10	6	5	3	5	8	6	2	4	6	13	2
ESTUDIANTE11	4	3	6	4	7	4	2	6	8	13	2
ESTUDIANTE12	8	6	4	5	5	3	2	4	8	11	3
ESTUDIANTE13	5	4	5	5	6	2	1	3	5	14	1
ESTUDIANTE14	7	3	7	4	7	6	3	6	8	12	2
ESTUDIANTE15	8	4	4	5	7	2	4	8	4	14	1
ESTUDIANTE16	2	6	4	7	10	3	6	4	8	9	3
ESTUDIANTE17	5	8	7	8	5	6	5	8	2	14	1

Tabla 14 Once variables del cuestionario 16 factores de personalidad.

En la figura 15-17 se muestran las gráficas de cada una de las personas del grupo de estudio donde se puede apreciar la distribución individual de sus datos.

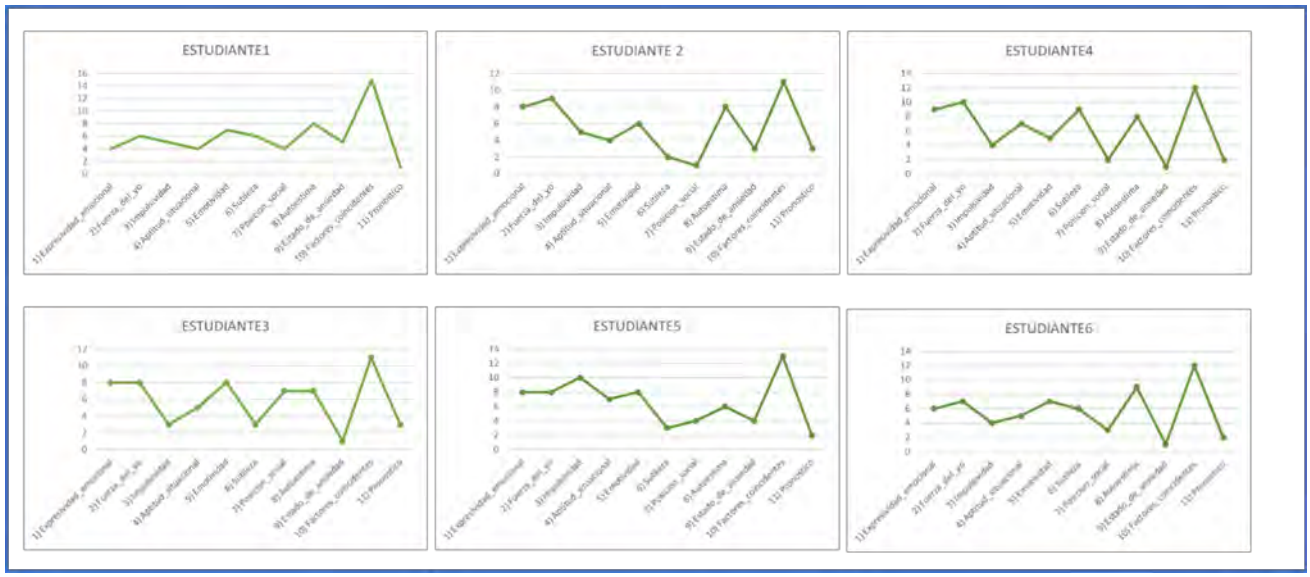


Figura 15 Relación de variables relevantes con cada individuo del grupo de estudio.

Figura 16 Relación de variables relevantes con cada individuo del grupo de estudio.

Figura 17 Relación de variables relevantes con cada individuo del grupo de estudio.

Una vez que introducimos los datos nuevos al modelo este nos arrojó la predicción como se muestra en la tabla 15 donde se hace una comparación de lo que predijo el modelo.

