simetría

MDPI

Artículo

Un modelo de predicción de conjuntos para estudiantes potenciales Recomendación usando aprendizaje automático

Lijuan Yan 1,2,* y Yanshen Liu 1,2

1 Centro Nacional de Investigación de Ingeniería para E-Learning, Universidad Normal de China Central, Wuhan 430079, China; yanshenliu@mail.ccnu.edu.cn Centro de Investigación Hubei para la Información
2 Educativa, Universidad Normal de China Central, Wuhan 430079, China * Correspondencia: yanlijuan@mails.ccnu.edu.cn

Recibido: 29 de marzo de 2020; Aceptado: 20 de abril de 2020; Publicado: 3 mayo 2020

Resumen: La predicción del rendimiento de los estudiantes se ha convertido en un tema candente de investigación. La mayoría de los modelos de predicción existentes se construyen mediante un método de aprendizaje automático. Están interesados en la precisión de la predicción, pero prestan menos atención a la interpretabilidad. Proponemos un modelo de conjunto de apilamiento para predecir y analizar el desempeño de los estudiantes en la competencia académica. En este modelo, el desempeño de los estudiantes se clasifica en dos clases categóricas simétricas. Para mejorar la precisión, se establecen en el primer nivel tres algoritmos de aprendizaje automático, que incluyen la máquina de vectores de soporte (SVM), el bosque aleatorio y AdaBoost, y luego se integran mediante regresión logística a través del apilamiento. Se aplicó un análisis de importancia de características para identificar variables importantes. Los datos experimentales se recopilaron durante cuatro años académicos en la Universidad de Hankou. De acuerdo con estudios comparativos sobre cinco métricas de evaluación (precisión, recuperación, F1, error y área bajo la curva característica operativa del receptor (AUC) en este análisis, el modelo propuesto generalmente funciona mejor que los modelos comparados. Las variables importantes identificadas a partir del análisis son interpretables, se pueden utilizar como guía para seleccionar a los estudiantes potenciales.

Palabras clave: conjunto; modelo de predicción; desempeño de los estudiantes; aprendizaje automático

1. Introducción

Asistir a competencias académicas es una forma efectiva de evaluar el desempeño de la enseñanza y el aprendizaje, y tiene un impacto positivo en la motivación académica y los hábitos de estudio de los estudiantes en educación [1]. La competencia académica es una serie de actividades para encontrar y resolver problemas a través de actividades prácticas fuera de las aulas, lo cual es una medida efectiva para identificar y formar jóvenes talentos [2–5].

Por lo tanto, asistir a competencias académicas es muy beneficioso para los estudiantes. Las competencias académicas pueden mejorar la eficacia de estudio de los estudiantes durante las actividades de estudio y mejorar la eficacia colectiva de los estudiantes, así como su conciencia de colaboración y comunicación.

En China, las universidades orientadas a la aplicación otorgan gran importancia a la competencia académica. Se asignan tutores para entrenar a los estudiantes antes de la competencia. Sin embargo, hay dos dificultades durante la organización de la competición. Los estudiantes pierden la oportunidad de ganar el premio por no participar en el concurso. A los tutores les resulta difícil seleccionar a los estudiantes potenciales ya que el número de estudiantes es demasiado grande. Para lograr buenos resultados en la competencia, es muy importante seleccionar a los estudiantes potenciales. La observación del tutor o las pruebas de práctica pueden seleccionar a los estudiantes que tienen un rendimiento superior. Estos métodos están sesgados por los conocimientos y experiencias de los tutores. Además, no son adecuados si hay demasiados estudiantes involucrados en la selección.

En este estudio se propone un modelo de predicción para predecir y analizar el desempeño de los estudiantes en la competencia académica. Al aplicar la minería de datos a los datos de antecedentes y de comportamiento de los estudiantes, el modelo predice los resultados de la competencia y luego identifica las características clave que afectan la predicción.

resultados Este trabajo tiene valores prácticos de referencia para los tutores en la conducción y el fomento de los estudiantes potenciales para asistir a las competencias académicas.

2. Trabajo relacionado

En los últimos años, se ha vuelto cada vez más popular aplicar los métodos de aprendizaje automático para la predicción del desempeño de los estudiantes en diferentes escenarios educativos, y las fuentes de datos utilizadas en la predicción han cubierto muchos temas. Se propuso un método para predecir los "Estudiantes en riesgo" [6–8] o el rendimiento al final del aprendizaje [9] a través del análisis de datos de comportamiento de aprendizaje en línea. Algunos investigadores analizaron datos en el sistema de gestión de cursos o datos de información demográfica de los estudiantes para identificar los factores que afectan el rendimiento académico, y los estudios de investigación revelaron que los factores principales incluyen la familia lingüística [10], el hábito del sueño [11], el uso de la computadora [12,13], y así sucesivamente. Otra investigación se centró en los datos de comportamiento recopilados en el aula, que se utilizaron para analizar las correlaciones entre los puntajes de las pruebas y los diferentes cursos, y predijeron una probabilidad de aprobación de las materias específicas [14,15] o una probabilidad calificada de grado. Con base en los resultados de estos análisis, los tutores pueden tomar las medidas correspondientes para optimizar la eficiencia del aprendizaje de los estudiantes. En cuanto a la construcción del modelo de predicción del rendimiento académico, muchos investigadores tienden a adoptar métodos tradicionales de aprendizaje automático, como la regresión logística, el árbol de decisiones, la red neuronal artificial (ANN) y la máquina de vectore Se ha confirmado que estos métodos pueden mejorar el rendimiento de la predicción según los estudios actuales. Por ejemplo, Kotsiantis et al. [16] aplicó seis algoritmos (C4.5, propagación inversa, bayesiano ingenuo, 3 vecinos más cercanos, regresión logística y optimización mínima secuencial) para entrenar el conjunto de datos para predecir los malos resultados, encontraron que el algoritmo bayesiano ingenuo tenía un mejor rendimiento en la precisión satisfactoria. Romero et al. [17] comparó el rendimiento de diferentes métodos de aprendizaje automático (árboles de decisión, inducción de reglas difusas y redes neuronales) para predecir las calificaciones finales de los estudiantes. En el estudio, aplicaron técnicas de preprocesamiento de discretización y reequilibrio para obtener mejores modelos clasificadores.

El aprendizaje conjunto [18] es el proceso mediante el cual se generan e integran estratégicamente múltiples modelos (como clasificadores o expertos) para resolver un problema de inteligencia computacional. Utiliza múltiples algoritmos de aprendizaje para obtener un mejor rendimiento de predicción que el rendimiento obtenido con un solo algoritmo. Se ha aplicado a una amplia gama de temas en clasificación, regresión, selección de características y detección de puntos anormales. Por ejemplo, Beemer et al. [19] propusieron un enfoque de aprendizaje conjunto para estimar los efectos del tratamiento individualizado (ITE) para caracterizar a los estudiantes en riesgo y evaluar el éxito y la retención de los estudiantes bajo estrategias de intervención. El trabajo de Ade Roshani y Deshmukh PR [20] aplicó un conjunto incremental que consta de tres clasificadores (naive Bayes, K-star y SVM) y utilizó un esquema de votación para predecir la carrera de los estudiantes. Kotsiantis et al. [21] aplicó los métodos de conjunto para predecir el éxito de los estudiantes en el aprendizaje a distancia con tres técnicas diferentes (WINNOW, naive Bayes y 1 vecino más cercano). Estos estudios han confirmado que es más probable que el modelo de conjunto alcance una mayor precisión que el algoritmo único.

En resumen, en la actualidad se han realizado muchos estudios para predecir el rendimiento académico de los estudiantes en aulas tradicionales o plataformas de aprendizaje en línea. Estos estudios presentaron resultados muy interesantes y razonables. El estudio sobre la predicción en la competencia académica ha sido motivo de preocupación para los investigadores.

3. Contribución y Estructura del Papel

Se puede recopilar mucha información relevante en una competencia académica, como la información demográfica de un estudiante, información de comportamiento (generada por el estudio diario de los estudiantes y el desempeño de participación en competencias). Es casi imposible para un tutor descubrir si tiene una conexión con un estudiante solo por experiencia personal cuando se enfrenta a una cantidad considerable de datos.
Por lo tanto, el aprendizaje automático es una forma práctica de resolver este problema.

En este estudio, recopilamos y construimos un conjunto de datos sobre el desempeño de los estudiantes en la competencia académica. El conjunto de datos contiene toda la información sobre los estudiantes participantes y sus

resultados de la competencia (ganar o perder la competencia). Se utilizó el aprendizaje supervisado para descubrir estructuras en el conjunto de datos, y se aplicaron los métodos de aprendizaje automático para entrenar un modelo que pueda explicar los datos. Al mismo tiempo, usamos este modelo para predecir nuevos datos e identificar las características clave que afectan el desempeño de los estudiantes a través del análisis de importancia de características. Los resultados del análisis se pueden proporcionar a los tutores o gerentes, para que se utilicen como referencias para la selección de candidatos.

Sean {x1, x2, · · · xn} el conjunto de n variables características de un estudiante y y represente los resultados de la competencia del estudiante. La y tiene dos valores posibles, ganar o perder la competencia, que se codifica como 1 y 0, respectivamente. Entonces, dada la perspectiva del aprendizaje automático, la selección de candidatos para la predicción del desempeño de los estudiantes en competencias académicas puede formularse como un problema de clasificación binaria al agrupar a los estudiantes en dos clases simétricas. Las relaciones entre las variables características y la variable objetivo pueden describirse como la Ecuación (1).

$$y = f(x1, x2, \cdots, xn) + ÿ. \quad (1)$$

donde f representa una función desconocida, que es el modelo de predicción que pretendemos entrenar usando los datos recopilados, ÿ es el error entre el valor real de la variable objetivo y para cada estudiante y el valor predicho del modelo de predicción. El objetivo del problema de clasificación binaria es entrenar un modelo que pueda predecir la salida de la variable objetivo y cuando se le da una serie de variables de entrada x. La predicción ypre de un nuevo estudiante xnew se puede lograr mediante ypre = f(xnew).

En este estudio, se propone un modelo de conjunto novedoso basado en el apilamiento para predecir el rendimiento de los estudiantes. en concursos académicos. Este estudio trata de responder a dos preguntas: • ¿Cómo

podemos mejorar el rendimiento de predicción del modelo con una estrategia de conjunto? • Los resultados

predictivos deben ser comprensibles. Entonces, ¿cómo podemos identificar características significativas? ¿Qué sugerencias se pueden hacer para los tutores/gerentes en función de los resultados del análisis?

En este documento, la Sección 4 presenta en detalle el modelo de conjunto propuesto. En la Sección 5, describimos el flujo de análisis detallado del modelo de aprendizaje de conjuntos propuesto para un conjunto de datos reales. Se verificó la validez y robustez del modelo, y el modelo realizó el análisis de importancia de las características. Los resultados experimentales se presentan y discuten en la Sección 5. En la Sección 6, concluimos y presentamos una breve perspectiva para futuros estudios.

