

Selección de subconjuntos de características para el pronóstico del aprovechamiento académico

Feature Subset Selection for Predicting Academic Achievement

Torres Soto, M.D¹, Torres Soto, A.²

¹ Dpto. de Sistemas de Información, Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes
C/ Avenida Universidad #940. 20100 Aguascalientes, Aguascalientes. México.

² Dpto. de Ciencias de la Computación, Centro de Ciencias Básicas, Universidad Autónoma de Aguascalientes
C/ Avenida Universidad #940. 20100 Aguascalientes, Aguascalientes. México.

¹dolores.torres@edu.uaa.mx, ²aurora.torres@edu.uaa.mx

Fecha de recepción: 30 de julio de 2024

Fecha de aceptación: 27 de agosto de 2025

Resumen. En este trabajo se probaron varios mecanismos de selección de subconjuntos de características para predecir el aprovechamiento escolar de estudiantes de nivel preparatoria haciendo uso de una base de datos de 395 estudiantes con 30 variables. Los mecanismos de selección de variables utilizados fueron: los selectores 1.ANOVA con F-test, 2.Regresión logística con regularización Lasso, 3. Eliminación recursiva de características con RFE y Random_Forest, 4.Modelo de una sola característica usando Random_Forest. 5.El conjunto de las 30 variables iniciales. Se diseñó una máquina de soporte vectorial para poner a prueba los conjuntos de variables propuestos por cada mecanismo y se encontró que, el realizar esta tarea previa al algoritmo de aprendizaje favorece la calidad del modelo utilizado. El mejor mecanismo de selección encontró que la cantidad de variables ideal es 16 usando eliminación recursiva de variables con Random_Forest, además se observó que existe una fuerte consistencia entre todos los mecanismos, pues seleccionaron variables de la misma índole e incluso las mismas variables con técnicas muy distintas. El mejor modelo alcanzó una precisión del 93.93%.

Palabras Clave: Selección de Subconjuntos de Características, Umbral de Varianza, Puntuación Anova F-test, Modelo de una Sola Variable.

Summary. In this study, several feature subset selection mechanisms were tested to predict the academic performance of high school students using a dataset of 395 students with 30 variables. The variable selection mechanisms employed were: 1.ANOVA with F-test, 2. Logistic Regression with Lasso Regularization, 3. Recursive Feature Elimination (RFE) with Random Forest, 4. Single Feature model using Random Forest, 5. The full set of 30 initial variables. A Support Vector Machine (SVM) was designed to evaluate the variable sets proposed by each mechanism, and it was found that performing this preliminary task prior to the learning algorithm improved the quality of the model. The best selection mechanism determined that the ideal number of variables was 16, using recursive feature elimination with Random Forest. Moreover, strong consistency was observed among all mechanisms, as they selected variables of the same nature and even the same variables despite applying very different techniques. The best model achieved an accuracy of 93.93%.

Keywords: Feature Subset Selection, Variance Threshold, ANOVA F-test Score, Single-Variable Model.

1 Introducción

El objetivo de este trabajo es encontrar el subconjunto de características adecuado para ser insumo de un clasificador creado con una máquina de soporte vectorial con la finalidad de predecir el aprovechamiento escolar de un estudiante de nivel preparatoria con una buena probabilidad (más de un 80%). Se probaron varios mecanismos de selección de subconjuntos de características que generaron diferentes conjuntos de variables para llevar a cabo el modelo de predicción. Además de obtener un conjunto de variables muy bueno, se pudo identificar que todos los algoritmos convergen en cierto conjunto de características predictoras. A lo largo de este documento, el lector podrá encontrar información general del ámbito de aprovechamiento escolar, selección de subconjuntos de características, la metodología usada, los principales resultados y un conjunto de conclusiones interesantes.

2 Aprovechamiento escolar

El aprovechamiento o logro escolar, se refiere al nivel de conocimientos de un alumno comparado con la norma de edad y nivel académico. Normalmente para valorarlo, éste es medido mediante una evaluación; sin embargo; como todos los problemas complejos, se trata de un fenómeno multifactorial; y aunque se manifiesta a lo largo de un periodo específico, a través de una calificación cuantitativa, no es el producto de una única aptitud. El aprovechamiento escolar involucra la personalidad, motivación, capacidad intelectual, interés personales, hábitos de estudio, autoestima, entorno familiar, entre otros aspectos del alumno [1]. Estudios de diferentes áreas, han identificado los siguientes como factores que impactan en el desempeño de los estudiantes [1] [2] [3] [4]:

- Factores socioeconómicos, culturales y familiares
- Factores escolares y académicos
- Factores Psicológicos y Personales
- Factores Sociales y del Clima Escolar

El resultado de las pruebas estandarizadas que se aplican en México tales como el Programa Internacional para la Evaluación de Estudiantes (PISA) de la OCDE (Organización para la Cooperación y el Desarrollo Económico) [5]; y los Exámenes de Calidad y Logro Educativo-Media Superior (EXCALE-MS); así como estudios cualitativos gubernamentales; revelan que el sistema educativo mexicano enfrenta actualmente el desafío de mejorar sustancialmente el aprendizaje de los estudiantes de educación media superior; pues un gran número de alumnos no alcanza los objetivos en lectura y matemáticas, o posee un dominio elemental; lo que les hace vulnerables para continuar estudios superiores [6]. De acuerdo con uno de los reportes del Senado de la República Mexicana, alrededor del 40% de los estudiantes que ingresan a la educación media superior no la concluyen, y la mayor parte de los estudiantes la abandonan en los primeros grados [6].