	Generacion	Fecha de examen	Nombre	Especialidad	Sexo	Expresividad_emocional	Fuerza_del_yo	Impulsividad	Aptitud_situacional	Emotividad	Sutileza	Posicion_social	Autoestima	Estado_de_ansiedad	Factores_coincidentes	Pronostico	Meses	Clase	Modelo
1	2018-1	15-ene-21	ESTUDIANTE 1	IS	0	4	6	5	4	7	6	4	8	5	15	1	36	TERMINA	NOTERMINA
2	2018-1	27-ene-20	ESTUDIANTE 2	IS	1	8	9	5	4	6	2	1	8	3	11	3	24	TERMINA	TERMINA
3	2018-1	17-jul-20	ESTUDIANTE 3	IS	0	8	8	3	5	8	3	7	7	1	11	3	30	TERMINA	TERMINA
4	2018-1	02-feb-21	ESTUDIANTE 4	IA	0	9	10	4	7	5	9	2	8	1	12	2	37	TERMINA	NOTERMINA
5	2017-2	16-jul-21	ESTUDIANTE 5	SD	0	8	8	10	7	8	3	4	6	4	13	2	47	TERMINA	TERMINA
6	2019-2	16-jul-21	ESTUDIANTE 6	CL	0	6	7	4	5	7	6	3	9	1	12	2	23	TERMINA	TERMINA
7	2017-2	10-ene-20	ESTUDIANTE 7	IS	1	8	4	6	7	5	2	2	6	8	13	2	30	TERMINA	TERMINA
8	2017-2	27-ene-20	ESTUDIANTE 8	IA	1	5	6	6	7	5	3	3	6	3	13	2	30	TERMINA	TERMINA
9	2017-2	06-feb-20	ESTUDIANTE 9	SD	0	6	2	6	5	7	4	5	7	7	15	1	31	TERMINA	TERMINA
10	2017-2	06-feb-20	ESTUDIANTE 10	IS	0	6	5	3	5	8	6	2	4	6	13	2	31	TERMINA	NOTERMINA
11	2018-1	07-feb-20	ESTUDIANTE 11	IS	0	4	3	6	4	7	4	2	6	8	13	2	25	TERMINA	TERMINA
12	2018-1	28-feb-20	ESTUDIANTE 12	IS	1	8	6	4	5	5	3	2	4	8	11	3	25	TERMINA	TERMINA
13	2020-1	?	ESTUDIANTE 13	IS	1	5	4	5	5	6	2	1	3	5	14	1	34	PENDIENTE	NOTERMINA
14	2020-1	?	ESTUDIANTE 14	SD	1	7	3	7	4	7	6	3	6	8	12	2	34	PENDIENTE	NOTERMINA
15	2020-2	?	ESTUDIANTE 15	IA	0	8	4	4	5	7	2	4	8	4	14	1	28	PENDIENTE	NOTERMINA
16	2020-2	?	ESTUDIANTE 16	SHI	0	2	6	4	7	10	3	6	4	8	9	3	28	PENDIENTE	NOTERMINA
17	2020-2	?	ESTUDIANTE 17	CI	0	5	8	7	8	5	6	5	8	2	14	1	28	PENDIENTE	NOTERMINA

Tabla 15 Resultados de la predicción de los nuevos estudiantes ingresados al sistema.

La eficiencia terminal del grupo experimental obtenidos con el estatus original comparado con el modelo se muestra en la figura 18. El modelo de eficiencia terminal seleccionado clasifico correctamente a catorce estudiantes lo cual se introduce en una precisión del 82.3.