4. Método

Diferentes algoritmos de aprendizaje arrojan resultados diferentes en los problemas de regresión o clasificación. Como los resultados de aprendizaje de diferentes algoritmos son diferentes, es posible mejorar el rendimiento de predicción final para cada algoritmo, de modo que se pueden obtener mejores resultados con los algoritmos de aprendizaje combinados en comparación con los resultados obtenidos usando un solo algoritmo [22]. El aprendizaje en conjunto está diseñado para impulsar el rendimiento predictivo al combinar las predicciones de múltiples algoritmos.

El aprendizaje de conjuntos se ha utilizado comúnmente en el aprendizaje automático en una variedad de tareas de clasificación y regresión para mejorar el rendimiento mediante la agrupación de algoritmos individuales. Se han propuesto varios tipos de métodos de conjunto, como votación, promediación, embolsado, aumento y apilamiento. El apilamiento es un método de conjunto heterogéneo eficiente. Ha sido ampliamente utilizado en competencias de minería de datos en los últimos años. Se puede considerar como un súper perceptrón multicapa. Cada capa incluye uno o varios modelos, y la siguiente capa aprende de los resultados de la capa anterior del modelo. En el aprendizaje automático, se utilizan muchos modelos para resolver un problema de clasificación binaria, incluido el algoritmo de regresión , el algoritmo de árbol de decisión, el algoritmo basado en kernel, el algoritmo del método bayesiano, el algoritmo de agrupamiento , etc. El apilamiento se puede integrar fácilmente con diferentes clasificadores o modelos de regresión para mejorar la robustez y la generalización en un solo modelo.

4.1. El modelo de conjunto propuesto

En este estudio, construimos un modelo de conjunto utilizando apilamiento de 2 capas. Los algoritmos de aprendizaje utilizados en la primera capa se denominan Base-learner, y el algoritmo de la segunda capa se denomina

el Meta-aprendiz respectivamente. El meta-aprendiz se utiliza para combinar los resultados de predicción de todos los aprendices básicos.

Los procedimientos de construcción del modelo de predicción pueden describirse mediante los siguientes pasos: Paso 1. Dividir el conjunto de datos en conjuntos de entrenamiento y conjuntos de prueba; Paso 2. Construya un nuevo conjunto de datos basado en el resultado de los estudiantes de Base; Paso 3. Entrene al Meta-aprendiz para generar el resultado de la predicción final basado en el conjunto de datos recién construido.

El modelo de apilamiento de 2 capas se puede usar para combinar algoritmos de aprendizaje automático para aumentar la precisión predictiva. Al construir el modelo de conjunto de apilamiento, la selección de Base-learner y Metal-learner afecta el rendimiento del modelo.

La premisa de utilizar el apilamiento para mejorar el efecto de clasificación es que el alumno básico debe tener un buen rendimiento de predicción. En general, el apilamiento funcionará mejor si los algoritmos que se combinan son, en algún sentido, muy diferentes entre sí. Por lo tanto, cuanto mayores sean las diferencias en el principio de clasificación entre estos alumnos de base, más complementarios pueden ser en el proceso de un conjunto. Los resultados de la clasificación se optimizarán en consecuencia. Estas estrategias se aplicaron para seleccionar a los estudiantes de base en la primera capa. Los clasificadores Random forest (RF) [23–25], SVM [26–28] y AdaBoost [29,30] se eligieron en este estudio para ser los aprendices básicos debido a sus mejores rendimientos de clasificación. En la actualidad, se adoptan comúnmente como modelos predictivos para la predicción del rendimiento de los estudiantes y tienen un buen rendimiento. Los algoritmos RF y AdaBoost se utilizaron para la clasificación, ya que ambos son excelentes algoritmos de conjunto basados en árboles de decisión, mientras que RF produce múltiples árboles de decisión basados en un subconjunto seleccionado al azar de muestras y variables de entrenamiento.
RF [27] no necesita asumir una distribución de datos y puede manejar miles de variables de entrada sin eliminación de variables. Sin embargo, la estructura de árbol de RF es inestable y puede sobreajustar los datos de entrenamiento y, por lo tanto, su rendimiento de generalización es deficiente. El SVM es un modelo de aprendizaje automático potente y flexible y puede realizar una clasificación lineal o no lineal. El SVM es particularmente adecuado para la clasificación de datos complejos con conjuntos de datos de tamaño pequeño o mediano. Aunque los clasificadores SVM lineales son eficientes y funcionan bien en muchas aplicaciones, solo funcionan para los datos separables linealmente y son muy sensibles a los valores atípicos. El AdaBoost enfatiza la adaptabilidad modificando con frecuencia los pesos de la muestra y agregando clasificadores débiles al impulso. El AdaBoost es sensible a datos con ruido y valores atípicos, y es menos susceptible al problema de sobreajuste que la mayoría de los demás algoritmos de aprendizaje. Estos tres modelos tienen sus ventajas y desventajas cuando se aplican a la predicción del desempeño de los estudiantes. El Apéndice A de este documento proporciona elaboraciones detalladas de los algoritmos SVM, RF y AdaBoost. El alumno base en el apilamiento utilizó un modelo bien realizado en busca de un aprendizaje adecuado durante el entrenamiento de datos. Por lo tanto, el apilamiento es más propenso a sobreajustar los datos de entrenamiento. Para reducir el riesgo de sobreajuste, el Meta-aprendiz tiende a seleccionar modelos simples, como la regresión logística , la regresión de lazo, etc. Elegimos la regresión logística como el Meta-aprendiz en nuestro estudio. La regresión logística es el modelo básico para la predicción de una variable aleatoria dependiente dicotómica. La regresión logística no tiene una alta precisión general y es propensa a no ajustarse bien a los datos de entrenamiento. Además, no es adecuado para tratar características no lineales [31]. Sin embargo, la regresión logística es una buena opción para la predicción del éxito en un curso o programa [32]. El Apéndice A de este documento proporciona una elaboración detallada de los algoritmos de regresión logística.

Finalmente, los clasificadores SVM, RF y AdaBoost se utilizaron como alumnos base y se adoptó la regresión logística como el alumno meta en este estudio. Se construyó un modelo de conjunto usando apilamiento como se ilustra en la Figura 1.

La variable característica x representa los datos del estudiante, y los datos se clasifican en tres tipos principales de información. El primer tipo es la información de antecedentes del estudiante, el segundo tipo son los datos de comportamiento del estudiante durante la competencia y el último tipo es el rendimiento académico diario del estudiante . Estas variables sin procesar se dividen luego en dos tipos, uno es numérico y el otro es categórico, como se enumera en la Tabla 1.

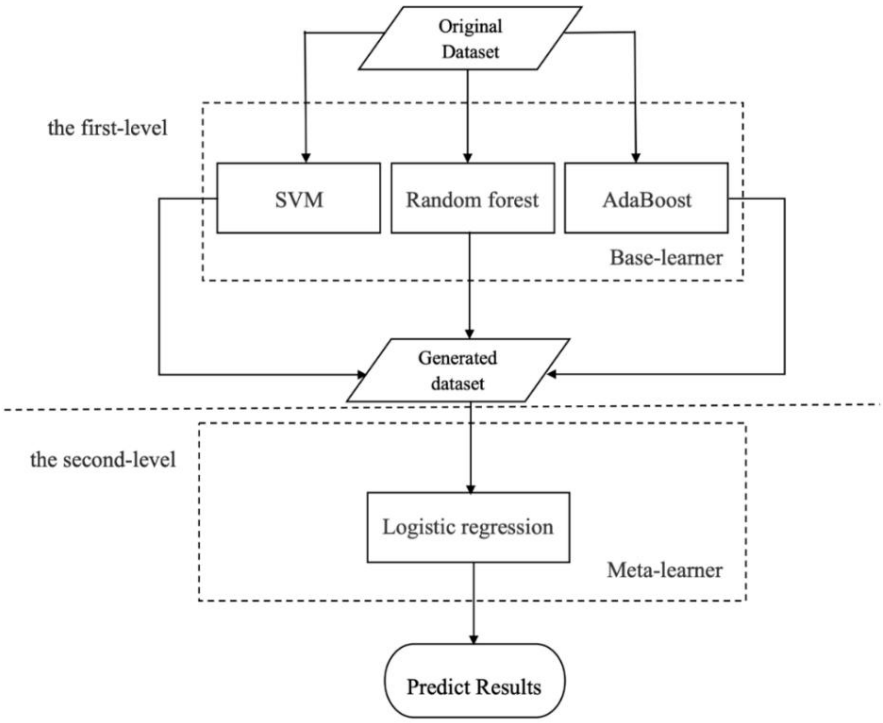


Figura 1. El modelo de predicción por conjuntos propuesto.

Figura 1. El modelo de predicción por conjuntos propuesto.

La variable característica *x* representa los datos del estudiante, y los datos se clasifican en tres principales tipos de información. El primer tipo es la información de antecedentes del estudiante, el segundo tipo son los datos del comportamiento del estudiante durante la competencia, y el último tipo es el rendimiento académico diario del estudiante. Estas variables sin procesar se dividen luego en dos tipos, uno es numérico y el otro de género Masculino, Femenino uno es categórico, como se enumera en la Tabla 1.

Tabla 1. Descripción de las variables crudas. En la columna Tipo, la N denota una variable numérica y la C denota una variable categórica.

Categoría	Variable	Tipo	Descripción
Fondo	género	C	género C Masculino, Femenino
	source_l	C	Lugar de residencia de los participantes en el examen de ingreso a la universidad
	competición_n	norte	especialidad Informática, Derecho, Inglés, E-commerce, etc.
	inscripción_c	C	Categoría temática (materias, literatura, historia, ingeniería, etc.)
	pares_b	C	Calificación en el momento de la primera participación en el concurso de la asignatura.
	competencia_g	C	Grado en el momento de la primera participación en el competencia temática.
	competencia_t	C	Estudiante de primer año, estudiante de segundo año, estudiante de tercer año, mayor
	competencia_n	norte	Número total de participantes en competencias académicas
Conducta	pre_c_GPA	norte	GPA para todos los semestres antes de la participación en el competencia
	peer_bc	C	Si un compañero de esta carrera ha participado en la misma competencia en el pasado? Si o no
	c_GPA	norte	GPA del semestre de competencia
	competencia_t	C	Ya sea para participar en la misma competencia varias veces? Si o no
GPA (punto de calificación Promedio)			
Resultado	resultado_c	C	Si ganar en la competencia académica

Al principio, cada estudiante base proporciona una probabilidad de que una muestra pertenezca a cada clase, P (y = 1 x) o P(y = 0 x). Cada modelo de clasificación tiene dos probabilidades de predicción de atributos, por lo que seis predicciones las probabilidades se obtienen en total para cada muestra de estudiantes. En segundo lugar, la probabilidad del atributo predicho los valores de Base-learner se proporcionan para el Meta-learner como su entrada. La etiqueta de clase real y aprender y luego obtener el modelo de predicción final M. Al final, la etiqueta de clase final de predicción los resultados serán dados por el **Mi.**

Al principio, cada alumno base proporciona una probabilidad de que una muestra pertenezca a cada clase, (= 1 |) o (= 0 |). Cada modelo de clasificación tiene dos probabilidades de predicción de atributos, por lo que seis predicciones las probabilidades se obtienen en total para cada muestra de estudiantes. En segundo lugar, el atributo predicho se usa para los valores de probabilidad del alumno base se proporcionan para el alumno meta como su entrada. La etiqueta de clase real se usará para aprender y luego obtener el modelo de predicción final. Al final, la etiqueta de clase final de los resultados de predicción estará dada por el ÿ.