El valor de la predicción del rendimiento escolar es de especial importancia para mejorarlo, pues permite identificar y apoyar a los estudiantes en riesgo antes de que enfrenten dificultades académicas. El predecir el aprovechamiento, también permite a las instituciones educativas implementar intervenciones estratégicas a tiempo para mejorar los resultados académicos y reducir las tasas de deserción [7].

Por otro lado, la falta de predicción del desempeño académico puede ocasionar vulnerabilidad y dificultad para alcanzar las metas de los estudiantes, pues si éstos no son conscientes de que están en desventaja, difícilmente actuarán en consecuencia.

Debido al enorme éxito que han demostrado tener los sistemas inteligentes; no es de sorprender que la predicción del rendimiento académico se apoye en modelos de aprendizaje automático. Estos modelos utilizan variables como el historial académico y factores sociodemográficos, económicos y culturales para identificar a estudiantes en riesgo. Las técnicas más precisas incluyen los árboles de decisión (DT), las máquinas de vectores de soporte (SVM) y las redes neuronales artificiales (ANN) [7].

3 Selección de Subconjuntos de Características

La selección de subconjuntos de características (Feature Subset Selection “FSS”), es el proceso de extracción del subconjunto de las características más relevantes de un conjunto de datos, para usarlas al construir y entrenar un modelo de aprendizaje automático (Machine Learning) [8, 9]. La importancia de este proceso radica en el hecho de que al trabajar con una menor cantidad de características, mejora el rendimiento de los modelos de Inteligencia Artificial (IA), reduciendo sus demandas computacionales.

Una "característica" se refiere a una propiedad o característica medible e individual: un atributo específico de los datos que ayuda a describir el fenómeno observado [8]. Por ejemplo, cuando se habla de personas, el conjunto de datos normalmente se constituye de características como "edad", "peso", "sexo", "nombre", etc.

La selección de características está íntimamente ligada con el rendimiento e interpretabilidad de los modelos de aprendizaje automático; pues el número de variables, su calidad y las interacciones entre ellas, hacen que los algoritmos “aprendan” más fácilmente y esto determina aspectos como la eficiencia, la precisión y capacidad de comprensión del modelo generado. En la figura 1 se mencionan algunas razones relevantes para no omitir este paso en el procesamiento de la información.

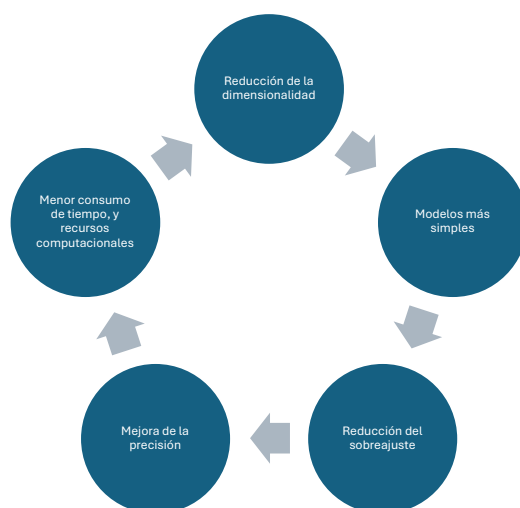


Figura 3. Clasificación de los FSSM.

Respecto de la mejora en la precisión predictiva, las características irrelevantes pueden debilitar el desempeño de los modelos al introducir “Ruido y sesgos” [10]. La selección de un subconjunto “óptimo” de características dota al modelo con una mejor capacidad de predicción.

Otro problema común de los modelos entrenados con todas las características que describen un fenómeno, es el conocido como sobreajuste (overfitting); este problema aparece cuando un modelo aprende demasiado bien los datos de entrenamiento, quedando imposibilitado para generalizar a datos nuevos [8]. La eliminación de características redundantes e irrelevantes disminuye el sobreajuste, permitiendo al modelo que generalice mejor con datos no vistos [10]. Adicionalmente, al trabajar con un número menor de características, el tiempo requerido, los requisitos de almacenamiento de información y el costo computacional para ejecutar los algoritmos, también se reducen, impactando en su eficiencia [11].

Por otro lado, el aumento en la cantidad de variables de entrada en los conjuntos de datos de alta dimensión puede causar problemas graves a muchos algoritmos de aprendizaje automático [10]. La selección de características aborda este problema al reducir los que se conoce como “la maldición de la dimensionalidad” de los datos, lo que facilita que el algoritmo identifique patrones y realice predicciones precisas [12].

