Análisis Predictivo del Rendimiento Académico Estudiantil

Identificación de Factores Clave y Modelos Predictivos para el Éxito Estudiantil

Equipo de Investigación en Ciencia de Datos

Invalid Date

1. Análisis Exploratorio de Datos (EDA) 1.1 Descripción del Conjunto de Datos El análisis utiliza el conjunto de datos de Rendimiento Estudiantil que contiene 395 estudiantes con 33 atributos que describen demografía, contexto social y registros académicos. La variable objetivo se derivó de la nota final (G3):

Aprobado: Nota final 10 (265 estudiantes, 67.1%)

Reprobado: Nota final $<$ 10 (130 estudiantes, 32.9%)

1.2 Análisis Univariado Distribución de Variables Numéricas Clave:

Table 1: Estadísticas Descriptivas de Variables Numéricas Clave

Variable	Media	Mediana	Desviacion_Estandar	Minimo	Maximo	Asimetria
Edad	16.71	17.0	1.29	15	22	0.46
Ausencias	5.53	4.0	7.75	0	75	3.64
Fallas Previas	0.32	0.0	0.73	0	3	2.37
Educación de la Madre	2.70	3.0	1.10	0	4	-0.32
Educación del Padre	2.55	2.5	1.07	1	4	-0.03
Tiempo de Estudio	2.04	2.0	0.82	1	4	0.63
Salidas Sociales	3.08	3.0	1.12	1	5	0.12
Alcohol Diario	1.50	1.0	0.94	1	5	2.17

Hallazgos Clave del Análisis Univariado: Distribución de Edad: La mayoría de estudiantes (moda = 17) están en sus últimos años de adolescencia

Ausentismo: Alta variabilidad con algunos estudiantes faltando hasta 75 clases

Fallas Previas: 68% de estudiantes no tienen fallas previas

Educación Parental: La educación de la madre es ligeramente mayor que la del padre en promedio

1.3 Análisis Bivariado con la Variable Objetivo Variables Numéricas vs. Estado Académico:

Table 2: Valores Promedio por Estado Académico

Variable	Media_Aprobados	Media_Reprobados	Diferencia	Direccion
Fallas Previas	0.12	0.73	0.61	Reprobados tienen más
Ausencias	4.42	7.65	3.23	Reprobados tienen más
Educación de la Madre	2.85	2.38	0.47	Aprobados tienen más educaci
Edad	16.58	17.02	0.44	Reprobados son mayores
Salidas Sociales	2.97	3.34	0.37	Reprobados salen más

Variables Categóricas vs. Estado (Pruebas Chi-Cuadrado):

Table 3: Pruebas Chi-Cuadrado para Variables Categóricas

Variable	Valor_P	Significancia	Interpretacion
higher (desea educación superior)	0.02385	**	Fuertemente relacionado con éxito académico
guardian (tutor)	0.03277	**	Tipo de tutor influye en resultados
paid (clases pagadas extra)	0.05645	*	Clases extra pueden ayudar marginalmente
romantic (en relación)	0.09125	.	Relaciones románticas tienen efecto marginal
Mjob (trabajo madre)	0.09994	.	Trabajo de la madre tiene alguna influencia
internet (acceso a internet)	0.10624	NS	Acceso a internet no es estadísticamente significat

**: $p < 0.05$, *: $p < 0.10$, .: $p < 0.15$, NS: No Significativo

2. Caso 1: Todas las Variables sin Balanceo 2.1 Configuración del Caso 1 Variables: Todas las 30 variables disponibles

Balanceo: Sin balanceo de clases (distribución natural 67.1% Aprobado / 32.9% Reprobado)

Objetivo: Establecer línea base de rendimiento con todas las características

2.2 Resultados del Caso 1

Table 4: Resultados del Caso 1: Todas las Variables sin Balanceo

Modelo	AUC	Precision	Sensibilidad	Especificidad	Posicion
Gradient Boosting	**0.6425**	**63.27%**	**25.00%**	**81.82%**	**1**
Regresión Logística	0.6335	61.22%	31.25%	75.76%	2
Random Forest	0.6127	71.43%	40.62%	86.36%	3
Árbol de Decisión	0.6125	64.29%	37.50%	77.27%	4