4.2. Validación cruzada de 10 × 10 veces

Symmetry 2020, 12, x PARA REVISIÓN POR PARES 6 de 17

4.2. Validación cruzada de 10 × 10 veces

En la validación cruzada de 10 veces, el conjunto de datos se divide aleatoriamente en 10 partes iguales. A su vez, nueve de ellos se toman como conjunto de entrenamiento y el otro se utiliza como conjunto de prueba. El valor medio de los resultados predictivos de los 10 turnos se considera como el resultado de la evaluación del modelo [33].

Mientras entrena a un meta-aprendiz, los resultados de la prueba del conjunto de entrenamiento deben caracterizarse por alumnos de la capa anterior. Si entrenamos al alumno y luego lo predecimos en un conjunto de entrenamiento, esto provocará una fuga de etiquetas. Una fuga de etiquetas significa que los modelos de entrenamiento y los resultados de predicción muestran la información personal de los participantes que se encuentran en el conjunto de datos. Se debe evitar una fuga en la etiqueta cuando se aplica el apilamiento. Para evitar la fuga de etiquetas, usamos otra validación cruzada de 10 veces en cada conjunto de entrenamiento. Como se muestra en la Figura 2, cada estudiante base utiliza una validación cruzada de 10 veces para generar un nuevo característica en este estudio.

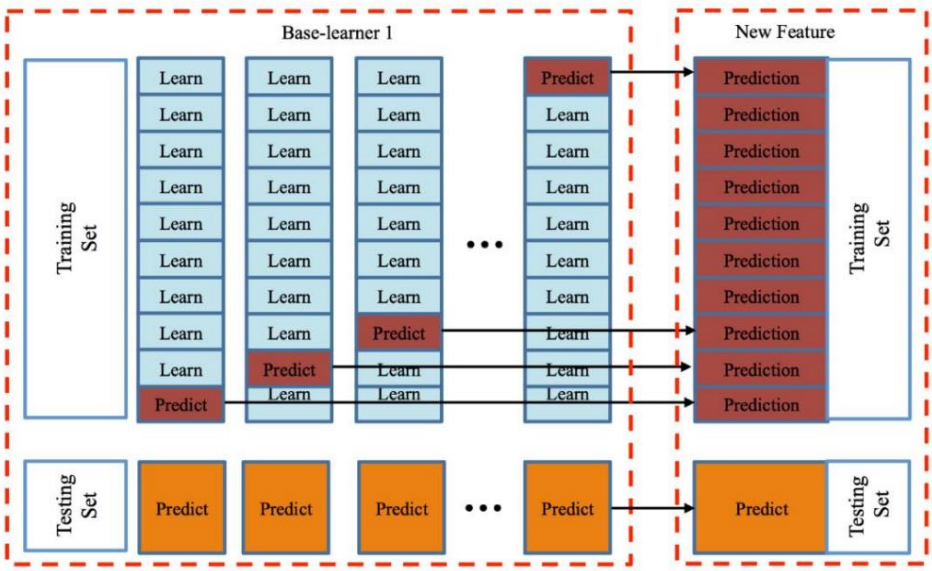


Figura 2. El modelo de validación cruzada de 10 × 10 veces utilizado como alumno base en este estudio.

Dado que el conjunto de muestra inicial es (yn, xn), n = 1, 2, · · · , donde yn es una etiqueta de clase (yn {0, 1}) y representa la variable característica. El conjunto de muestras se divide aleatoriamente en 10 subconjuntos que el mismo tamaño y los subconjuntos se definen como {S1, S2, · · · S10}. Elegimos uno de estos subconjuntos como cada vez, y los subconjuntos restantes se consideraron como conjuntos de entrenamiento. Usamos la cruz de 10 veces cada 10 veces como se muestra en la Figura 2 para cada conjunto de entrenamiento original. Tomando el aprendiz base como ejemplo, cada ronda de entrenamiento y prueba producirá dos resultados posibles, Pk (y = 0 xi) y Pk (y = 1 xi). Aquí , Pk (y = 0 xi) y Pk (y = 1 xi) están dados por el k-ésimo alumno base, y (= 0|) y (= 1|) están dados por el aprendiz Baseth y ellos (= 1|) . Aquí el (= 0|) y (= 1|) están representan las probabilidades del resultado de la predicción de la i-ésima muestra perteneciente a la clase con etiqueta y = 0 (perdedor en competencia) e y = 1 (ganador en competencia), respectivamente. Una vez que la predicción obtengan los resultados de la predicción de los tres estudiantes base, todas las probabilidades se concatenarán juntas para formar un nuevo vector de características.

Como se muestra en la Figura 3, el vector de características se expresa en forma de una matriz Ntr × 6, representada por L en este estudio. En la matriz, Ntr es el número de muestras en el conjunto de entrenamiento, y cada Base por *L* en este estudio. En la matriz, el alumno aporta dos dimensiones.

$$\begin{bmatrix} 1(=0|\ 1) & 1(=1|\ 1) & 2(=0|\ 1) & 2(=1|\ 1) & 3(=0|\ 1) & 3(=1|\ 1) \\ 1(=0|\ 2) & 1(=1|\ 2) & 2(=0|\ 2) & 2(=1|\ 2) & 3(=0|\ 2) & 3(=1|\ 2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1(=0|\ ) & 1(=1|\ ) & 2(=0|\ ) & 2(=1|\ ) & 3(=0|\ ) & 3(=1|\ ) \end{bmatrix}$$

Figura 3. El vector de características en forma de matriz Ntr×6.

5. Estudios experimentales

5.1. Recopilación de datos

5. Estudios experimentales

5.1. Recopilación de datos

Los datos se recopilaron de la Universidad de Hankou e incluyeron los registros de estudiantes de pregrado que habían participado en al menos una competencia académica nacional de 2015 a 2018 Los datos se recopilaron de la Universidad de Hankou e incluyeron los registros de estudiantes de pregrado

año académico. En total, 684 registros de competencia que contienen 486 estudiantes están presentes en los datos. Que hayan participado en al menos un concurso académico nacional de 2015 a 2018

Son 86 concursos académicos que abarcan múltiples categorías, como ingeniería, literatura, ciencia, año. En total, 684 registros de competencia que contienen 486 estudiantes están presentes en los datos. Existen

arte, y así sucesivamente. Además, se incluyen algunos concursos integrales, como concursos académicos de innovación 86 que cubren múltiples categorías, como ingeniería, literatura, ciencia, arte,

y competencia empresarial, competencia de habilidades, etc. Consideramos el primer premio, el segundo premio, y así sucesivamente. Además, se incluyen algunos concursos integrales, como el de innovación y

competencia empresarial, competencia de habilidades, etc. Consideramos el 1er premio, el 2do premio y el el 3er premio otorgado en los concursos como ganador en este estudio

Para cada alumno fueron definidas once variables, conforme descrito en la Tabla 1. Diez de ellas son 3º premio otorgado en los concursos como vencedores en este estudio.

variables características, y se derivan de la información demográfica y del comportamiento Para cada estudiante se definieron once variables, como se describe en la Tabla 1. Diez de ellas son

información de los estudiantes. La información anterior se obtuvo de las variables características del estudio diario de los estudiantes , y se derivan de la información demográfica y del comportamiento.

desempeño y comportamiento en las competencias participantes. Estas variables fueron tomadas como insumos de la información de los estudiantes. La información anterior se recopiló del rendimiento de estudio diario de los estudiantes.

modelo de predicción. El resultado de la competencia se definió como la variable objetivo. El enmascaramiento de datos fue y el comportamiento en las competencias participantes. Estas variables se tomaron como entradas de la predicción.

aplicado al conjunto de datos para enmascarar información privada o confidencial antes del análisis. modelo. El resultado de la competencia se definió como la variable objetivo. Se aplicó el enmascaramiento de datos a

el conjunto de datos para enmascarar información privada o confidencial antes del análisis.

5.2. Preprocesamiento de datos y análisis de descripción.

5.2. Preprocesamiento de datos y análisis de descripción

El preprocesamiento de datos se llevó a cabo en los siguientes pasos: Al principio, siete de estas variables sin procesar

eran variables categóricas, incluidos género, especialidad, categoría de materia, etc. Se convirtieron en El preprocesamiento de datos se llevó a cabo en los siguientes pasos: Al principio, siete de estas variables sin procesar

variables categóricas, que incluyen género, especialidad, categoría de materia, etc. Se convirtieron en los valores numéricos por codificación entera. Por ejemplo, Freshman, sophomore, junior y senior son

numéricos faltantes por codificación de enteros. Por ejemplo, los estudiantes de primer año, segundo año, tercer año y último año son establecido en 1, 2, 3 y 4, respectivamente. En segundo lugar, se utilizaron conjuntos de elementos frecuentes para sustituir los valores

establecido en 1, 2, 3 y 4, respectivamente. En segundo lugar, se utilizaron conjuntos de elementos frecuentes para sustituir los valores perdidos. valores.

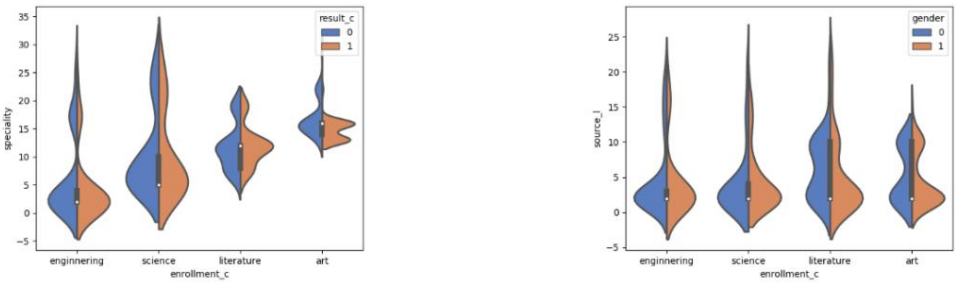
Las distribuciones de densidad de especialidad y matrícula_c se agruparon según los resultados de

distribuciones de densidad de especialidad y matrícula_c se agruparon según los resultados de la competencia como se muestra en la Figura 4a. Las distribuciones de densidad de fuente_l y inscripción_c fueron

la competencia como se muestra en la Figura 4a. Las distribuciones de densidad de fuente_l y inscripción_c fueron agrupados según el género como se muestra en la Figura 4b. En general, la inscripción_c y el género

agrupados según el género como se muestra en la Figura 4b. En general, la inscripción_c y el género de los estudiantes están distribuidos uniformemente y los estudiantes de ingeniería son propensos a participar en competencias.

de los estudiantes están distribuidos uniformemente y los estudiantes de ingeniería son propensos a participar en competencias.