Adicionalmente, la identificación de las características más relevantes ayuda a identificar las propiedades medibles individuales que son más influyentes para el proceso que se observa. Este aspecto es también muy valioso para la “interpretabilidad” de los modelos; que se refiere a la capacidad de los humanos para entender cómo funcionan y por qué hacen ciertas predicciones. En este sentido, la selección de características permite obtener modelos más simples y fáciles de entender, monitorear, auditar, depurar y explicar, lo que es crucial en el campo de la Inteligencia Artificial explicable (XAI). “La XAI es una disciplina que busca desarrollar modelos y sistemas que puedan explicar de manera comprensible su comportamiento y procesos de toma de decisiones” [13]. Esto es crucial en entornos críticos, como el sector médico, donde las decisiones automatizadas pueden tener un impacto significativo en la vida de las personas [14, 15].

Aunque la selección de características y su interpretación (incluida la importancia de las características) son pasos de modelado distintos con objetivos diferentes, se complementan. Por un lado la selección de características es un paso de preprocesamiento o restricción del modelo, que se enfoca en reducir la dimensionalidad del espacio de entrada; mientras que la importancia de las características se centra en clasificar y cuantificar la contribución de las características a las predicciones o al rendimiento del modelo.

En resumen, la selección de características es un proceso fundamental en Aprendizaje Automático, que no solo optimiza el rendimiento del modelo al hacerlo más rápido, preciso y menos propenso al sobreajuste, sino que también mejora su interpretabilidad al simplificar su estructura y hacerla más comprensible para los seres humanos.

3.1 Métodos de Selección de Subconjuntos de Características

Los métodos de selección de subconjuntos de características (Feature Subset Selection Methods “FSSM”) generalmente se clasifican en las tres categorías mostradas en la figura 2.

Hay otros factores importantes que tener en cuenta, que son los que se detallan a continuación:

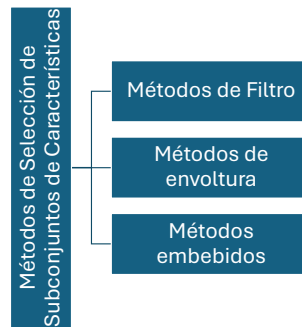


Figura 2. Importancia de la Selección de Subconjuntos de Características.

“Los métodos de filtro (Filter Methods), actúan como un paso de preprocesamiento, seleccionando características independientemente del algoritmo de Machine Learning” [15]. Clasifican las características según puntuaciones estadísticas que determinan la correlación con la variable de resultado. Algunos ejemplos de estos métodos incluyen la Correlación de Pearson, LDA (Análisis Discriminante Lineal), ANOVA (Análisis de Varianza) y Chi-cuadrada. Estos métodos son menos intensivos computacionalmente hablando y resistentes al sobreajuste; sin embargo, no toman en consideración las interacciones entre las características, por lo que tienden a seleccionar variables redundantes; razón por la que se usan a manera de preprocesamiento para establecer un ranking de características.

“Los métodos de envoltura (Wrapper Methods), utilizan un modelo predictivo para asignar una ponderación a los subconjuntos de características” [9]. Estos métodos entrenan un nuevo modelo para cada subconjunto de características. A diferencia de los anteriores, son computacionalmente intensivos y tienen riesgo de sobreajuste para el modelo; sin embargo, son capaces de detectar interacciones entre variables y frecuentemente obtienen el mejor subconjunto de características para un tipo de modelo particular. Estos métodos llevan ese nombre porque "envuelven" un algoritmo de predicción y lo usan como "caja negra" para puntuar los diferentes subconjuntos de características [10]. Frecuentemente usan metaheurísticas porque la evaluación de todos los conjuntos de características tiene una complejidad exponencial. Ejemplos de estos métodos son: el Análisis lógico combinatorio (Teoría de testores)[16] [17], Selección de características secuencial (Sequential feature selection - SFS) [18], Regresión paso a paso (Stepwise regression), Búsqueda exhaustiva (Exhaustive search) [19], Eliminación recursiva de características (RFE) por mencionar algunos.

Finalmente los métodos embebidos (Embedded Methods), se caracterizan por realizar la selección de características como parte del proceso de construcción del modelo [9]; es decir, que la selección de características y la clasificación (o regresión) se realizan de manera simultánea. La complejidad computacional de estos métodos se encuentra entre los de filtro y los de envoltura. Algunos ejemplos de algoritmos que caen en esta categoría son: Regresión LASSO (Least Absolute Shrinkage and Selection Operator) [20] y Árboles de decisión (DT) [21].

4 Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés), son un conjunto de algoritmos de aprendizaje supervisado, utilizadas para la solución de problemas de aprendizaje supervisado (clasificación y regresión) [22].

Las SVM fueron desarrolladas por Vladimir Vapnik y su equipo en los laboratorios de AT&T Bell alrededor de 1995 [22]. Aunque originalmente estos algoritmos fueron diseñados para problemas de clasificación; con el transcurso de los años se han reformulado para aplicarse en diferentes áreas como: reconocimiento de caracteres, reconocimiento del habla, detección de intrusos e incluso en Bioinformática.

En el ámbito educativo, también han sido usadas con diferentes propósitos, como por ejemplo: en la predicción del rendimiento académico de los estudiantes (problema que se afronta en este trabajo), en el diseño de programas de aprendizaje personalizados, en la evaluación de competencias, en la mejora de la toma de decisiones educativas, mitigar el riesgo de deserción, entre otros [23].