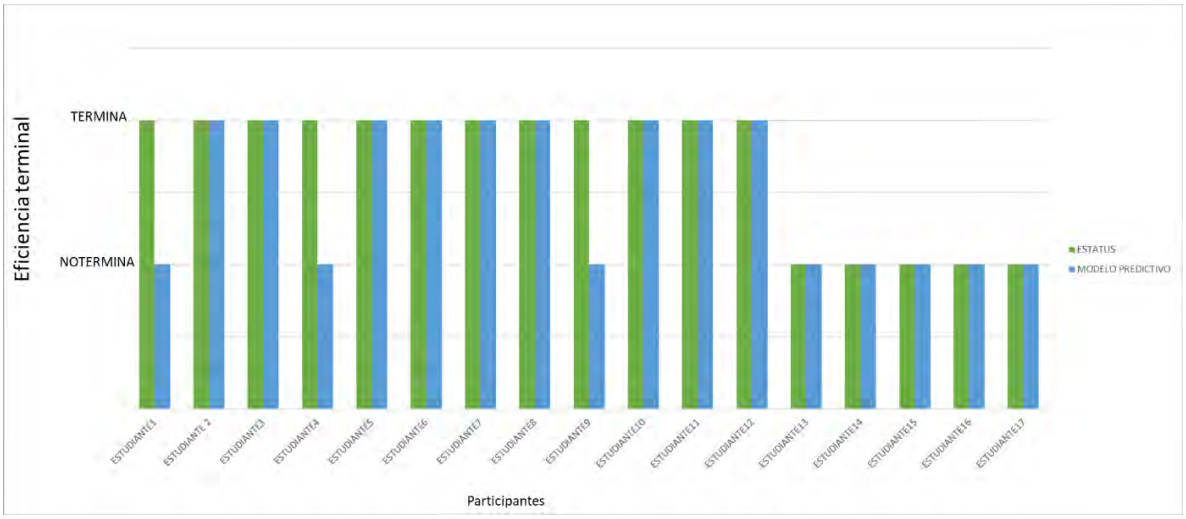


Figura 18 Comparación de los resultados del modelo predictivo y el estatus actual del estudiante.

La tabla 14 y figura 19 muestran los resultados obtenidos de este experimento.

Correctly Classified Instances	**14**	**82.3529**
Incorrectly Classified Instances	3	17.6471

Tabla 16 Porcentaje de instancias clasificadas

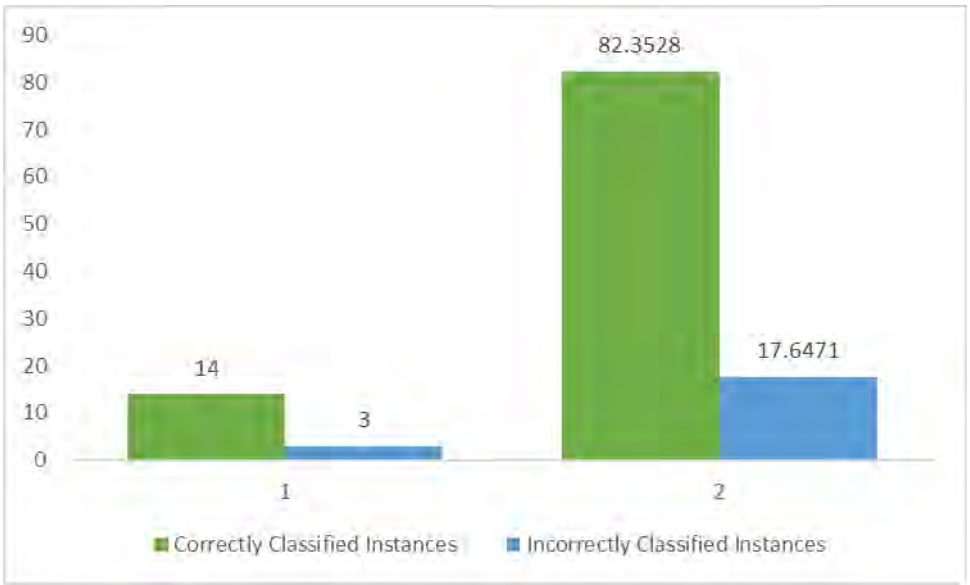


Figura 19 Comparación de los resultados.

La tabla 17 se representa la matriz de confusión para visualizar en donde se está equivocando el modelo.

a	b	classified as
9	3	a=TERMINA
0	5	b=NOTERMINAR

Tabla 17 Matriz de confusión.

Conclusiones y trabajos futuros

Este trabajo de investigación se centró en determinar la relación que existe entre los factores de personalidad y la eficiencia terminal de los programas de posgrado. Para lograr esto se propuso un modelo predictivo basado en k-NN que utiliza el cuestionario de personalidad 16PF para anticipar la eficiencia terminal de los estudiantes de maestría en el Centro de Investigación y Desarrollo Tecnológico (CENIDET).

Como parte del proceso se utilizó el procesamiento de datos para mejorar la calidad de los datos, se seleccionaron atributos relevantes y se equilibraron las clases. Posteriormente, se desarrollaron experimentos utilizando diversos algoritmos de minería de datos y se evaluaron en métricas clave como precisión, puntuación F1 y sensibilidad.