(a) **(B)**

Figura 4. (a) Distribuciones de densidad de especialidad y matrícula_c agrupadas según los resultados de la competencia, **(b)** distribuciones de densidad de fuente_l y matrícula_c agrupadas según el género.

Figura 4. (a) Distribuciones de densidad de especialidad y matrícula_c agrupadas según los resultados de la competencia, **(b)** distribuciones de densidad de fuente_l y matrícula_c agrupadas según el género.

La Tabla 2 muestra los resultados del análisis descriptivo de los datos preprocesados de los cuatro años académicos.

La Tabla 2 muestra los resultados del análisis descriptivo de los datos preprocesados de los cuatro años académicos.

La asimetría se utilizó para calcular la dirección de la asimetría y el grado de distribución de datos estadísticos, y es un

La asimetría se utilizó para calcular la dirección de la asimetría y el grado de distribución de datos estadísticos, y es un característica digital del grado de asimetría de la distribución de datos estadísticos. Como se muestra en la Tabla 2, la mayoría

es una característica digital del grado de asimetría de la distribución de datos estadísticos. Como se muestra en la Tabla 2, la mayoría de los estudiantes eligieron participar en competencias académicas durante el segundo y tercer año. Diferente

la mayoría de los estudiantes eligió participar en competencias académicas durante los años de segundo y tercer año. Diferentes estudiantes tienen un número diferente de participaciones como se muestra en la Tabla 2. Por ejemplo, participa en más

los estudiantes tienen un número diferente de participaciones como se muestra en la Tabla 2. Por ejemplo, un estudiante puede de una competencia académica

el estudiante puede participar en más de una competencia académica.

Las distribuciones de frecuencia de pre_c_GPA y c_GPA se muestran en la Figura 5a,b, respectivamente.

En resumen, tanto el pre_c_GPA como el c_GPA están bien descritos por la distribución normal. En total, 159 de **Tabla 2.** Estadísticas descriptivas de los datos preprocesados.

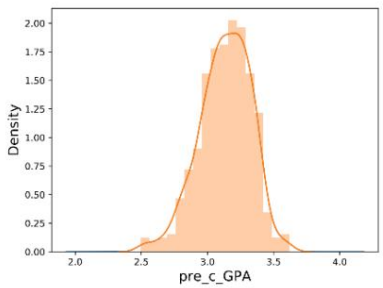
Participaron 486 estudiantes ganados en los concursos académicos nacionales.

Variable	desv.	mín.	25%	50%	75%	máx.
género	0,45	0.50	0.00	0,00	0,00	1,00	1.00
competencia_g	2.24		0.79	1.00	2,00	2,00	3,00	4,00	
competencia_n	1.60	peer_b	1.28	1.00	1,00	1,00	2,00	11,00	
		0.76	0.43	0.00	1,00	1,00	1,00	1.00

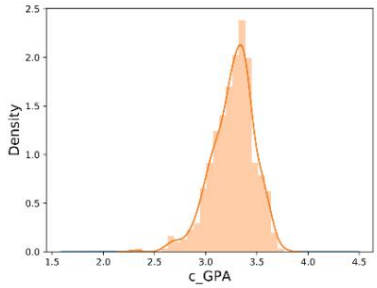
Tabla 2. Estadísticos descriptivos de los datos preprocesados.

Variable	Desv estándar media	mín.	25%	50%	75%	máx.
género	competencia_t 0,13	0,34 0,00	0,00 0,00 0,00	0,00 0,00	1.00	1.00
competencia_g 2.24 1.00 1.00	**c_GPA 0.21** competencia_c 1.60	3,62	2,00 2,00 3,00	2,50 3,01	3,15 3,27	4.00
resultado_c 0.55 competencia_t 0.13 0.00 0.00	respectivamente. El resultante	3,77	1,00 1,00 2,00	2,31 3,14	3,29 3,40	11.00
el c_GPA como el 2.31 estudiantes que los	estudiantes que los	1,00	1,00 1,00 1,00	0,00 0,00	0,00 1,00	1.00
académicas nacionales.	ganaron en las competencias		0,00 0,00 0,00	1,00		
	0,34					
	0,19		3,01 3,15 3,27	3,62		
3.27	0,21		3,14	3,29	3,40	3,77
0.36	0,55	0.00	0.00	0.00	1.00	1.00

y se muestran en la Figura 5a,b, están bien descritos por la distribución normal.



(a)



(B)

Figura 5. (a) La distribución de frecuencias de ; **(b)** la distribución de frecuencias de la **Figura 5. (a)** La distribución de frecuencias de pre_c_GPA; **(b)** la distribución de frecuencias de c_GPA.

5.3. Indicadores de rendimiento de clasificación

5.3 Indicadores de rendimiento de clasificación

Los modelos considerados en este estudio fueron evaluados por cinco medidas de rendimiento: Los modelos considerados en este estudio fueron evaluados por cinco medidas de desempeño: precisión, recuperación, F1, error y área bajo la curva característica operativa del receptor (AUC). Las salidas de los clasificadores de clasificadores se resumen en cuatro grupos: Los estudiantes que ganaron la competencia fueron correctamente etiquetado como ganador ( ), los estudiantes que perdieron la competencia fueron etiquetados incorrectamente como ganadores (FP); los estudiantes que perdieron la competencia son etiquetados correctamente como perdedores ( ); los estudiantes que ganaron fueron etiquetados incorrectamente como perdedores (FN).

La *Precisión* y la *Recuperación* se calculan mediante las Ecuaciones (2) y (3) de la siguiente manera:

$$\text{Precisión} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recordar} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

La ecuación (2) indica la métrica de precisión como la fracción de estudiantes que ganaron la competencia sobre los estudiantes predichos como el ganador. De hecho, a mayor número de *FP*, menor precisión de la clasificación. Además, para investigar la capacidad del modelo de clasificación para predecir todos los ganadores, usamos la métrica de recuperación como se define en la Ecuación (3). En otras palabras, la métrica de recuperación representa la fracción de los estudiantes que están correctamente etiquetados como ganadores sobre el número total de ganadores. Es la fracción de los estudiantes que están correctamente etiquetados como ganadores sobre el número total de ganadores. vale la pena mencionar que más estudiantes etiquetados como ganadores darán lugar a un recuerdo alto y bajo vale la pena mencionar que más estudiantes etiquetados como ganadores darán lugar a un recuerdo alto y bajo precisión. El promedio armónico de precisión y recuperación, llamado F1, también se considera en este estudio de precisión. El promedio armónico de precisión y recuperación, llamado F1, también se considera en este estudio y se describe como:

$$F1 = 2 \times \frac{\text{Precisión} \times \text{Recuperación}}{\text{Precisión} + \text{Recuperación}} \quad (4)$$

Recuperación La *F1* estima la calidad de la clasificación tanto para el ganador como para el perdedor, simultáneamente. La ecuación (5) a continuación define la medida *Error,* lo que significa la proporción de estudiantes etiquetados incorrectamente como ganadores y estudiantes etiquetados incorrectamente como perdedores sobre todos los estudiantes participantes.

$$= \frac{+}{+ + +} \quad (5)$$

Una curva característica operativa del receptor (ROC) es un enfoque gráfico para analizar el rendimiento de un clasificador. Utiliza un par de estadísticas (tasa de verdaderos positivos y tasa de falsos positivos) para caracterizar el rendimiento de un algoritmo de clasificación. El gráfico resultante se puede utilizar para comparar la

La F1 estima la calidad de la clasificación tanto para el ganador como para el perdedor, simultáneamente. Ecuación (5) a continuación define la medida Error, que significa la proporción de estudiantes mal etiquetados como ganador y los estudiantes etiquetados erróneamente como perdedores sobre todos los estudiantes participantes.

$$\text{Error} = \frac{FP + FN}{TP + TN + FP + FN} \quad (5)$$

Una curva característica operativa del receptor (ROC) es un enfoque gráfico para analizar la rendimiento de un clasificador. Utiliza un par de estadísticas (tasa de verdaderos positivos y tasa de falsos positivos) para caracterizar el rendimiento de un algoritmo de clasificación. El gráfico resultante se puede utilizar para comparar la desempeño relativo de diferentes clasificadores y para determinar si un clasificador se desempeña mejor que adivinanzas al azar. La curva ROC no se ve afectada por los cambios en la distribución de la muestra. Las AUC El valor representa el área bajo la curva ROC. Cuanto mayor sea el valor de AUC , mejor será la clasificación el algoritmo es. El AUC es equivalente a la probabilidad del caso de que un resultado positivo seleccionado al azar ejemplo se clasifica más alto que un ejemplo negativo seleccionado al azar [34].

5.4. Resultados

En este estudio, se desarrolló un modelo de conjunto que utiliza el apilamiento utilizando el lenguaje Python dentro de el marco de la biblioteca Sklearn en PyCharm. PyCharm es un entorno de desarrollo integrado de Python con un conjunto de software que puede ayudar a los usuarios a mejorar su eficiencia al desarrollar en Python. El El modelo de conjunto se construyó utilizando RF, SVM y AdaBoost como sus tres estudiantes base, y Se seleccionó la regresión logística como el Meta-aprendiz. Para probar la eficacia y la estabilidad de la modelo de conjunto propuesto, se llevaron a cabo 10 experimentos de predicción. Después de realizar las 10 rondas del modelo de conjunto propuesto, se obtuvieron los resultados de predicción, tal como se enumeran en la Tabla 3. El promedio el valor de las 10 rondas se calculó como el rendimiento de predicción general del conjunto propuesto modelo. Como se muestra en la Tabla 3, la Precisión, Recuperación, F1, Error y AUC son 0.8550, 0.8600, 0.85, 0.1460, y 0,9185, respectivamente.

Tabla 3. Resultados de las pruebas del modelo de conjunto propuesto a partir de 10 ejecuciones.