El principio de funcionamiento de la SVM radica en la separación óptima de datos mediante un hiperplano o un conjunto de hiperplanos en un espacio de dimensionalidad muy alta; separando de los puntos de una clase, de los de la(s) otra(s), tal como se muestra en la figura 3. Como se aprecia en esta figura, el

objetivo de las SVM es encontrar el hiperplano que tenga la máxima distancia (margen) con los puntos que están más cerca de él. A los puntos más cercanos al hiperplano se les denomina vectores soporte, y son los únicos ejemplos que intervienen en la construcción del hiperplano óptimo [22]. Para conjuntos de datos linealmente separables, el hiperplano óptimo equidista del ejemplo más cercano de cada clase.

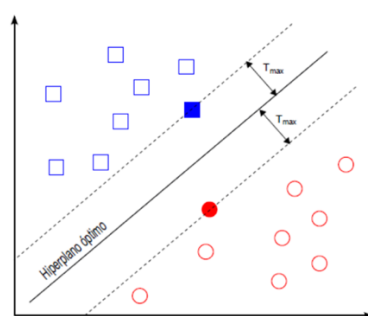


Figura 3. Hiperplano de separación óptimo para 2 clases [22].

Cuando se debe lidiar con datos no lineales, es necesario emplear el “truco del kernel”; lo que implica transformar el espacio de entrada original (“espacio de entradas”), a uno nuevo de mayor dimensionalidad (“espacio de características”) [22]; frecuentemente obtenido mediante una función no lineal. La siguiente figura el caso de un espacio de entrada original y el espacio transformado.

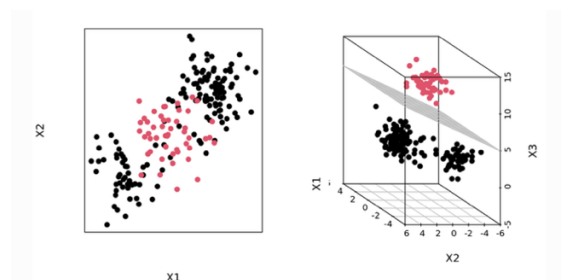


Figura 4. Espacio de entrada original (izquierda) y espacio transformado (derecha) [24].

Como se aprecia en la figura 4, en el nuevo espacio (espacio de características) se vuelve posible la separación lineal. Las funciones Kernel más comunes incluyen el lineal, polinómico y gaussiano.

Aunque algunos estudios han encontrado que otros algoritmos como los Árboles Aleatorios (Random Forest), pueden tener mayor precisión en ciertas tareas de predicción de rendimiento académico [3], las SVM son muy efectivas para manejar datos complejos y no lineales. Adicionalmente, su diseño les permite trabajar con datos en espacios de muy alta dimensionalidad y mediante la implementación de conceptos como "margen blando" son capaces de evitar el sobreajuste.

Finalmente, también se ha destacado que las SVM no requieren de un conocimiento profundo de los datos; lo que las hace de fácil aplicación en diferentes contextos [3].

5 Metodología

La presente investigación consta de varios experimentos que tienen el propósito de pronosticar el grado de aprovechamiento escolar de un estudiante de nivel preparatoria haciendo uso de máquinas de soporte vectorial con diferentes grupos de variables. Como primer paso de la metodología, se realizó un preprocesamiento de la base de datos “Student Performance” [25] validada por Cortez en 2008. Posteriormente, se implementaron 4 mecanismos de selección de subconjuntos de características para aprendizaje automático y se probaron los diferentes grupos con una máquina de soporte vectorial. Finalmente, se hace la comparación de cada subconjunto y se selecciona el mejor. Esta metodología general, se muestra en la fig. 5.

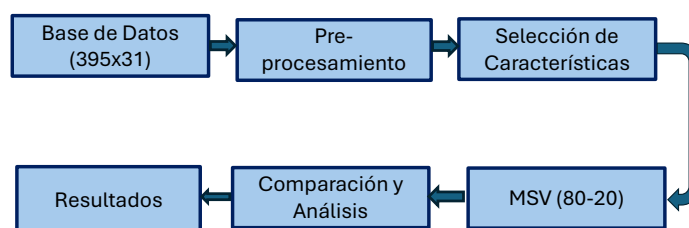


Figura 5. Metodología.

Como parte del preprocesamiento, se realizó una discretización de las variables de la base de datos original, se conjuntaron las 3 variables de resultado en una sola categoría dicotómica que contiene el aprovechamiento del estudiante con valores bueno y malo. Posteriormente, se realizó un análisis de la varianza para eliminar las características cuya varianza sea 0. Lo que implica que esa o esas variables no aportan información para diferenciar clases. Ninguna de las 30 variables fue eliminada en esta etapa de la metodología y como resultado de ella, se tuvo la base de datos preparada para los siguientes experimentos con 395 casos y 30 variables predictoras más la clase como se muestra en la figura 6.

