https://repository.uaeh.edu.mx/revistas/index.php/investigium/issue/archive

Boletín Científico INVESTIGIUM de la Escuela Superior de Tizayuca

InvESTigium

ISSN: 2448-4830

Publicación semestral, Vol. 9, No. Especial (2023) 120-128

Segundo Congreso de Investigación e Innovación en Tendencias Globales
"Responsabilidad Social y Sustentabilidad"
24 – 27 de octubre de 2023

Rasgos de personalidad que inciden en la predicción de eficiencia terminal de estudiantes de posgrado

Personality traits that have influence in the prediction of student graduation rate

Alicia Martínez-Rebollar[a], Hugo Estrada-Esquivel[b], Ernesto Echeverria-Ignacio[c], Ana L. Islas-Avila[d]

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

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 Random Forest 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 Clave:

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

Introducción

En la actualidad, las organizaciones suelen utilizar instrumentos de selección de personal para determinar los mejores aspirantes para ocupar un puesto. Algunos instrumentos son las pruebas psicológicas de personalidad cuya finalidad es seleccionar personas capaces de adaptarse a los cambios, los cuales además tengan iniciativa, empatía, respuesta frente a situaciones de crisis, madurez, motivación y estabilidad emocional. El constructo de personalidad se define como las tendencias estables de una persona para comportarse de una manera determinada en diversas situaciones. En otras palabras, se trata de un patrón complejo de características comportamentales que se manifiestan en prácticamente todas las áreas de funcionamiento de un individuo, incluyendo tendencias generales en la

[a] *Autor de Correspondencia, TECNM/CENIDET, https://orcid.org/0000-0002-1071-8599, Email: alicia.mr@cenidet.tecnm.mx*

[b] *TECNM/CENIDET, https://orcid.org/0000-0002-1466-7581, Email: hugo.ee@cenidet.tecnm.mx*

[b] *TECNM/CENIDET, https://orcid.org/0009-0008-4550-7899, Email: m21ce010@cenidet.tecnm.mx*
[d] *TECNM/CENIDET, https://orcid.org/0009-0003-9172-6076, Email: m23ce066@cenidet.tecnm.mx*

Fecha de recepción: 04/10/2023, Fecha de aceptación: 05/10/2023, Fecha de publicación: 05/12/2023

DOI: https://doi.org/10.29057/est.v9iEspecial.11796

percepción, emoción, pensamiento, acción y relaciones con otros. La investigación en variables psicológicas de personalidad se ha vuelto cada vez más relevante, ya que ha contribuido a identificar patrones normales o anormales de comportamiento en diferentes situaciones [1].

El estudio de la personalidad se ha enfocado en la identificación de rasgos característicos de las personas, y existen diversas estrategias para evaluar este constructo. Uno de los instrumentos más confiables en las pruebas psicológicas es el cuestionario de 16FP (16 factores de personalidad) desarrollado por Cattell [2], el cual es un cuestionario que está construido desde un punto de vista no patologicista [3].

En el contexto del proceso de selección de estudiantes en el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), se busca medir tanto las capacidades aptitudinales como las actitudinales de los aspirantes. Por esta razón, se aplica el cuestionario de 16 FP a todos los aspirantes que deseen ingresar al posgrado en cada generación. Este cuestionario es uno más de los requisitos de ingreso que deben cumplir los aspirantes y que permite analizar la idoneidad de los candidatos para realizar un posgrado. Sin embargo, a pesar de los múltiples requisitos y evaluaciones que se aplican a los aspirantes, no todos los estudiantes logran titularse en el tiempo establecido para la duración del programa de posgrado, lo que afecta la eficiencia terminal de los programas.