Ejecutar índice	Etiqueta de clase	Precisión	Recuerdo	F1	Error	ABC
1	0	0.8400	0.8800	0.8600	0.1429	0.9202
	1	0.8700	0.8400	0.8500		
2	0	0.8400	0.8700	0.8600	0.1480	0.9206
	1	0.8600	0.8400	0.8500		
3	0	0.8500	0.8700	0.8600	0.1429	0.9157
	1	0.8600	0.8600	0.8500		
4	0	0.8400	0.8600	0.8500	0.1531	0.9191
	1	0.8500	0.8400	0.8400		
5	0 1	0.8400	0.8700	0.8600	0.1480	0.9199
		0.8600	0.8400	0.8500		
6	0	0.8400	0.8800	0.8600	0.1429	0.9159
	1	0.8700	0.8400	0.8500		
7	0	0.8500	0.8700	0.8600	0.1480	0.9192
	1	0.8600	0.8500	0.8500		
8	0	0.8500	0.8900	0.8700	0.1378	0.9160
	1	0.8800	0.8400	0.8600		
9	0	0.8500	0.8800	0.8600	0.1378	0.9196
	1	0.8700	0.8500	0.8600		
10	0 1	0.8500	0.8800	0.8700	0.1378	0.9187
		0.8700	0.8500	0.8600		
mínimo		0.8400	0.8400	0.8400	0.1378	0.9157
máx.		0.8800	0.8900	0.8700	0.1531	0.9206
Cra		0.8550	0.8600	0.8565	0.1439	0.9185

Los valores de AUC se clasifican en tres niveles de rendimiento con los siguientes umbrales: AUC > 0,9 (excelente), 0,7 < AUC < 0,9 (regular) y AUC < 0,7 (pobre). Los resultados que se muestran en la Tabla 3 indican que la El modelo de conjunto propuesto tiene un mejor rendimiento de predicción.

Para obtener una mayor claridad del rendimiento de la predicción para el modelo propuesto, un estudio comparativo se llevó a cabo, en el que se seleccionaron varios algoritmos para la comparación de rendimiento. Estos seleccionados los algoritmos únicos se usan comúnmente en los estudios existentes para la predicción del rendimiento de los estudiantes, incluyendo SVM, árbol de decisión, regresión logística y Bernoulli Naive Bayes (BernoulliNB). El

Los resultados experimentales del estudio comparativo se resumen en la Tabla 4.

Tabla 4. Resultados de clasificación de comparación con algoritmos únicos.

Tabla 4. Resultados de clasificación de comparación con algoritmos únicos.

Método	Etiqueta de clase	Precisión	Recuperación	F1	Error	ABC
MVS	0	0,8202	0,8384	0,8058	0,2041	0.8444
	1	0,7826	0,7526	0,7849		
Árbol de decisión	0	0,7419	0,7273	0,7539	0.2395	0.7605
	1		0,7938	0,7662		
Regresión logística	0		0,6970	0.7188	0.2275	0.7725
	1	0,7087	0,7526	0.7300		
BernoulliNB	0	0,7500	0,6667	0,7059	0.2351	0.7649
	1	0,6944	0,7732	0.7317		
Modelo propuesto	0	**0,8710**	**0,8351**	**0.8614**	**0.1429**	**0.9138**
	1			**0.8526**		

Método Etiqueta de clase Precisión Recuperación F1 Error ABC
MVS 0 0,7757 0,8058
1 0,8202 0,8202 0,8384 0,8058 0,2041 0.8444
0 0,7826 0,7404 0,8384 0,7526 0,7849
Árbol de decisión 1 0,7419 0,7419 0,7273 0,7273 0,7539 0.2395 0.7605
0 0,7938 0,7938 0,7662
Regresión logística 1 0,6970 0,7188 0.2275 0.7725
0 0,7087 0,7087 0,7526 0,7526 0.7300
BernoulliNB 1 0,7500 0,7500 0,6667 0,6667 0,7059 0.2351 0.7649
0 0,6944 0,6944 0,7732 0,7732 0.7317
Modelo propuesto 1 **0,8447** **0,8447** **0,8788** **0,8788**
0 **0,8710** **0,8351** **0.8614** **0.1429** **0.9138**
1 **0.8710** **0.8351** **0.8526** **0.1429** **0.9138**

Los resultados revelan que el rendimiento de predicción del modelo de conjunto propuesto en este estudio es mejor que los otros cuatro modelos con un solo algoritmo cada uno. Los resultados de predicción en este estudio son se utilizan para la selección de candidatos de concurso, por lo tanto, el modelo se centra en los méritos de los utilizados indicadores etiquetados como 1. La *Precisión* del modelo propuesto obtiene el valor más alto de 0.8710, que se etiqueta seguido por el modelo SVM. Usando el modelo propuesto, la probabilidad de que los estudiantes estén correctamente en etiquetado como 1 es 87.1%. El valor de *Recall* en el modelo propuesto es 0,8351, al que le sigue el árbol de decisión. Entre los cinco modelos, el modelo propuesto tiene el mejor desempeño con el valor AUC de 0.9138. Al mismo tiempo, el modelo propuesto tiene la tasa de error más baja en comparación con otros algoritmos.

De acuerdo con la comparación de las curvas ROC, como se muestra en la Figura 6, las áreas bajo la curva ( ) de los cinco modelos son mucho mayores que 0,5.

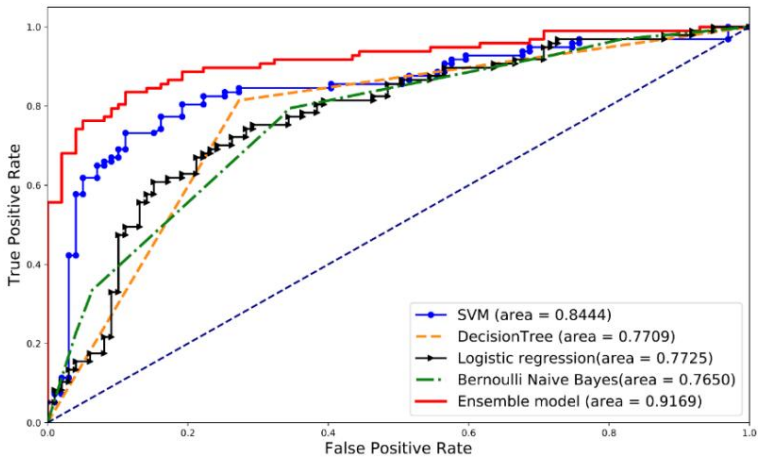


Figura 6. Curva característica de funcionamiento del receptor del modelo propuesto y otros algoritmos comparados.

Figura 6. Curva característica de funcionamiento del receptor del modelo propuesto y otros comparados algoritmos

Realizamos un estudio de comparación entre el modelo de conjunto propuesto y varias curvas populares y ROC algoritmos de conjunto, que incluyen aumento de gradiente, RF, AdaBoost y XGBoost. Los resultados promedio y las curvas ROC de 10 rondas se enumeran en la Tabla 5 y la Figura 7, respectivamente.

Realizamos un estudio de comparación entre el modelo de conjunto propuesto y varios algoritmos de conjunto, que incluyen aumento de gradiente, RF, AdaBoost y XGBoost. Los resultados promedio de 10 rondas que se enumeran en la Tabla 5 y la Figura 7, respectivamente.

Tabla 5. Comparación entre el método propuesto y otros algoritmos de conjunto.

Modelo	Media ± Desv. estándar				
	Precisión	Recuerdo	F1	Error	ABC

Tabla 5. Comparación entre el método propuesto y otros algoritmos de conjunto.

Modelo	Media ± Desv. estándar				
	Precisión	**Recuerdo**	**F1**	**Error**	**ABC**
Aumento de gradiente	0,8295 ± 0,0430	0,8028 ± 0,0502	0,8144 ± 0,0315	0,1765 ± 0,0000	0,8901 ± 0,0222
RandomForest	0,8471 ± 0,0407	0,7997 ± 0,0588	0,8210 ± 0,0336	0,1832 ± 0,0003	0,8826 ± 0,0257
AdaBoost	0,7724 ± 0,0433	0,7835 ± 0,0428	0,7763 ± 0,0223	0,2296 ± 0,0000	0,8374 ± 0,0236
XGBoost	0,8365 ± 0,0389	0,8026 ± 0,0652	0,8170 ± 0,0355	0,1837 ± 0,0000	0,8910 ± 0,0246
Modelo propuesto	0,8600 ± 0,0375	0,8508 ± 0,0431	0,8543 ± 0,0257	0,1439 ± 0,0000	0,9207 ± 0,0177

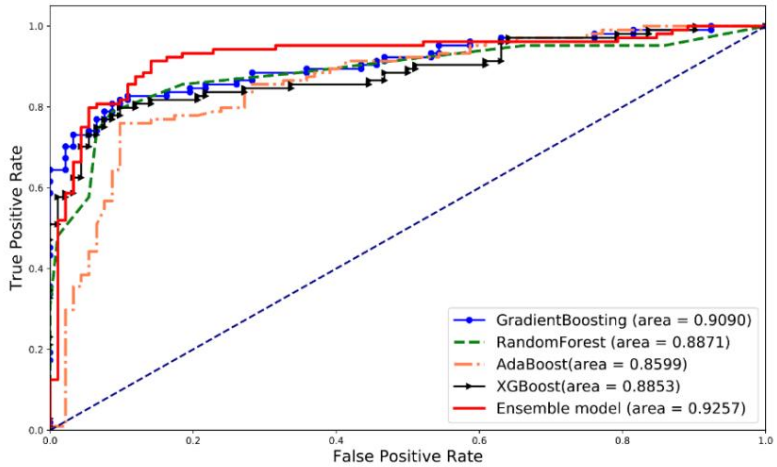


Figura 7. Curvas características de funcionamiento del receptor del modelo propuesto y otros algoritmos de conjunto.

Los resultados de la Tabla 5 y la Figura 7 muestran que el modelo propuesto logra el mejor desempeño en esos cinco indicadores. A continuación, se utilizó una prueba t independiente para comparar estos modelos, como se muestra en la Tabla 6. Las diferencias resultantes entre los modelos se asumieron como estadísticamente significativas cuando $p < 0,05$. Estos resultados muestran las ventajas estadísticamente significativas del método propuesto en comparación con los otros cuatro modelos, particularmente en *F1, Error* y . El modelo propuesto es más eficiente que AdaBoost en y

Tabla 6. Comparación del valor p entre el modelo propuesto y otros algoritmos de conjunto.