Se presenta un modelo para predecir la eficiencia terminal de los estudiantes de posgrado del CENIDET. Basado en las características del cuestionario de 16 factores, el modelo identifica qué estudiantes tienen probabilidades de completar la maestría. Para determinar los mejores modelos, se evaluaron seis algoritmos diferentes, y se encontró que el modelo de predicción k-NN tiene una precisión del 80.69% con los siguientes atributos: Expresividad Emocional, Fuerza del Yo, Impulsividad, Aptitud Situacional, Emotividad, Sutileza, Posición Social, Autoestima, Estado de Ansiedad, Factores Coincidentes y Pronóstico. La investigación ha permitido asociar estos atributos con la probabilidad de que un estudiante termine o no la maestría.

Además, se observó que los estudiantes con altos niveles de Autoestima y Fuerza del Yo tienen mayores probabilidades de éxito. Los resultados sugieren que el soporte emocional y las habilidades interpersonales son cruciales para el rendimiento académico. Este modelo no solo sirve para la predicción, sino que también puede orientar el diseño de intervenciones personalizadas para apoyar a los estudiantes en riesgo. A futuro, se planea integrar este modelo en el sistema de gestión académica del CENIDET para facilitar la toma de decisiones informadas por parte de los tutores y administradores. La validación continua y la expansión del conjunto de datos mejorarán aún más la precisión del modelo.

Los resultados confirman que existe una correlación entre la personalidad y el hecho que los estudiantes terminen en tiempo sus programas de posgrado. Esto permitió generar una metodología sólida para la predicción de resultados académicos basada en características de personalidad. Los factores relevantes para determinar si un estudiante terminará en el tiempo establecido en su programa de posgrado son los siguientes: expresividad emocional, fuerza del

yo, impulsividad, aptitud situacional, emotividad, sutileza, posición social, autoestima, estado de ansiedad, factores coincidentes y finalmente el pronóstico.

Estos hallazgos tienen implicaciones significativas en la mejora de los procesos de selección y apoyo a estudiantes de posgrado, pero es necesario experimentar con un volumen de datos mayor, además de poder ampliarse a alumnos de otros niveles educativos.

Referencias

Aregbesola, M. K., & Griva, I. (2022). A Fast Algorithm for Training Large Scale Support Vector Machines. *Journal of Computer and Communications*, *10*(12), 1–15. https://doi.org/10.4236/jcc.2022.1012001

Balogun, A. O., Basri, S., Mahamad, S., Abdulkadir, S. J., Almomani, M. A., Adeyemo, V. E., Al-Tashi, Q., Mojeed, H. A., Imam, A. A., & Bajeh, A. O. (2020). Impact of feature selection methods on the predictive performance of software defect prediction models: An extensive empirical study. *Symmetry*, *12*(7). https://doi.org/10.3390/sym12071147

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. In *International Journal of Medical Informatics* (Vol. 77, Issue 2, pp. 81–97). https://doi.org/10.1016/j.ijmedinf.2006.11.006

Benedick, P. Lou, Robert, J., & Traon, Y. Le. (2021). A systematic approach for evaluating artificial intelligence models in industrial settings. *Sensors*, *21*(18). https://doi.org/10.3390/s21186195

Boyle, G. J. (n.d.). *A Review of the Factor Structure of the Sixteen Personality Factor Questionnaire and the Clinical Analysis Questionnaire*. https://www.researchgate.net/publication/234728797

Boyle, G. J., & Lennon, T. J. (1994). Examination of the reliability and validity of the Personality Assessment Inventory. In *Journal of Psychopathology and Behavioral Assessment* (Vol. 16, Issue 3). http://epublications.bond.edu.au/hss_pubs/792

Boyle, G. John., Matthews, Gerald., & Saklofske, D. H. (2008a). *The SAGE handbook of personality theory and assessment*. SAGE Publications.

Boyle, G. John., Matthews, Gerald., & Saklofske, D. H. (2008b). *The SAGE handbook of personality theory and assessment*. SAGE Publications.

Breiman, L. (2001a). *Random Forests* (Vol. 45).

Breiman, L. (2001b). Statistical Modeling: The Two Cultures. In *Statistical Science* (Vol. 16, Issue 3).