La eficiencia terminal es definida por la Secretaría de Educación Pública (SEP) 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 de acuerdo con la duración del plan de estudios. De esta manera, diversos factores inciden en la eficiencia terminal, como la rigidez y especialización excesiva de los planes de estudio, así como métodos de enseñanza y evaluación [4]. Por esta razón, es necesario no solo evaluar la idoneidad de los estudiantes para ingresar a la maestría, sino también identificar cualquier factor que pueda afectar la finalización del posgrado. En el CENIDET, se cuenta con más de 35 años impartiendo maestrías y doctorados, por lo que existe un gran número de cuestionarios del 16 FP aplicados en todas las generaciones. Por esta razón surge la necesidad de aplicar técnicas de Aprendizaje Máquina para descubrir las características de personalidad que tienen los egresados para culminar sus estudios en tiempo y forma.

El objetivo de este trabajo de investigación es desarrollar un modelo predictivo que utilice la información histórica del cuestionario 16FP de los estudiantes del CENIDET para predecir aquellos que terminarán sus estudios y aquellos que no terminarán. Esta información será utilizada para determinar si un estudiante seleccionado podrá terminar sus estudios de forma más independiente o si, en caso contrario, se trata de un estudiante que requerirá de mayor atención y apoyo por parte de los directores de tesis para que concluya con éxito sus estudios.

El artículo se encuentra organizado de la siguiente manera: en la Sección 1 se presentan los fundamentos teóricos y el estado del arte. La Sección 2 aborda la metodología empleada para desarrollar el modelo predictivo, describiendo en detalle las diferentes fases que la componen. Los resultados obtenidos de la evaluación del modelo se exponen en la Sección 3. Finalmente, en la Sección 4 se presentan las conclusiones derivadas del trabajo de investigación.

Fundamentos teóricos y estado del arte

En esta sección se presenta el estado del arte, así como los principales conceptos utilizados en el trabajo de investigación.

Estado del arte

El cuestionario de los 16 factores de personalidad ha sido empleado en diversas instituciones educativas, como en el caso de estudiantes de Psicología de la Universidad Autónoma de México. En este caso el estudio se enfocó en la descripción de las diferencias en los factores de personalidad que distinguen a los estudiantes en su trayectoria académica en el campo de la Psicología. La investigación involucró la aplicación del cuestionario 16 FP a una muestra de 272 estudiantes en turnos matutino y vespertino. Los resultados resaltan diferencias en la manifestación de factores de personalidad, tanto en aspectos negativos como positivos, en relación con el grado académico y el turno de estudio de los participantes [5].

La personalidad también ha sido analizada utilizando el instrumento de los Cinco Grandes Factores de Personalidad (Extraversión, Apertura, Amabilidad, Estabilidad Emocional y Responsabilidad) en la educación. Tal es el caso del presentado en el trabajo de investigación [6], en donde se identificaron las diferencias entre los estudiantes de tres programas académicos y se prevé la elección académica adecuada para los futuros estudiantes analizando sus rasgos de personalidad. Los hallazgos de este estudio demuestran que el modelo propuesto tiene un gran potencial para ayudar a los futuros estudiantes a tomar decisiones informadas y elegir las opciones de estudio superiores adecuadas, al tiempo que aumenta la motivación académica.

Otro estudio realizado en la escuela de Nivel Medio Superior de Celaya [7] tuvo el objetivo de conocer y

analizar los factores que inciden en la eficiencia terminal en el Nivel Medio Superior del país y en particular del estado de Guanajuato. En este trabajo se aplicó una encuesta a 227 alumnos, demostrando la relación entre eficiencia terminal y aspectos como la relación con la familia, las aspiraciones académicas, el nivel máximo de estudios dentro de la familia, plan de estudios, etc.

Fundamentos Teóricos

A continuación, se presentan los principales conceptos y teorías empleadas en esta investigación, como son: personalidad, el cuestionario de los 16 factores de personalidad y el aprendizaje automático.

Personalidad:

La personalidad comprende las características psicológicas arraigadas de una persona, incluyendo pensamientos, sentimientos y comportamientos que perduran en el tiempo, haciéndonos únicos. Esto influye en nuestras reacciones ante diversas situaciones [8] y, hasta cierto punto, puede predecir cómo nos comportaremos. Debido a la complejidad humana y la variedad de factores que influyen en nuestro comportamiento, puede resultar imposible identificar un solo predictor de conducta.