Modelo	Valor p (este estudio frente a otros algoritmos)					
	Precisión	**Recuerdo**	**F1**	**Error**	**ABC**	
Aumento de gradiente	0.1277	0.0568	0.0810	0.0400	0.0614	0.0053
RandomForest	0.6565	0,1277	0,0072	0,0000	0,0000	0.0022
AdaBoost	0.0004	0,6565	0,0022	0,1004	0,0274	0.0000
XGBoost	0.2401	0.0004	0,0000	0,0000	0,0072	0.0102
		0.2401	0,1004	0,0274	0,0000	0,0000
					0,0102	

5.5. Análisis de importancia de características

5.5. Análisis de importancia de características

Para mejorar la interpretabilidad del modelo, llevamos a cabo un análisis de importancia de características en este estudio. La contribución de cada variable al desempeño de la predicción se muestra en la Tabla 7. A

Para mejorar la interpretabilidad del modelo, llevamos a cabo un análisis de importancia de características en este análisis de importancia de características que se aplicó para identificar características significativas basadas en un conjunto de datos reales. El estudio. La contribución de cada variable al rendimiento de la predicción se muestra en la Tabla 7. Una característica Los estudiantes básicos generan puntajes de importancia de características de diferentes maneras. La importancia característica de la Se aplicó un análisis de importancia para identificar características significativas basadas en un conjunto de datos reales. El SVM lineal base es de hecho el vector de peso, que contiene los coeficientes, y estos coeficientes definen los estudiantes generan puntajes de importancia de características de diferentes maneras. La importancia característica del vector lineal y ortogonal al hiperplano. El modelo RF mide la importancia de la característica por SVM es de hecho el vector de peso, que contiene los coeficientes, y estos coeficientes definen un calculando el error fuera de bolsa (OOB) correspondiente. El AdaBoost genera puntajes de características por vector ortogonal al hiperplano. El modelo RF mide la importancia de la característica calculando la reducción total normalizada en el error cuadrático medio, que se produce por esa calculando el error fuera de bolsa (OOB) correspondiente. AdaBoost genera puntajes de función por función con la suma de todos los niveles de importancia de la función igual a uno. Los resultados de estos métodos son calcular la reducción total normalizada en el error cuadrático medio, que se produce por ese normalizado, y el valor promedio de los tres puntajes normalizados es el nivel final de importancia de la característica. característica con la suma de todos los niveles de importancia de la característica igual a uno. Los resultados de estos métodos están normalizados y el valor promedio de las tres puntuaciones normalizadas es el nivel final de importancia de la característica. La Figura 8 muestra la clasificación de las características importantes calculadas a partir de diferentes modelos, y las características se muestran en diferentes colores.

Tabla 7. La contribución de cada variable a la reducción.

Symmetry **2020**, *12*, x PARA REVISIÓN POR PARES 12 de 17

La Figura 8 muestra la clasificación de las características importantes calculadas a partir de diferentes modelos, y las características se muestran en diferentes colores.

competencia_t 2.3491 0.1190 0.0400 0.5954

inscripción_c 0.5936 **SVM** 0,0764 0.0400 0.2204

género 0.0640 **Bosque** aleatorio 0.0679

peer_b 0.1639 0.4920 0,0394 0,1331 0.0200

pre_c_GPA 0,0867 0,0572 0.1800 0.5221

competencia_g 2.3491 source_l 0.9936 0,1539 0,1461 0.0200 0,0200 0.3337

competencia_n 0.1086 0,0764 0.2400 0,1800 0.0730

competencia_t 0.0240 0.0400 0,6574 0.6804

inscripción_c 0.0400 0.5954

género 0.4920 especialidad pares_b 0.0163 0,1068 0.1800 0.0200 0,0128 0.2204

acuerdo con los puntajes promedio que se 0,0867 0.0200 0.4387 0.0679

en importancia de características asignadas respectivamente 0.2400 0.3337

muestran en la Figura 8. pre_c_GPA 0.1086 y son los cuales géneros 0.0800 0.6574

fuente_l competencia del estudiante y GPA juegan los roles más 0.0809 0.2128

importantes en el desempeño de la predicción del modelo, mientras que la formación académica es menos importante. 0.0163 0.1068 0.1800 0.4387

Tabla 7. La contribución de cada variable a la predicción.

Características	SVM	Bosque aleatorio	AdaBoost	Puntaje promedio
competencia_t	2.3491	0,1190	0.0400	0.5954
inscripción_c	0.5936	0,0764	0.0400	0.2204
género	0.0640	0,0394	0.0200	0.0679
peer_b	0.1639	0.4920	0,1331	0.1800
pre_c_GPA	0,0867	0,0572	0.1800	0.5221
competencia_g	2.3491	source_l 0.9936	0,1539	0,1461
competencia_n	0.1086	0,0764	0.2400	0,1800
competencia_t	0.0240	0.0400	0.0400	0,6574
inscripción_c	0.0400	0.5954		
género	0.4920	especialidad pares_b	0.0163	0,1068
acuerdo con los puntajes promedio que se	0,0867	0.0200	0.4387	0.0679
en importancia de características asignadas respectivamente	0.2400	0.3337		
muestran en la Figura 8. pre_c_GPA 0.1086 y son los cuales géneros	0.0800	0.6574		
fuente_l competencia del estudiante y GPA juegan los roles más	0.0809	0.2128		
importantes en el desempeño de la predicción del modelo, mientras que la formación académica es menos importante.	0.0163	0.1068	0.1800	0.4387

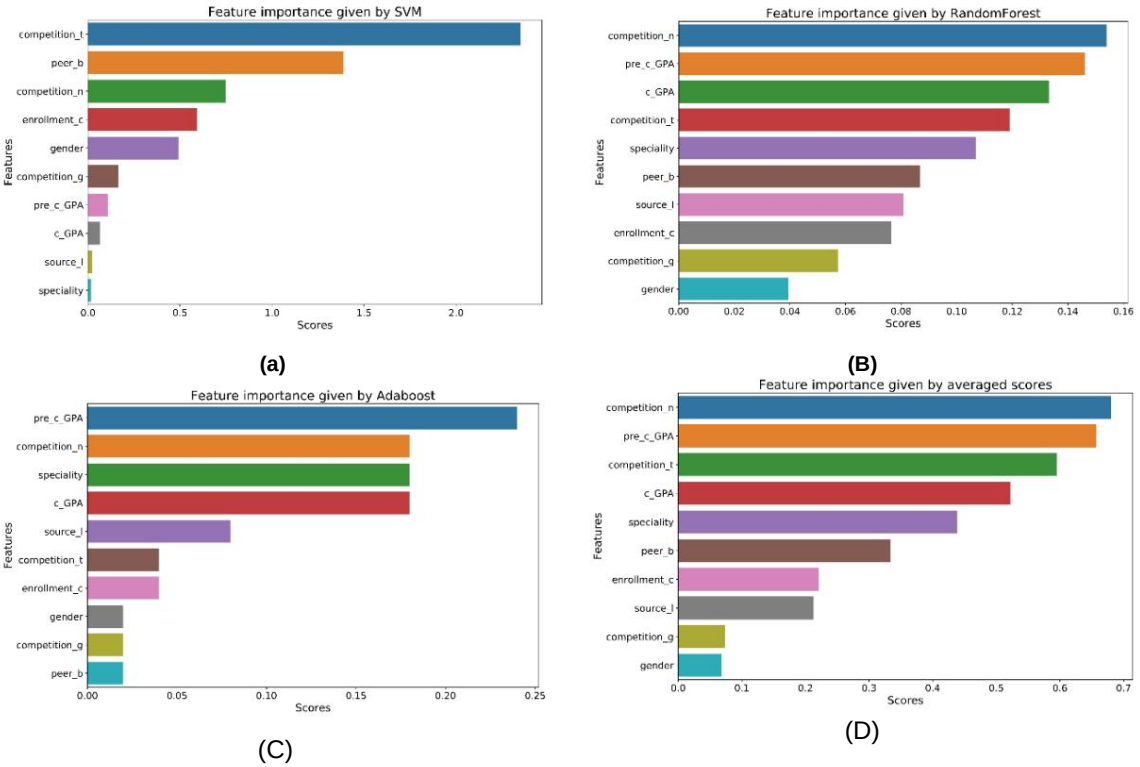


Figura 8. Niveles de importancia de las características otorgadas por la máquina de vectores de soporte (SVM) (a), bosque aleatorio (b), AdaBoost (c), y los valores promedio de las puntuaciones de esos tres modelos (d).

Figura 8. Niveles de importancia de las características otorgadas por la máquina de vectores de soporte (SVM) (a), bosque aleatorio (b), AdaBoost (c), y los valores promedio de las puntuaciones de esos tres modelos (d).

De acuerdo con los puntajes promedio que se muestran en la Figura 8, competition_n, pre_c_GPA, competition_t y c_GPA son los cuatro primeros en cuanto a importancia de características. Los resultados muestran el comportamiento competitivo del alumno y GPA juegan los papeles más importantes en el rendimiento de predicción del modelo, mientras que el académico competencia académica. En este modelo, se implementaron varios algoritmos (SVM, RF y AdaBoost) con un rendimiento de predicción preciso como alumnos base y un algoritmo relativamente simple (regresión logística) se adoptó como el Meta-aprendiz para reducir el riesgo de sobreajuste, y utilizamos Validación cruzada de 10 × 10 veces para evitar la fuga de etiquetas.

La predicción del rendimiento académico de los estudiantes ha sido durante mucho tiempo un tema de investigación importante. En esto estudio, proponemos un modelo de conjunto que utiliza el apilamiento de 2 capas para predecir el rendimiento de los estudiantes en el fondo es menos importante.

6. Discusión y Conclusión

La predicción del rendimiento académico de los estudiantes ha sido durante mucho tiempo un tema de investigación importante. En esto estudio, proponemos un modelo de conjunto que utiliza el apilamiento de 2 capas para predecir el rendimiento de los estudiantes en este modelo, varios algoritmos (SVM, RF y AdaBoost) con precisión.

La investigación empírica basada en los datos recopilados de la Universidad de Hankou muestra que la competencia académica. En El modelo de conjunto propuesto tiene un mejor rendimiento de predicción en comparación con otros modelos. A través de la predicción, el rendimiento se implementó como estudiantes base y un algoritmo relativamente simple el análisis de importancia de las características, encontramos que el comportamiento competitivo y el GPA de los candidatos juegan el papel más importante en la predicción de los resultados de la competencia, y la formación académica no es tan importante como se esperaba. En el pasado, los tutores y gerentes tendían a prestar más atención a los antecedentes académicos y el GPA de los estudiantes, pero ignoraban su comportamiento al participar en competiciones Hasta cierto punto, el comportamiento de los estudiantes en una competencia puede reflejar la comprensión cognitiva de los estudiantes sobre la competencia y las características psicológicas de la personalidad. Por ejemplo, el total

(regresión logística) se adoptó como metaaprendizaje para reducir el riesgo de sobreajuste, y utilizamos una validación cruzada de 10 × 10 veces para evitar la fuga de etiquetas.