Acto seguido, se realizó la etapa de selección de subconjuntos de características, que consistió en 4 implementaciones:

1. Selector ANOVA F-test. Un selector ANOVA F-test en scikit-learn es un método estadístico para evaluar y seleccionar variables en problemas de clasificación midiendo qué tan bien cada característica (columna) separa las clases de la variable objetivo. Consiste en separar en grupos (conforme a la clase y luego realizar el cálculo del puntaje F como sigue:

$$F = \frac{\text{Varianza dentro de la clase}}{\text{Varianza entre clases}} \quad (1)$$

La variable con un puntaje F mayor, es la más importante para hacer clasificación.

No	Variable	Valor				
1	Escuela	1=Gabriel Pereira	2=Mousinho da			
2	Género	1=Femenino	2=Masculino			
3	Edad	Numerico				
4	Tipo de Localidad	1=Urbana	2=Rural			
5	Miembros Familia	1= Menor o igual a 3	2= Mas de 3			
6	Estatus Pabres	1=Juntos	2=Separados			
7	Grado Estudios Madre	0=Ninguno	1=Primaria (4°)	2=5° Primaria a 3° Secundaria	3=Preparatoria	4= Superior
8	Grado Estudios Padre	0=Ninguno	1=Primaria (4°)	2=5° Primaria a 3° Secundaria	3=Preparatoria	4= Superior
9	Trabajo Madre	1=Educación	2=Salud	3=Servicios(admivo, policiaco,	4=Casa	5=Otro
10	Trabajo Padre	1=Educación	2=Salud	3=Servicios(admivo, policiaco,	4=Casa	5=Otro
11	Razon Escuela	1=Cercanía	2=Reputacion	3=Por sus cursos	4=Otra	
12	Tutor Académico	1=Madre	2=Padre	3=Otro		
13	Tiempo a Escuela	1=Menos 15 min.	2=Entre 15 y 30 min.	3=Entre 30 min. y 1 hora	4= Más de 1 hora	
14	Tiempo Estudio Semanal	1=Menos de 2 horas	2=Entre 2 y 5 horas	3=Entre 5 y 10 horas	4= Más de 10	
15	Clases Reprobadas	Numérica hasta 3	4 = mas de 3			
16	Apoyo Extra-Clase	1=Si	2=No			
17	Apoyo Familiar Educativo	1=Si	2=No			
18	Clases Particulares Pagadas	1=Si	2=No			
19	Actividades Extra-	1=Si	2=No			
20	Asistió a Preescolar	1=Si	2=No			
21	Desea Nivel Licenciatura	1=Si	2=No			
22	Tiene Internet	1=Si	2=No			
23	Tiene Relación Romántica	1=Si	2=No			
24	Relación Familiar Buena	1=Muy Mala	2=Mala	3=Regular	4=Buena	5=Excelente
25	Cuenta con Tiempo Libre	1=Muy Poco	2=Poco	3=Regular	4=Mucho	5=Muchísim
26	Frecuencia Salida con	1=Muy Poco	2=Poco	3=Regular	4=Mucho	5=Muchísim
27	Ingesta de Alcohol Diaria	1=Muy Baja	2=Baja	3=Regular	4=Alta	5=Muy Alta
28	Ingesta de Alcohol Semanal	1=Muy Baja	2=Baja	3=Regular	4=Alta	5=Muy Alta
29	Buena Salud	1=Si	2=No			
30	Faltas a Clase	Numérico de 0 a 93				
31	Clase	1=Buen	2=Aprovechamiento			

Figura 6. Variables del Estudio con sus Valores.

2. Selector basado en Regresión Logística con Regularización Lasso. Como la regularización Lasso es susceptible a la falta de estandarización de las variables, para utilizar este selector primero se desarrolló una estandarización Z en todas las variables predictoras usando el siguiente cálculo:

$$x_{\text{estandarizado}} = \frac{x - \mu}{\sigma} \quad (2)$$

donde:

μ es la media de la característica en el conjunto de entrenamiento.
 σ es la desviación estándar de esa característica.

Hecho lo anterior, se utilizó un selector de características basándose en la importancia que un modelo les asigna, para nuestro caso, el modelo usado fue una regresión logística. Se utilizó la regularización Lasso que en este contexto es una técnica para evitar sobreajuste (overfitting) penalizando los coeficientes del modelo ayudando a simplificarlo y a mejorar su capacidad de generalización. Este mecanismo se cataloga dentro del conjunto de técnicas embebidas de selección de características porque el selector basado en Regresión Logística con regularización Lasso (penalty='l1') hace la selección dentro del propio entrenamiento del modelo

3. Selector basado en eliminación recursiva de características con el RFE usando Random Forest. RFE (Recursive Feature Elimination) con Random Forest sirve para elegir automáticamente el subconjunto de variables más útil para el modelo. Es un método wrapper: entrena un modelo, mide la importancia de cada variable y va eliminando recursivamente las menos importantes hasta quedarse con un número óptimo.
4. Selector basado en Modelo de una sola variable usando Random Forest. Una buena manera de determinar si una variable es predictiva consiste en entrenar un modelo de machine learning usando solamente esa variable. Luego, usamos el rendimiento del modelo para ranquear las variables y seleccionar las variables mejor ranqueadas. Este mecanismo se clasifica dentro de las técnicas filter de selección de subconjuntos de característica.