Cuestionario de 16 factores de personalidad (16PF):

El Cuestionario 16PF es una medida ampliamente utilizada de la personalidad normal de los adultos que se desarrolló a partir de la investigación factorial de los elementos estructurales básicos de la personalidad. Se basa en la teoría de la personalidad multinivel de Cattell [2], y mide 16 factores primarios, 5 factores globales o de segundo estrato, y 2 factores de tercer estrato.
El trabajo presentado por Boyle [9] presenta una vista general de la investigación del cuestionario de 16 factores de personalidad, en la cual muestra que es una medida de rango normal que ha resultado ser eficaz en una variedad de entornos en los que se necesita una evaluación en profundidad de la persona en su totalidad. El objetivo principal es describir las características básicas del cuestionario 16PF y mostrar pruebas de la utilidad en entornos clínicos, de asesoramiento, industriales-organizativos, educativos y de investigación. Asimismo, tiene el objetivo de describir sus usos, aplicaciones y para concluir hacer comparación de las escalas globales del cuestionario 16PF con otros modelos.
En el aspecto formativo, el cuestionario 16PF resulta valioso para identificar factores importantes en el desempeño educativo. La Tabla 1 muestra de manera organizada los factores que influyen en la eficiencia terminal:

Escala		Los polos bajo (-) y alto (+)
Afabilidad	A-	Fría, impersonal, distante
	A+	Cálida, afable, generosa y atenta a los demás
Razonamiento	B-	De pensamiento concreto
	B+	De pensamiento abstracto
Estabilidad	C-	Reactiva y emocionalmente cambiante
	C+	Emocionalmente estable, adaptada y madura
Dominancia	E-	Deferente, cooperativa y evita conflictos
	E+	Dominante, asertiva y competitiva
Animación	F-	Seria, reprimida, cuidadosa
	F+	Animosa, espontánea, activa y entusiasta
Atención normas	G-	Inconformista, muy suya e indulgente
	G+	Atenta a las normas, cumplidora y formal
Atrevimiento	H-	Tímida, temerosa y cohibida
	H+	Atrevida, segura en lo social y emprendedora
Sensibilidad	I-	Objetiva, nada sentimental y utilitaria
	I+	Sensible, esteta y sentimental
Vigilancia	L-	Confiada, sin sospechas y adaptable
	L+	Vigilante, suspicaz, escéptica y precavida
Abstracción	M-	Práctica, con los pies en la tierra, realista
	M+	Abstraída, imaginativa, idealista
Privacidad	N-	Abierta, genuina, llana y natural
	N+	Privada, calculadora, discreta y no se abre
Aprensión	O-	Segura, despreocupada y satisfecha
	O+	Aprensiva, insegura y preocupada
Apertura al cambio	Q1-	Tradicional y apegada a lo familiar
	Q1+	Abierta al cambio, experimentadora y analítica
Autosuficiencia	Q2-	Seguidora y se integra en el grupo
	Q2+	Autosuficiente, individualista y solitaria
Perfeccionismo	Q3-	Flexible y tolerante con el desorden o las faltas
	Q3+	Perfeccionista, organizada y disciplinada
Tensión	Q4-	Relajada, plácida y paciente
	Q4+	Tensa, enérgica, impaciente e intranquila

Tabla 1. Escalas primarias del cuestionario 16PF

Aprendizaje Automático:

El aprendizaje automático o Machine Learning se ha definido como el campo de estudio que otorga a las computadoras la capacidad de aprender sin ser programadas explícitamente [10]. El aprendizaje automático es una técnica de análisis de datos que enseña a las computadoras a hacer lo que es natural para los humanos y los animales: aprender de la experiencia. Los algoritmos de aprendizaje automático utilizan métodos computacionales para "aprender" información directamente de los datos sin depender de una ecuación predeterminada como modelo. Los algoritmos mejoran su rendimiento de forma adaptativa a medida que aumenta el número de instancias disponibles para el aprendizaje.
Los algoritmos de aprendizaje automático se pueden dividir en las siguientes cuatro categorías según la cantidad y el tipo de supervisión que necesitan durante el entrenamiento: aprendizaje supervisado, no supervisado, semi-supervisado y reforzado [11].
En el aprendizaje supervisado, los datos de entrenamiento que se alimentan a los algoritmos de

aprendizaje automático incluyen las soluciones deseadas, llamadas etiquetas o clases. Los algoritmos de aprendizaje automático supervisado ampliamente utilizados tanto para la clasificación como para la regresión incluyen k-vecino más cercano (k-NN), clasificadores Bayes ingenuos, máquinas de vectores de soporte (SVM), redes neuronales, árboles de decisión, bosques aleatorios, regresión lineal y regresión logística. En este trabajo de investigación se hace uso de este tipo de aprendizaje supervisado.

En el aprendizaje no supervisado, el conjunto de datos de entrenamiento no se encuentra etiquetado. El aprendizaje semi-supervisado es una clase de tareas y técnicas de aprendizaje supervisado que utilizan datos sin etiquetar para el entrenamiento. El aprendizaje por refuerzo es la tarea de conseguir que un agente que pueda observar el entorno seleccione y realice acciones y obtenga a cambio recompensas o sanciones en forma de recompensas negativas.

Metodología propuesta

En esta sección se presenta la metodología propuesta para el desarrollo del modelo predictivo que nos permitió identificar los rasgos de personalidad que influyen en que un estudiante pueda culminar sus estudios de posgrado.

La metodología de solución propuesta se compone de tres fases principales: (1) Recolección de datos, (2) Pre-procesamiento de datos, y (3) Construcción del modelo predictivo. La Figura 1 resume las tres fases de la metodología propuesta.

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.

Fase 1. Recolección de datos

La recolección de los datos es el núcleo de la primera fase de la metodología de solución propuesta. En esta fase se consideran las tres fuentes de los datos: 1. la lista de estudiantes aceptados a maestría por generación, 2. el cuestionario de 16 factores de personalidad y 3. la lista de estudiantes graduados. En la Tabla 2 se muestran las variables que se recolectaron de las diferentes fuentes de información.

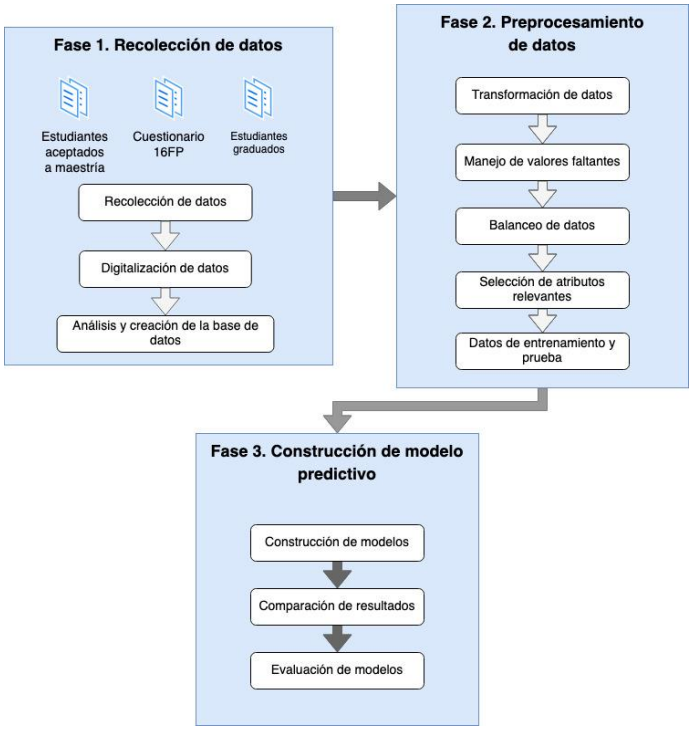

Figura 1. Metodología propuesta para el desarrollo del modelo predictivo

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

Tabla 2. Características de los atributos de la base de datos.