Čehovin, L., & Bosnić, Z. (2010). Empirical evaluation of feature selection methods in classification. *Intelligent Data Analysis*, *14*(3), 265–281. https://doi.org/10.3233/IDA-2010-0421

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In *Journal of Artificial Intelligence Research* (Vol. 16).

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, *26-June-2016*, 2201–2206. https://doi.org/10.1145/2882903.2912574

CS-GY 6923 Machine Learning - Kannan. (n.d.).

Cuéllar Saavedra, M., & Espinoza, B. (2006a). Revista de la Educación Superior. *Revista de La Educación Superior*, *XXXV*(3), 7–27. http://www.redalyc.org/articulo.oa?id=60413901

Cuéllar Saavedra, M., & Espinoza, B. (2006b). Revista de la Educación Superior. *Revista de La Educación Superior*, *XXXV*(3), 7–27. http://www.redalyc.org/articulo.oa?id=60413901

Gallardo et al 2019 Eficiencia terminal. (n.d.).

García-Martínez, C., & Ventura, S. (2020). Multi-view genetic programming learning to obtain interpretable rule-based classifiers for semi-supervised contexts. Lessons learnt. *International Journal of Computational Intelligence Systems*, *13*(1), 576–590. https://doi.org/10.2991/ijcis.d.200511.002

Goodfellow, I., Bengio, Y., & Courville, A. (n.d.). *Deep Learning*.

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. (2022). *A Review of Safe Reinforcement Learning: Methods, Theory and Applications*. http://arxiv.org/abs/2205.10330

Gupta, D., & Rani, R. (2019). A study of big data evolution and research challenges. *Journal of Information Science*, *45*(3), 322–340. https://doi.org/10.1177/0165551518789880

Horvitz, E., & Mulligan, D. (2015). Data, privacy, and the greater good. *Science*, *349*(6245), 253–255. https://doi.org/10.1126/science.aac4520

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, *39*(1), 1503–1509. https://doi.org/10.1016/j.eswa.2011.08.040

John, G. H. (n.d.). *338 Estimating Continuous Distributions in Bayesian Classifiers*. http://robotics.stanford.edu/-gjohn/

John, O. P., Naumann, L. P., & Soto, C. J. (n.d.). *Paradigm Shift to the Integrative Big Five Trait Taxonomy History, Measurement, and Conceptual Issues*.

Jollife, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 374, Issue 2065). Royal Society of London. https://doi.org/10.1098/rsta.2015.0202

Kohavi, R. (n.d.). *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. http//roboticsStanfordedu/"ronnyk

Leevy, J. L., Hancock, J., & Khoshgoftaar, T. M. (2023). Comparative analysis of binary and one-class classification techniques for credit card fraud data. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00794-5

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. In *ACM Computing Surveys* (Vol. 50, Issue 6). Association for Computing Machinery. https://doi.org/10.1145/3136625

Mach-Król, M., & Hadasik, B. (2021). On a certain research gap in big data mining for customer insights. In *Applied Sciences (Switzerland)* (Vol. 11, Issue 15). MDPI AG. https://doi.org/10.3390/app11156993

Mariscal, Á., Mendivil, M. /, Jessica, T., Govea, N., & De Jesús, M. (n.d.). *BAJA CALIFORNIA, EN ENSENADA*.

Mitchell, T. M. (2006). *The Discipline of Machine Learning*.

Nie, J.-Y., Institute of Electrical and Electronics Engineers, & IEEE Computer Society. (n.d.). *2017 IEEE International Conference on Big Data : proceedings : Dec 11- 14, 2017, Boston, MA, USA*.

Nissar, I., Rizvi, D. R., Masood, S., & Mir, A. N. (2019). Voice-based detection of parkinson's disease through ensemble machine learning approach: A performance study. *EAI Endorsed Transactions on Pervasive Health and Technology*, *5*(19). https://doi.org/10.4108/eai.13-7-2018.162806

Papageorgiou, G., Grant, S. W., Takkenberg, J. J. M., & Mokhles, M. M. (2018). Statistical primer: How to deal with missing data in scientific research? *Interactive Cardiovascular and Thoracic Surgery*, *27*(2), 153–158. https://doi.org/10.1093/icvts/ivy102

Pietrzak, D., Page, B., Korcuska, J. S., & Gorman, A. B. (2015). An examination of the relationship of the 16PF fifth edition to a multidimensional model of self-esteem. *SAGE Open*, *5*(4). https://doi.org/10.1177/2158244015611453

Revista146_S5A1ES. (n.d.).