Además de los modelos mencionados, se hizo un primer experimento sin hacer uso de selección de variables. Se tomaron las 30 variables listadas en la tabla 1. Validadas por

Los 4 conjuntos de datos (de cada mecanismo de selección de subconjuntos de características y el conjunto original) fueron probados en una máquina de soporte vectorial con las características que cada modelo seleccionó, se hizo la comparación correspondiente y se obtuvieron resultados y conclusiones.

6 Resultados

Los principales resultados de usar los 5 conjuntos de características mencionados en la metodología se muestran en las siguientes secciones.

6.1 ANOVA F-Test

El selector ANOVA F-test, permitió conservar 15 de las 30 variables como de interés para el problema de pronóstico del aprovechamiento académico de estudiantes de nivel preparatoria. En la Tabla 1. Podrán revisarse las variables y su puntaje obtenido con este selector.

Tabla 3. Resultados del Selector ANOVA F-test.

Orden Importancia	1	2	3	4	5	6	7	8
ANOVA F-test	37.28	15.24	11.98	10.04	7.946	6.989	6.908	4.754
No. variable	15	16	7	8	26	21	30	13
Variable	Clases Reprobadas	Apoyo Extra-Clase	Grado Estudios Madre	Grado Estudios Padre	Frecuencia Salida con Amigos	Desea Nivel Licenciatura	Faltas a Clase	Tiempo a Escuela

Orden Importancia	9	10	11	12	13	14	15
ANOVA F-test	4.503	4.131	3.381	2.848	2.698	2.691	2.515
No. variable	9	28	2	5	22	4	14
Variable	Trabajo Madre	Ingesta de Alcohol Semanal	Género	Miembros Familia	Tiene Internet	Tipo de Localidad Urbana / Rural	Tiempo Estudio Semanal

De este modelo se conservaron las primeras 15 variables por ser las que obtuvieron un puntaje de más de 2.0 que fue donde se estableció el punto de corte.

6.2 Regresión Logística con Regularización Lasso

El selector de Regresión Logística con Regularización Lasso conservó 25 de las 30 variables como de interés para el problema de pronóstico del aprovechamiento académico de estudiantes de nivel preparatoria. En la Tabla 2. se pueden revisar las variables conservadas.

Tabla 2. Resultados del Selector ANOVA F-test.

No. Variable	6	7	10	11	12	13	14	15	16	19	20	21	22
Variable	Estatus Pabres	Grado Estudios Madre	Trabajo Padre	Razon Escuela	Tutor Académico	Tiempo a Escuela	Tiempo Estudio Semanal	Clases Reprobadas	Apoyo Extra-Clase	Actividades Extra-curriculares	Asistió a Preescolar	Desea Nivel Licenciatura	Tiene Internet

No. Variable	24	25	26	28	29	2	8	17	5	23	27	30
Variable	Relación Familiar Buena	Cuenta con Tiempo Libre	Frecuencia Salida con Amigos	Ingesta de Alcohol Semanal	Buena Salud	Género	Grado Estudios Padre	Apoyo Familiar Educativo	Miembros Familia	Tiene Relación Romántica	Ingesta de Alcohol Diaria	Faltas a Clase

Este modelo rechazó la incorporación de 5 variables.

6.3 Selector basado en Eliminación Recursiva de Características con el RFE usando Random Forest

El selector basado en eliminación recursiva de características con RFE y Random Forest, conservó 16 de las 30 variables como de interés para el problema de pronóstico del aprovechamiento académico de estudiantes de nivel preparatoria. En la Tabla 3. se pueden consultar las variables conservadas.

Tabla 3. Variables conservadas por el Selector de RFE con Random Forest.

No. Variable	2	4	5	6	8	12	13	14
Variable Seleccionada RFE con Random Forest	Género	Tipo de Localidad	Miembros Familia	Estatus Pabres	Grado Estudios Padre	Tutor Académico	Tiempo a Escuela	Tiempo Estudio Semanal

No. Variable	15	16	17	21	22	24	26	28
Variable Seleccionada RFE con Random Forest	Clases Reprobadas	Apoyo Extra-Clase	Apoyo Familiar Educativo	Desea Nivel Licenciatura	Tiene Internet	Relación Familiar Buena	Frecuencia Salida con Amigos	Ingesta de Alcohol Semanal

En este modelo, se quedaron fuera del conjunto un total de 14 variables.

6.4 Selector una sola variable usando Random Forest

El selector que sometió cada variable de manera independiente al modelo usando Random Forest nos permitió conservar 10 de las 30 variables como de interés para el problema de pronóstico del aprovechamiento académico de estudiantes de nivel preparatoria. En la Tabla 4. se pueden revisar las variables conservadas.

Tabla 4. Variables Conservadas por el Selector de una sola variable usando Random Forest.

Orden de Selección	1	2	3	4	5	6	7	8	9	10
Random Forest de 1 variable	0.700	0.690	0.680	0.670	0.660	0.655	0.655	0.655	0.655	0.650
No. de Variable	15	8	22	14	27	13	16	10	7	19
Variable	Clases Reprobadas	Grado Estudios Padre	Tiene Internet	Tiempo Estudio Semanal	Ingesta de Alcohol Diaria	Tiempo a Escuela	Apoyo Extra-Clase	Trabajo Padre	Grado Estudios Madre	Actividades Extra-curriculares

Con este selector, se decidió poner el punto de corte en una accuracy de 0.65 y dejando las 20 variables restantes fuera del predictor.