La base de datos se elaboró digitalizando la información de 24 generaciones pasadas en las cuales tuvimos los datos completos de cada uno de ellos como se muestra en la tabla 2. En total obtuvimos 344 registros de estas tres fuentes de datos.

Fase 2. Pre-procesamiento 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:

1) *Transformación de datos:* Esta tarea consistió 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.

2) *Tratamiento de valores faltantes:* Normalmente, los datos no están limpios y, a menudo, puede tener valores corruptos u omisos (no se presentó la situación de datos omisos). 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 esta técnica porque los resultados con las otras técnicas media, promedio y regresión, arrojan un dato de tipo decimal y la imputación por mediana nos da un valor entero.

3) *Balanceo de clases:* 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.

En la Tabla 3 se muestra en la primera columna las clases de nuestro conjunto de datos. La segunda columna corresponde al número de instancias de cada clase. La columna 3 muestra el número de instancias obtenidas después de aplicar la técnica SMOTE para el balanceo de clases.

4) *Selección de atributos relevantes*: La selección de atributos nos permitió seleccionar un subconjunto de atributos relevantes para ser utilizados en la construcción del modelo predictivo.

Clase	***Núm. De instancias***	***Núm instancias despues de aplicar SMOTE***
TMAS30M	*44*	*113*
TMENOS31M	*114*	*114*
TMENOS27M	*55*	*113*
NTERMINO	*59*	*113*
TOTAL	*272*	*453*

TMAS30M: Termina en más de 30 meses; TMENOS31M: Termina en menos de 31 meses; TMENOS27M: Termina en menos de 27 meses; NTERMINO: No termina.

Tabla 3. Resultados obtenidos después de aplicar la Técnica SMOTE en nuestro conjunto de datos.

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

- *Componentes principales o PCA:* Usa una aproximación por componentes principales para

reducir la dimensionalidad del conjunto de características. Al probar este modelo seleccionaron 11 atributos relevantes de 22 atributos.

- *CorrelationAttributeEval:* Evalúa la correlación (Pearson) entre un atributo y la clase de destino. Los atributos relevantes son aquellos que tienen una correlación positiva o negativa de moderada a alta (cercana a -1 o 1) usando 0.2 de correlación como punto de corte. No muestra atributos relevantes porque no hay una correlación entre ellos.
- *Classifiersubsetevaluator:* Evalúa subconjuntos de atributos en datos de entrenamiento o en un conjunto de prueba de reserva separado. En este método se utiliza un clasificador para estimar el "mérito" de un conjunto de atributos. En este modelo evaluó 11 atributos relevantes y se tomó la decisión de que este método nos favorecía, ya que los datos que tomó fueron 11 de los 16FP. Ver Tabla 4.

Atributos relevantes
Expresividad emocional
Fuerza del yo
Impulsividad
Aptitud situacional
Emotividad
Sutileza
Posición social
Autoestima
Estado de ansiedad
Factores coincidentes
Pronóstico

Tabla 4. Atributos relevantes

5) Datos de entrenamiento y prueba: La división de datos de entrenamiento y prueba se llevó a cabo después del equilibrio de clases y la extracción de características para evitar modelos sesgados y estimaciones demasiado optimistas, por lo cual dividimos el conjunto de datos en 80% para entrenamiento y 20% para pruebas.

Esto garantiza que el selector o clasificador de características nunca vea los datos de prueba. Utilizamos una validación cruzada de 5 veces con una proporción de 1/5 para probar los datos.

Fase 3. Construcción del modelo predictivo

Se desarrollaron varios experimentos para generar el modelo predictivo de eficiencia terminal a partir del cuestionario de 16 factores de personalidad. Los experimentos se realizaron con seis de diez algoritmos de minería de datos principales identificados por la Conferencia Internacional de minería de datos [12].