Roberts, B. W., & Mroczek, D. (2008). Personality trait change in adulthood. *Current Directions in Psychological Science*, *17*(1), 31–35. https://doi.org/10.1111/j.1467-8721.2008.00543.x

Rochin Berumen, F. L. (2021). Deserción escolar en la educación superior en México: revisión de literatura. *RIDE Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, *11*(22). https://doi.org/10.23913/ride.v11i22.821

Ross Quinlan, by J., Kaufmann Publishers, M., & Salzberg, S. L. (1994). *Programs for Machine Learning* (Vol. 16).

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). *Behavioral Development and Construct Validity: The Principle of Aggregation* (Vol. 94, Issue 1). American Psychological Association, Inc.

Sardinha, J. A. R. P., Garcia, A., Lucena, C. J. P., & Milidiú, R. L. (n.d.). *A Systematic Approach for Including Machine Learning in Multi-Agent Systems*.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. In *SN Computer Science* (Vol. 2, Issue 3). Springer. https://doi.org/10.1007/s42979-021-00592-x

Schermer, J. A., Krammer, G., Goffin, R. D., & Biderman, M. D. (2020). Using the 16PF to test the differentiation of personality by intelligence hypothesis. *Journal of Intelligence*, *8*(1). https://doi.org/10.3390/jintelligence8010012

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (n.d.). *Hidden Technical Debt in Machine Learning Systems*.

Semi-Supervised_Learning_Chapelle_O._et_al._Eds._2006_Book_reviews. (n.d.).

Shrifan, N. H. M. M., Akbar, M. F., & Isa, N. A. M. (2022). An adaptive outlier removal aided k-means clustering algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), 6365–6376. https://doi.org/10.1016/j.jksuci.2021.07.003

Singh Rawat, S., & Kumar Mishra, A. (n.d.). *Review of Methods for Handling Class-Imbalanced in Classification Problems*.

Soria-Barreto, K., & Zúñiga-Jara, S. (2014). Aspectos determinantes del éxito académico de estudiantes universitarios. *Formacion Universitaria*, *7*(5), 41–50. https://doi.org/10.4067/S0718-50062014000500006

Tang, J., Alelyani, S., & Liu, H. (n.d.). *Feature Selection for Classification: A Review*.

Vinet, E. V. (2006). *El Inventario Clínico Para Adolescentes de Millon (MACI) y su Cap acidad Para Discriminar Entre Pob lació n General y Clínica Millon' s Adolescent Clinical Inventory (MACI) and its Cap ab ility to Discriminate Between General and Clinical Pop ulation* (Vol. 1).

Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., & Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and Information Systems*, *65*(1), 31–57. https://doi.org/10.1007/s10115-022-01772-8

Whare W-ananga, T., Hamilton, W., & Hall, M. A. (1999). *' " mvers1ty ofWaikato Correlation-based Feature Selection for Machine Learning*.

Witten, Frank, & Eibe. (n.d.). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*.

Yadav, S., & Shukla, S. (2016). Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. *Proceedings - 6th International Advanced Computing Conference, IACC 2016*, 78–83. https://doi.org/10.1109/IACC.2016.25

Yao, Y., Peng, L., & Tsakiris, M. C. (2024). Unlabeled Principal Component Analysis and Matrix Completion. In *Journal of Machine Learning Research* (Vol. 25). http://jmlr.org/papers/v25/22-0816.html.

Yu, L., & Liu, H. (n.d.). *Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution*.

Zambrano Cruz, R. (n.d.). *REVISIÓN SISTEMÁTICA DEL CUESTIONARIO DE PERSONALIDAD DE EYSENCK (EYSENCK PERSONALITY QUESTIONNARIE-EPQ) SYSTEMATIC REVIEW FROM EYSENCK PERSONALITY QUESTIONNARIE (EPQ)*.