Un resumen de los resultados obtenidos de los 5 mecanismos se muestra en la tabla 5. Podrá observarse que los mejores resultados para la máquina de soporte vectorial que se usa en la predicción del aprovechamiento escolar es: RFE. En esta tabla, se puede observar que el conjunto de las 30 variables predictoras es menos bueno que cualquiera de los conjuntos arrojados por los 4 mecanismos estudiados en este trabajo. De las técnicas de selección de variables Podemos observar que no es conveniente usar demasiadas pero tampoco usar demasiado pocas. Un punto adecuado se ubica en las 16 variables de 30 con el mecanismo de Eliminación Recursiva de Variables (RFE con Regresión Logística). En la tabla 6 se pueden ver los parámetros utilizados en la SVM. Además Podemos observar que los subconjuntos encontrados son consistentes en variables como: Reprobación de clases previas, si cuenta con internet, el grado de estudios del padre, de la madre o de los dos, Ingesta de alcohol diaria o semanal, Tiempo de estudio semanal, recibir apoyo extra-clase, el Tiempo que le lleva llegar de su casa a la Escuela, Actividades Extracurriculares o salir con amigos

Tabla 5. Resultados de los selectores de características.

Selector	Promedio Valicación Cruzada	Accuracy	Pecisión	F1	AUC
5 grupos					
1. ANOVA	0.8109	0.8101	0.8484	0.7887	0.8305
2. R.L. y Lasso	0.8578	0.8607	0.909	0.845	0.8812
3. RFE con RF	0.858	0.886	0.9393	0.8732	0.9065
4. Una Variable con RF	0.8107	0.8101	0.8484	0.7887	0.8305
5. Sin Selección	0.8422	0.7848	0.8181	0.7605	0.8052

Tabla 6. Parámetros de la Support Vector Machine

C	Kernel	Gamma
1.2	Poly	1
		Número de Variables

7 Conclusiones

Entre las principales conclusiones que podemos resaltar en este trabajo se encuentra el hecho de que la selección de subconjuntos de características debería considerarse como un quehacer necesario a la hora de realizar aprendizaje automático. En el caso de la predicción del aprovechamiento de estudiantes de nivel preparatoria, resultó de vital importancia. Aun tratándose de un conjunto de datos expresamente utilizado en aprendizaje automático, encontramos que todos los mecanismos de selección de subconjuntos de características usados [1. ANOVA, 2. Regresión Logística con regularización Lasso, 3. Eliminación recursiva de Variables con RFE y Random Forest y 4. Selección de una única variable con Random Forest] resultaron en pronosticadores de mejor calidad que el uso de todas las variables.

Por otro lado, encontramos que los mecanismos selectores de variables fueron muy congruentes en variables que consideramos de gran interés empírico como son:

Reprobación de clases previas de matemáticas. Es un hecho que el comenzar a reprobar contenidos de matemáticas, genera una predisposición negativa del estudiante hacia esos contenidos. En contraposición, cuando un estudiante aprueba cursos previos, genera confianza en sí mismo y esto permite que tenga una Buena actitud hacia los contenidos de matemáticas

Contar con internet. Es interesante que todos los selectores de variables conservaron esta variable. De manera empírica es natural considerar que el contar con internet para apoyar sus actividades académicas es crucial para un estudiante de nivel preparatoria.

El grado de estudios del padre, de la madre o de los dos. En este sentido, es claro que cuando se tiene un ejemplo de éxito académico por parte de la madre, del padre o de los dos, el estudiante siente un mayor compromiso por lograr niveles acordes a los de sus progenitores.

El grado de ingesta de alcohol diaria o semanal. El ingerir alcohol en demasía puede provocar dificultad para codificar y retener nueva información. El alcohol interfiere con procesos de plasticidad sináptica. En el hipocampo, región clave para formar nuevos recuerdos. El consumo repetido de alcohol provoca alteraciones estructurales y menor neurogénesis hipocampal.

El tiempo de estudio semanal. Es claro que esta variable resulta de interés desde el punto de vista empírico desde el principio. Si un alumno estudia más, obtendrá mejores calificaciones que aquél que no dedica tiempo para estudiar.

El contar o no con apoyo extra-clase. En este sentido, el apoyo es el recibido por parte de profesore de la misma escuela o de familiares que dedican tiempo los estudiantes que así lo requieren.

El tiempo que le lleva llegar de su casa a la Escuela. En muchos casos, sobre todo para los estudiantes que se movilizan a la escuela desde una comunidad rural, es frecuente que vivan a distancias muy largas y que esto dificulte su asistencia regular a clases o incluso a la posibilidad de asistir a cursos extra-clase de la misma escuela.

El realizar Actividades Extracurriculares o salir con amigos. Es natural recordar que el ser humano es multifacético y que para un bienestar (necesario para conseguir un buen desempeño académico), es necesario complementar las actividades escolares con actividades de esparcimiento como son las actividades extra-curriculares ofrecidas por la misma escuela y una socialización con personas de su edad.