La investigación empírica basada en los datos recopilados de la Universidad de Hankou muestra que el modelo de conjunto propuesto tiene un mejor rendimiento de predicción en comparación con otros modelos. A través del análisis de importancia de las características, encontramos que el comportamiento de la competencia y el GPA de los candidatos juegan el papel más importante en la predicción de los resultados de la competencia, y la formación académica no es tan importante como se esperaba. En el pasado, los tutores y gerentes tendían a prestar más atención a los antecedentes académicos y el GPA de los estudiantes, pero ignoraban su comportamiento al participar en competencias. Hasta cierto punto, el comportamiento de los estudiantes en una competencia puede reflejar la comprensión cognitiva de los estudiantes sobre la competencia y las características psicológicas de la personalidad. Por ejemplo, el número total de participantes en competencias académicas puede reflejar la constancia de propósito de un estudiante en la competencia. Es particularmente importante prestar más atención al comportamiento de competencia del candidato para la selección de candidatos de competencia en el futuro.

En 2019, utilizamos este modelo para ayudar en la selección de estudiantes para la Copa Blue Bridge (Concurso Nacional de Diseño de Talento Profesional de Software y Emprendimiento patrocinado por el Centro de Intercambio de Talento del Ministerio de Industria y Tecnología de la Información) en la Universidad de Hankou. La Blue Bridge Cup es una competición de programación que se celebra todos los años. El desempeño de los participantes se clasifica de mayor a menor. El 10% superior es el primer premio, el 20% es el segundo premio y el 30% es el tercer premio. Alrededor de 30.000 estudiantes universitarios de todo el país participaron en la competencia en 2019. Es difícil ganar el premio y en años anteriores, la cantidad de participantes y premios de la Universidad de Hankou no fueron positivos. Importamos los datos de los estudiantes de tres grados (especialidad en computación) al modelo y aplicamos los resultados de predicción y análisis de características a la selección de estudiantes potenciales. Finalmente, 32 estudiantes participaron en la competencia y 11 de ellos ganaron el premio en 2019. Entre ellos, un estudiante recibió el primer premio, dos estudiantes recibieron el segundo premio y los demás ganaron el tercer premio en 2019. El número de ganadores y la proporción de ganadores aumentaron significativamente en comparación con los resultados de la competencia en años anteriores.

Se debe prestar más atención a algunos puntos de investigación en el estudio futuro. En primer lugar, el modelo debe aplicarse a otras universidades en el futuro para examinar más a fondo la estabilidad y la generalización del modelo. En segundo lugar, en la actualidad, es un modelo de clasificación binaria. Con el aumento del volumen de datos, podemos investigar múltiples modelos de clasificación. Se pueden clasificar más categorías de resultados de la competencia en el futuro para obtener conclusiones más significativas. Finalmente, el documento utiliza las medidas tradicionales, como Precisión, Recuperación, F1, Error y AUC para evaluar el rendimiento del modelo. Estas son medidas comúnmente utilizadas para problemas de clasificación en el aprendizaje automático. En trabajos futuros, es necesario encontrar nuevas medidas adecuadas para escenarios educativos específicos.

Contribuciones de los autores: Conceptualización, YL; metodología, LY y YL; recursos, LY; software, LY; supervisión, YL; validación, LY; redacción: borrador original, LY y YL Todos los autores han leído y aceptado la versión publicada del manuscrito.

Agradecimientos: Los autores agradecen al Centro de Investigación de Informatización Educativa de Hubei, Universidad Normal de China Central por brindar apoyo financiero y buenas instalaciones. Además, agradecen a la Universidad de Hankou por proporcionar los datos.

Conflictos de interés: Los autores declaran no tener ningún conflicto de interés.

Apéndice A

Apéndice A.1 Máquina de vectores de soporte (SVM)

El clasificador de máquina de vectores de soporte es un algoritmo clasificador binario que busca un hiperplano óptimo como una función de decisión en un espacio de alta dimensión [35–37]. La máquina de vectores de soporte (SVM) es un modelo de clasificación de dos clases. Su modelo básico se define como el clasificador lineal más espaciado en el espacio de características, es decir, la estrategia de aprendizaje de la máquina de vectores de soporte es maximizar el intervalo y, finalmente, puede transformarse en una optimización cuadrática de una función convexa.

SVM es el algoritmo de aprendizaje automático más común que puede usar técnicas kernel [38], y SVM tiene un buen rendimiento de generalización en pequeños conjuntos de entrenamiento de muestra, pero si la cantidad de datos es grande, el tiempo de entrenamiento de SVM será mayor. El rendimiento de la máquina de vectores de soporte depende principalmente de la selección de la función del kernel [39], por lo que para un problema práctico, la cuestión de cómo seleccionar la función del kernel adecuada de acuerdo con el modelo de datos real para construir el algoritmo SVM es crítica. En la actualidad, muchos parámetros de funciones del kernel dependen de la selección manual, con un cierto grado de arbitrariedad. En diferentes áreas problemáticas, las funciones del kernel deberían tener diferentes formas y parámetros [40]. Las funciones comunes del núcleo en SVM se enumeran como las siguientes ecuaciones:

$$\text{Polinomio} : k(X1, X2) = X \quad {}^{T}_{1}X2 \quad (A1)$$

$$\text{RBF} : k(X1, X2) = \exp \ddot{y} \, 2\ddot{y}^{2} \quad \frac{||X1 \ddot{y} X2||2}{!} \quad (A2)$$

$$\text{Laplaciano} : k(X1, X2) = \exp \ddot{y} \, 2\ddot{y}^{2} \quad \frac{||X1 \ddot{y} X2||}{} \quad (A3)$$

$$\text{Sigmoide} : k(X1, X2) = \tanh a \, X \quad {}^{T}_{1} X2 \ddot{y} \text{ segundo } \mathbf{yo}, \text{ un, segundo} > 0 \quad (A4)$$

Apéndice A.2 Bosque aleatorio

RF [41] se refiere a un método de aprendizaje conjunto para entrenar, clasificar y predecir datos de muestra mediante el uso de árboles de decisión múltiples cuyas salidas se agregan por votación mayoritaria. RFs [27] no necesita asumir la distribución de datos, puede manejar miles de variables de entrada sin eliminación de variables. Al mismo tiempo, da estimaciones de qué variables son importantes en la clasificación.

El algoritmo común para construir RF se describe a continuación: Paso 1.

Seleccionar aleatoriamente K características entre el total de m características, donde K m, luego aleatoriamente seleccione J muestras entre el total de n muestras;

Paso 2. Con las características K sobre las muestras J, calcular el nodo d utilizando el mejor punto de división;

Paso 3. Dividir el nodo en nodos secundarios utilizando la mejor división.

Paso 4. Repita los pasos anteriores del 1 al 3 hasta alcanzar el número de nodos.

Paso 5. Construya el bosque repitiendo los pasos 1 a 4 por q veces para que se creen q árboles.

Apéndice A.3 AdaBoost

AdaBoost se utiliza mejor para aumentar el rendimiento de los árboles de decisión en problemas de clasificación binaria. AdaBoost es sensible a datos con ruido y valores atípicos. De lo contrario, es menos susceptible al problema de sobreajuste que la mayoría de los algoritmos de aprendizaje. El flujo de trabajo del algoritmo AdaBoost para resolver el problema de clasificación binaria se puede describir de la siguiente manera: Paso 1. Inicializar la distribución del peso de los datos de entrenamiento;

$$D(1) = (\ddot{y}11, \ddot{y}12, \cdots, \ddot{y}1m); \ddot{y}1i = \frac{1}{\text{metro}}; \text{ yo} = 1, 2, \cdots, \text{metro} \quad (A5)$$

Paso 2. Obtenga el clasificador Gk (x) por tren en un conjunto de datos con una distribución de peso Dk ; Paso 3. Calcular la tasa de error de clasificación de Gk (x)

$$ek = P \, (Gk \, (xi) \, , yi = Xm \underset{yo=1}{} \ddot{y}kil \, (Gk \, (xi) \, , yi) \quad (A6)$$

Paso 4. Calcular los coeficientes de Gk (x)

$$ak = \frac{1}{2} \log \frac{1 - ek}{eek} \quad (A7)$$

Paso 5. Actualice la distribución del peso del conjunto de datos

$$ÿk+1,i = \frac{ÿki}{Zk} \exp(ÿakyiGk(xi)); yo = 1, 2, \cdots , \text{metro} \quad (A8)$$

donde Zk es un factor de normalización definido como sigue:

$$Zk = \sum_{yo=1}^{m} ÿki \exp(ÿakyiGk(xi)); yo = 1, 2, \cdots , \text{metro} \quad (A9)$$

Paso 6. Para k = 1, 2, . . . , K, repita los pasos 2 a 5 hasta que se entrenen K clasificadores débiles;

Paso 7. Salida del clasificador final.

$$f(x) = \text{signo}\left(\sum_{k=1}^{A} akGk(x)\right). \quad (A10)$$

Apéndice A.4 Regresión logística

Schumacher [32] señaló que la regresión logística es una buena opción para la predicción del éxito en un curso o programa. Es el modelo básico de predicción de una variable aleatoria dependiente dicotómica. La regresión logística describe la relación entre una variable dependiente dicotómica y un conjunto de variables predictoras. Las variables predictoras pueden ser numéricas o categóricas (variables ficticias). Este modelo se utiliza para la predicción de la probabilidad de ocurrencia de un evento ajustando los datos a una curva logística.

El modelo de regresión logística se puede expresar como:

$$\text{logit}(y) = c0 + c1x1 + c2x2 + \cdots + ckxk \quad (A11)$$

donde (x1, x2, · · · , xk) son variables independientes, y es la variable dependiente. (c1, c2, · · · , ck) son coeficientes que se ajustan y usando la técnica de máxima verosimilitud y logit(y) = ln $\frac{y}{1-y}$.

Con un límite numérico determinado (el valor predeterminado suele ser 0,5), los casos con probabilidades superiores a este valor se clasifican como 1 (éxito), mientras que los casos inferiores a este valor se clasifican como 0 (fracaso). Sin embargo, la precisión general de la regresión logística no es alta y es fácil que no se ajuste bien, de modo que no puede manejar muy bien las características no lineales [31].

Referencias

1. Van Nul, WTD; Roach, VA Cara a cara: El papel de la competencia en la educación universitaria. Anat. ciencia Educ. **2015,** 8, 404–412. [Referencia cruzada]
2. Campbell, HWJR; Walberg, HJ La teoría de un sistema cuántico general interactuando con un sistema disipativo lineal. Ana. física **2000,** 547–607. [Referencia cruzada]
3. Campbell, JR; Walberg, HJ Estudios de Olimpiadas: Las competencias brindan alternativas para desarrollar talentos que sirvan a los intereses nacionales. Roeper Rev. **2010,** 33, 8–17. [Referencia cruzada]
4. Goldstein, D.; Wagner, H. Programas extracurriculares, competencias, olimpiadas escolares y programas de verano. En t. Manob. Res. desarrollo Regalo. Talento **1993,** 33, 593–604.
5. Urhahne, D.; Ho, LH; Parchmann, I.; Nick, S. Intentando predecir el éxito en la ronda de clasificación de la Olimpiada internacional de química. Alta habilidad Semental. **2012,** 23, 167–182. [Referencia cruzada]
6. Sandeep, MJ Alerta temprana de estudiantes en riesgo académico: una iniciativa de análisis de código abierto. J. Aprende. Anal. **2014,** 1, 6–47.
7. Bouzayane, S.; Saad, I. Predicción semanal de los estudiantes de mooc en riesgo utilizando un conjunto aproximado basado en dominancia Acercarse. lect. Cómputo de notas. ciencia **2017,** 10254, 160–169.
8. Botelorenzo, ML; Gomezsanchez, E. Predicción de la disminución de los indicadores de compromiso en un mooc. En Actas de la Séptima Conferencia Internacional sobre Análisis y Conocimiento del Aprendizaje en LAK, Vancouver, BC, Canadá, 13–17 de marzo de 2017; págs. 143–147.