Los otros cuatro algoritmos no fueron considerados por estar enfocados al agrupamiento y por tener características que no se adaptaron al problema abordado en este trabajo de investigación.

Una vez que se seleccionaron las variables se procedió a hacer el entrenamiento con los diferentes algoritmos de clasificación. La Tabla 5 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.

Para evaluar el desempeño de cada modelo creado se usaron las métricas de evaluación como se muestran en la Tabla 6.

	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 5. Comparación de los modelos creados

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 6. Evaluación de los 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.

Evaluación del modelo

Una vez se ha creado el modelo es necesario evaluarlo para determinar su efectividad en experimentos reales. 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.

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 un total de 17 registros de estudiantes del posgrado de las generaciones 2017 a 2020, algunos de los cuales se han graduado y otros aún no finalizan a pesar de haber concluido el periodo establecido en su programa de estudio. Estos datos fueron utilizados como entrada para el modelo, el cual clasificó en forma correcta 14 registros y 3 registros de forma incorrecta.

La Tabla 7 muestra que modelo clasificó de forma correcta 9 registros de estudiantes de la categoría "TERMINA" y clasificó en forma correcta 5 registros en la categoría "NOTERMINA". El modelo clasificó de forma incorrecta 3 registros en la categoría "NOTERMINA". Esto representa una precisión del 82.3529 del modelo propuesto.

Estudiante	Generación	Fecha de examen	Meses de estudio	Clase real	Predicción del modelo
1	2018-1	15-ene-21	36	TERMINA	NOTERMINA
2	2018-1	27-ene-20	24	TERMINA	TERMINA
3	2018-1	17-jul-20	30	TERMINA	TERMINA
4	2018-1	2-feb-21	37	TERMINA	NOTERMINA
5	2017-2	16-jul-21	47	TERMINA	TERMINA
6	2019-2	16-jul-21	23	TERMINA	TERMINA
7	2017-2	10-ene-20	30	TERMINA	TERMINA
8	2017-2	27-ene-20	30	TERMINA	TERMINA
9	2017-2	6-feb-20	31	TERMINA	TERMINA
10	2017-2	6-feb-20	31	TERMINA	NOTERMINA
11	2018-1	7-feb-20	25	TERMINA	TERMINA
12	2018-1	28-feb-20	25	TERMINA	TERMINA
13	2020-1	-	34	PENDIENTE	NOTERMINA
14	2020-1	-	34	PENDIENTE	NOTERMINA
15	2020-2	-	28	PENDIENTE	NOTERMINA
16	2020-1	-	28	PENDIENTE	NOTERMINA
17	2020-1	-	28	PENDIENTE	NOTERMINA

Tabla 7. Evaluación del modelo Random Forest

Un modelo predictivo nunca proporcionará el 100% de aciertos, incluso en algunas ocasiones los resultados pueden alejarse de los resultados esperados. Esto sucede porque, a pesar de tener un modelo que refleje en forma precisa un patrón de comportamiento en el pasado, este puede cambiar, por ejemplo, para nuestro caso de estudio, si se modifican los criterios de aceptación de estudiantes, o si se modifica la duración del plan de estudios. En este caso se requiere la creación de un nuevo modelo que refleje el nuevo comportamiento.

Conclusiones

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 Random Forest (RF) 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.

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

[1] Sánchez Gallego, N. J., Gómez Macías, C., & Zambrano Cruz, R. (2011). *Revisión sistemática del Cuestionario Factorial de Personalidad (16PF).* Pensando Psicología, 7(2), 11–23.

[2] Cattell, H., & Mead, A. (2008). *The sixteen personality factor questionnaire (16PF).* SAGE Publications Ltd, https://doi.org/10.4135/9781849200479

[3] Torres Viñals, M. (1994). *Instrumentos usuales en la evaluación clínica de adultos.* Barcelona: autora.