Finalmente, es importante comentar que para hacer una tarea como la de predecir el aprovechamiento académico con herramientas de aprendizaje supervisado (en este caso con máquinas de soporte vectorial), la selección de características de eliminación recursiva de variables es una excelente opción para que el pronosticador sea más acertado.

Referencias

- [1] Carlos IP De (2010) Factores que influyen en el bajo rendimiento escolar desde la visión de estudiantes de bachillerato. Universidad de Querétaro
- [2] Vargas-Sánchez JR (2023) Family Capital and School Achievement in Mexican Secondary Schools. *Revista Colombiana de Educacion* 237–256. <https://doi.org/10.17227/RCE.NUM88-13575>
- [3] Gutiérrez Villaverde H, Linares Barbero M, Agüero Correa A, Pérez Nuñez J (2022) Predicción de Rendimiento Académico de alumnos usando Machine Learning
- [4] Factores psicológicos condicionantes del rendimiento académico en escolares del sistema de educación media superior – CIPSICO 2024. <https://2024.cipsico.org/ponencia/factores-psicologicos-condicionantes-del-rendimiento-academico-en-escolares-del-sistema-de-educacion-media-superior/>. Accessed 10 Aug 2025

- [5] Organización de las Naciones Unidas Acerca de | OCDE. <https://www.oecd.org/en/about.html>. Accessed 10 Aug 2025
- [6] Rodolfo Ramírez Raymundo, Germán Sergio Benítez, Eduardo Weiss Horz, et al (2015) Desafíos de la Educación Media Superior
- [7] Andre R, Carpio-Mendoza D (2024) Predicción del rendimiento académico utilizando modelos de aprendizaje automático: Una revisión sistemática de la literatura. 593 Digital Publisher CEIT 9:1038–1054. <https://doi.org/10.33386/593DP.2024.6.2797>
- [8] What is Feature Selection? | IBM. <https://www.ibm.com/think/topics/feature-selection>. Accessed 7 Aug 2025
- [9] Canedo VB Calidad vs. cantidad: La importancia de la selección de características en el análisis de datos
- [10] Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Computers and Electrical Engineering 40:16–28. <https://doi.org/10.1016/j.compeleceng.2013.11.024>
- [11] Yu L, Liu H Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution
- [12] Ziad Ali M, Abdullah A, Mohamed Zaki A, et al (2024) Advances and Challenges in Feature Selection Methods: A Comprehensive Review. Journal of Artificial Intelligence and Metaheuristics (JAIM) 07:67–77. <https://doi.org/10.54216/JAIM.070105>
- [13] Cecilia Pezzini Directora Dra Claudia Pons L, Plata L (2024) “Inteligencia Artificial Explicable: Análisis de Metodologías y Aplicaciones” Trabajo final integrador para obtener la especialización en ingeniería de software. Universidad Nacional de la Plata
- [14] Aplicación de Técnicas de Explicabilidad (XAI) y Cuantificación de Incertidumbre (UQ) en la predicción del Síndrome Metabólico mediante Aprendizaje Automático. <https://openaccess.uoc.edu/items/7b4f93a5-a0b7-4c11-b8ec-89530b999ecc>. Accessed 8 Aug 2025
- [15] Métodos de Selección de Características - AprendeIA. <https://aprendeia.com/2019/01/04/metodos-de-seleccion-de-caracteristicas-machine-learning/>. Accessed 8 Aug 2025
- [16] Ruíz-Shucloper J, Alba-Cabrera E, Lazo-Cortés M (1995) Introducción a la teoría de testores. CINVESTAV-IPN 197-undefined
- [17] Torres MD (2010) Metaheurísticas híbridas en selección de subconjuntos de características para aprendizaje no supervisado. Universidad Autónoma de Aguascalientes
- [18] El Touati Y, Slimane J Ben, Saidani T (2024) Adaptive Method for Feature Selection in the Machine Learning Context. Engineering, Technology & Applied Science Research 14:14295–14300. <https://doi.org/10.48084/ETASR.7401>
- [19] Feature Selection in Machine Learning - Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/>. Accessed 9 Aug 2025
- [20] Operador de selección y contracción mínima absoluta: descripción general | Temas de ScienceDirect. <https://www.sciencedirect.com/topics/engineering/least-absolute-shrinkage-and-selection-operator>. Accessed 8 Aug 2025
- [21] Selección de Variables en Machine Learning con Python. <https://cienciadedatos.net/documentos/py67-seleccion-variables-machine-learning-python>. Accessed 9 Aug 2025
- [22] León EC (2016) Introducción a las máquinas de vector soporte (SVM) en aprendizaje supervisado. Universidad Zaragoza
- [23] Morales NO, García PAO (2024) Application of Artificial Intelligence Models in Standardized Tests for Optimizing Academic Performance in Higher Education. European Public and Social Innovation Review 9:. <https://doi.org/10.31637/EPSIR-2024-1605>
- [24] RPubs - SVM Máquinas de Vector Soporte (Support Vector Machines). https://rpubs.com/joaquin_ar/267926. Accessed 9 Aug 2025
- [25] Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5TG7T>.