9. Kennedy, G.; Ataúd, C.; De Barba, P.; Corrin, L. Prediciendo el éxito: cómo los conocimientos previos, las habilidades y las actividades de los alumnos predicen el rendimiento de mooc. En Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, Poughkeepsie, NY, EE. UU., 16 al 20 de marzo de 2015; págs. 136–140.
10. Mann, CM; Canny, BJ; Lindley, JM; Rajan, R. La influencia de la familia lingüística en el rendimiento académico en alumnos de 1 y 2 mbbs de año. Medicina. Educ. **2010,** 44, 786–794. [Referencia cruzada]
11. Johns, MW; Dudley, HAF; Masterton, JP Los hábitos de sueño, la personalidad y el rendimiento académico de estudiantes de medicina. Con. Educ. **1976,** 10, 158–162. [Referencia cruzada]
12. Carretero, SP; Greenberg, K.; Walker, MS El impacto del uso de la computadora en el rendimiento académico: evidencia de un ensayo aleatorio en la academia militar de los Estados Unidos. economía Educ. Rev. **2017,** 56, 118–132. [Referencia cruzada]
13. Vale, MW; Kim, W. Uso de ipads e ipods para el rendimiento académico y la participación de estudiantes de prek12 con discapacidades: una síntesis de investigación. Excepcionalidad **2017,** 25, 54–75. [Referencia cruzada]
14. Huang, S.; Fang, N. Predicción del rendimiento académico de los estudiantes en un curso de ingeniería dinámica: una comparación de cuatro tipos de modelos matemáticos predictivos. computar Educ. **2013,** 61, 133–145. [Referencia cruzada]
15. Al-Ghamdi, SA; Al-Bassiouni, AAM; Mustafá, HMH; Al-Hamadi, A. Simulación de rendimiento académico mejorado para un tema matemático utilizando el modelado de redes neuronales. Cómputo mundial. ciencia información Tecnología J. **2013,** 3, 77–84.
16. Kotsiantis, SB; Pierrakeas, C.; Pintelas, PE Predicción del rendimiento de los estudiantes en el aprendizaje a distancia utilizando técnicas de aprendizaje automático. aplicación Artefacto Intel. **2004,** 18, 411–426. [Referencia cruzada]
17. Romero, C.; Espejo, PG; Zafra, A.; Romero, JR; Ventura, S. Minería de uso web para predecir calificaciones finales de estudiantes que usan cursos de Moodle. computar aplicación Ing. Educ. **2013,** 21, 135–146. [Referencia cruzada]
18. Parij, D.; Polikar, R. Un enfoque de aprendizaje incremental basado en conjuntos para la fusión de datos. sist. Hombre cibernético. **2007,** 37, 437–450. [Referencia cruzada]
19. Beemer, J.; Cuchara, KM; El, L.; Ventilador, J.; Levine, RA Aprendizaje conjunto para estimar los efectos del tratamiento individualizado en los estudios de éxito estudiantil. Artefacto Intel. Educ. **2018,** 28, 315–335. [Referencia cruzada]
20. Adé, R.; Deshmukh, PR Un conjunto incremental de clasificadores como técnica para predecir la elección de carrera de los estudiantes. En Actas de la Primera Conferencia Internacional sobre Redes y Soft Computing de 2014 (ICNSC2014), Guntur, India, 19 y 20 de agosto de 2014; págs. 384–387.
21. Kotsiantis, SB; Patriarcas, K.; Xenos, MN Un conjunto incremental combinacional de clasificadores como técnica para predecir el desempeño de los estudiantes en la educación a distancia. Saber Sistema basado **2010,** 23, 529–535. [Referencia cruzada]
22. Kearns, MJ; Li, M.; Fórmulas booleanas de Valiant, LG Learning. J. ACM. **1994,** 41, 1298–1328. [Referencia cruzada]
23. Schalk, PD; mecha, DP; Turner, PR; Ramsdell, MW Evaluación predictiva del desempeño de los estudiantes para orientación estratégica temprana. En Proceedings of the 2011 Frontiers in Education Conference (FIE), Rapid City, SD, EE. UU., 12 al 15 de octubre de 2011.
24. Hardman, J.; Paucar-Cáceres, A.; Fielding, A. Predicción del progreso de los estudiantes en la educación superior mediante el uso del algoritmo de bosque aleatorio. sist. Res. Comportamiento ciencia **2013,** 30, 194–203. [Referencia cruzada]
25. Shamsi, MS; Lakshmi, J. Predicción del rendimiento de los estudiantes utilizando técnicas de minería de datos de clasificación. arxiv **2016,** arXiv:1606.05735.
26. Ishizue, R.; Sakamoto, K.; Washizaki, H.; Fukazawa, Y. Predictores de clasificación de habilidades y colocación de estudiantes para clases de programación utilizando actitud de clase, escalas psicológicas y métricas de código. Res. Practica Tecnología mejorar Aprender. **2018,** 13, 7. [Referencia cruzada]
27. Petkovic, D.; Sosnickpérez, M.; Okada, K.; Todtenhoefer, R.; Huang, S.; Miglani, N.; Vigil, A. Uso del clasificador de bosque aleatorio para evaluar y predecir el aprendizaje de los estudiantes sobre el trabajo en equipo de ingeniería de software. En Actas de la Conferencia IEEE Frontiers in Education (FIE) de 2016, Eire, PA, EE. UU., 12 al 15 de octubre de 2016; págs. 1 a 7.
28. Noori, R.; Karbassi, A.; Moghaddamnia, A.; Mano .; Zokaeiashtiani, MH; Farokhnia, A.; Gousheh, MG Evaluación de la determinación de las variables de entrada en el rendimiento del modelo svm utilizando pca, prueba gamma y técnicas de selección directa para la predicción mensual del flujo de la corriente. J. Hydrol. **2011,** 401, 177–189. [Referencia cruzada]
29. Han, M.; Tong, M.; Chen, M.; Liu, J.; Liu, C. Aplicación del algoritmo de conjunto en la predicción del rendimiento de los estudiantes. En Actas del 6º Congreso Internacional IIAI sobre Informática Aplicada Avanzada (IIAI-AAI) de 2017, Hamamatsu, Japón, 9 al 13 de julio de 2017; págs. 735–740.

30. Poh, N.; Smythe, I. ¿Hasta qué punto podemos predecir el desempeño de los estudiantes? Un estudio de caso en colegios de sudáfrica. En Actas del Simposio IEEE sobre Inteligencia Computacional y Minería de Datos (CIDM) de 2014, Orlando, FL, EE. UU., 9 al 12 de diciembre de 2014; págs. 416–421.
31. Allison, PD Regresión logística utilizando el sistema SAS: aplicación de la teoría. J. Chem. información Modelado **2019,** 53, 1689–1699.
32. Schumacher, M.; Rosner, R.; Vach, W. Redes neuronales y regresión logística. computar Estadística Análisis de datos. **1996,** 21, 661–682. [Referencia cruzada]
33. Kohavi, R. Un estudio de validación cruzada y arranque para la estimación de precisión y selección de modelos. Ijcai **1995,** 14, 1137–1145.
34. Fawcett, T. Una introducción al análisis de roc. Reconocimiento de patrones. Letón. **2006,** 27, 861–874. [Referencia cruzada]
35. Boser, BE; Guyon, I.; Vapnik, V. Un algoritmo de entrenamiento para clasificadores de margen óptimo. Actas del taller anual de Acm sobre teoría del aprendizaje computacional. 2008, págs. 144–152. Disponible en línea: http: //www.gautampendse.com/projects/bsvm/webpage/boser1992.pdf (consultado el 29 de marzo de 2020).
36. Vapnik, V. Teoría del aprendizaje estadístico; Willy: Nueva York, NY, EE. UU., 1998; pags. 16. Disponible en línea: http: //read.pudn.com/downloads161/ebook/733192/Statistical-Learning-Theory.pdf (consultado el 29 de marzo de 2020).
37. Cristianini, N.; Shawetaylor, J. Introducción a las máquinas de vectores de soporte y otros métodos de aprendizaje basados en el kernel; Prensa de la Universidad de Cambridge: Cambridge, Reino Unido, 2000; Disponible en línea: https://books.google.com.hk/ books?hl=en&lr=&id=_PXJn_cxv0AC&oi=fnd&pg=PR9&dq=37.%09Cristianini,+N.%3B+Shawetaylor, +J. +An+Introduction+to+Support+Vector+Machines+and+Other+Kernel-Based +Métodos+de+aprendizaje+ +. +Universidad+de+Cambridge+Prensa&ots=xSUk6D-r09&sig=cO32--yeujiGuwGA8wHfqWbnAOU&redir_ esc=y&hl=zh-CN&sourceid=cndr#v=onepage&q=37.%09Cristianini%2C%20N.%3B%20Shawetaylor% 2C%20J. %20An%20Introducción%20a%20Soporte%20Vector%20Máquinas%20y%20Otro%20Kernel Basado%20Aprendizaje%20Métodos%20%20.%20Cambridge%20Universidad%20Press&f=false (consultado el 29 de marzo de 2020).
38. Burges, CJC Un tutorial sobre máquinas de vectores de soporte para el reconocimiento de patrones. Datos mín. Saber Descubrir **1998,** 2, 121–167. [Referencia cruzada]
39. Aluko, RO; Daniel, EI; Oshodi, OS; Aigbavboa, C.; Abisuga, AO Hacia una predicción fiable del rendimiento académico de los estudiantes de arquitectura utilizando técnicas de minería de datos. J. Ing. Des. Tecnología **2018,** 16, 385–397. [Referencia cruzada]
40. Frohlich, H.; Chapelle, O.; Scholkopf, B. Selección de características para máquinas de vectores de soporte mediante algoritmo genético. En Actas de la 15.ª Conferencia Internacional IEEE sobre Herramientas con Inteligencia Artificial, Sacramento, CA, EE. UU., 3–5 de noviembre de 2003; págs. 142–148.
41. Breiman, L. Bosques aleatorios. Mach. Aprender. **2001,** 45, 5–32. [Referencia cruzada]