[4] León, F. N. G. (2019*). Factores que inciden en la eficiencia terminal del alumno en la escuela de nivel medio Superior de Celaya.* http://repositorio.ugto.mx/handle/20.500.12059/3473.

[5] Díaz Contreras, S., & Díaz Reséndiz, F. de J. (2017). *Factores de personalidad en estudiantes de psicología en México.* Enseñanza e Investigación en Psicología, 22(3), 353-363.

[6] N. Kamal, F. Sarker, M. S. H. Mukta and K. A.Mamun, *"Predictive Analysis of the Effects of Personality Traits on an Academic Program,"* 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), Dhaka, Bangladesh, 2020, pp. 168-172, doi: 10.1109/ICAICT51780.2020.9333455

[7] Gallardo L., F. N., Gutierrez C. M., Mandujano, S.E., Arreola R., E. (2019*). 3 factores que inciden en la eficiencia terminal del alumno en la escuela de nivel medio Superior de Celaya.* Revista Jóvenes en la Ciencia, Vol 5 (2019). ISSN 2395-9797. pp 1-5. http://repositorio.ugto.mx/handle/20.500.12059/3473.

[8] Fiske, D.W. (1949). *Consistency of the factorial structures of personality ratings from different sources*. Journal of Abnormal and Social Psychology, 44, 329-344..

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

[10] A. L. Samuel, *"Some Studies in Machine Learning Using the Game of Checkers,"* IBM J. Res. Dev., vol. 3, no. 3, pp. 210–229, Jul. 1959.

[11] M. Kang and N. J. Jameson, *"Machine Learning: Fundamentals,"* in Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things, M. G. Pecht and M. Kang, Eds. John Wiley & Sons, 2019, pp. 85–109.

[12] Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: practical machine learning tools and techniques*. Morgan Kaufmann.

Autores

Alicia Martínez Rebollar tiene un doctorado en Informática por la Universidad Politécnica de Valencia, España y un doctorado en Informática y Telecomunicaciones por la Universidad de Trento, Italia. Actualmente desempeña el puesto de Coordinadora de la maestría y doctorado en Ciencias de la Ingeniería del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET). Sus áreas de investigación incluyen el uso de técnicas de aprendizaje automático y cómputo evolutivo para dominios de agricultura de precisión, salud mental y análisis de comportamiento humano.
https://orcid.org/0000-0002-1071-8599

Hugo Estrada Esquivel tiene un doctorado en Informática por la Universidad Politécnica de Valencia, España y un doctorado en Informática y Telecomunicaciones por la Universidad de Trento, Italia. Ha sido investigador en el centro de investigación INFOTEC, en CONACYT y actualmente en el Departamento de Ciencias Computacionales del CENIDET. Sus líneas de investigación son Internet de las Cosas, minería de datos, cómputo en la nube y ciudades inteligentes, específicamente en el análisis de la movilidad vehicular en ciudades de México.
https://orcid.org/0000-0002-1466-7581

Ernesto Echeverria Ignacio realizó sus estudios de licenciatura en el Instituto Tecnológico de Ciudad Altamirano, donde obtuvo el grado de Licenciado en Ingeniería en Informática. Actualmente, se encuentra cursando la Maestría en Ciencias Computacionales en el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), con una especialización en la línea de investigación de Cómputo Inteligente y Ciencia

de Datos. Su principal líena de investigación es el uso de ténicas de descubriento de patrones para información de estudiantes de posgrado.
https://orcid.org/0009-0008-4550-7899

Ana Luisa Islas Ávila realizó sus estudios de licenciatura en el Instituto Tecnológico de Cuautla, donde obtuvo el grado de Ingeniera en Sistemas Computacionales. Actualmente se encuentra cursando la Maestría en Ciencias Computacionales en el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), con una especialización en la línea de investigación de Cómputo Inteligente y Ciencia de Datos. Su principal enfoque de investigación se centra en técnicas de aprendizaje automático orientada a la salud para la detección de patrones en datos médicos.
https://orcid.org/0009-0003-9172-6076