SVM Radial	0.5909	70.41%	31.25%	89.39%	5
KNN	0.5386	63.27%	25.00%	81.82%	6
Naive Bayes	0.5421	59.18%	31.25%	72.73%	7

2.3 Análisis del Caso 1 Fortalezas Identificadas: Gradient Boosting obtuvo el mejor AUC (0.6425)

Random Forest logró la mayor precisión general (71.43%)

SVM Radial alcanzó la mayor especificidad (89.39%)

Debilidades Identificadas: Baja sensibilidad general: Todos los modelos detectan menos del 50% de estudiantes reprobados

Sobreajuste potencial: 30 variables pueden introducir ruido en algunos modelos

KNN y Naive Bayes mostraron rendimiento inferior

Conclusiones del Caso 1: Los modelos complejos (Gradient Boosting, Random Forest) funcionan mejor con todas las variables

Existe un claro trade-off entre sensibilidad y especificidad

La baja sensibilidad sugiere necesidad de abordar el desbalance de clases

3. Caso 2: Todas las Variables con Balanceo 3.1 Configuración del Caso 2 Variables: Todas las 30 variables disponibles

Balanceo: Con sobremuestreo (up-sampling) para balancear clases

Objetivo: Evaluar impacto del balanceo manteniendo todas las características

3.2 Resultados del Caso 2

Table 5: Resultados del Caso 2: Todas las Variables con Balanceo

Modelo	AUC	Precision	Sensibilidad	Especificidad	Cambio_Sensibilidad
Regresión Logística	**0.6316**	**61.22%**	**50.00%**	**66.67%**	**+60.0%**
Gradient Boosting	0.6259	63.27%	40.62%	74.24%	+62.5%
Random Forest	0.6120	66.33%	40.62%	78.79%	0.0%
SVM Radial	0.6009	59.18%	37.50%	69.70%	+20.0%
Árbol de Decisión	0.5881	68.37%	31.25%	86.36%	-16.7%
KNN	0.5573	57.14%	53.12%	59.09%	+112.5%

3.3 Impacto del Balanceo en el Caso 2 Mejoras Observadas: Sensibilidad incrementada significativamente en la mayoría de modelos

Regresión Logística mejoró su sensibilidad de 31.25% a 50.00%

KNN duplicó su sensibilidad (de 25.00% a 53.12%)

Compensaciones Observadas: Reducción en AUC en la mayoría de modelos

Disminución en especificidad como trade-off por mayor sensibilidad

Precisión general se mantuvo similar o disminuyó ligeramente

Conclusiones del Caso 2: El balanceo mejora significativamente la detección de estudiantes reprobados

Los modelos más simples (Regresión Logística, KNN) se benefician más del balanceo

Existe un trade-off claro entre sensibilidad y especificidad

4. Proceso de Selección de Variables 4.1 Metodología de Selección Se emplearon tres métodos complementarios para identificar las variables más predictivas:

4.1.1 Pruebas Chi-Cuadrado (Variables Categóricas)

Table 6: Ranking de Variables Categóricas por Chi-Cuadrado

Variable	Valor_P	Significancia	Decision
higher	0.02385	Muy alta	Seleccionada
guardian	0.03277	Alta	Seleccionada
paid	0.05645	Moderada	Considerada
romantic	0.09125	Marginal	Considerada
Mjob	0.09994	Marginal	Considerada
internet	0.10624	Baja	No seleccionada

4.1.2 Algoritmo Boruta (Importancia de Todas las Variables)

Table 7: Variables Confirmadas Importantes por Boruta

Variable	Importancia_Media	Decision	Categoria
failures	22.90	Confirmada	Académica
absences	11.23	Confirmada	Asistencia
goout	8.09	Confirmada	Conductual
schoolsup	7.04	Confirmada	Apoyo
guardian	6.34	Confirmada	Familiar

Medu	4.22	Confirmada	Familiar
age	4.01	Confirmada	Demográfica

4.1.3 Eliminación Recursiva de Características (RFE) Método: Random Forest como wrapper

Variables óptimas identificadas: 9 variables

Top 5 variables: failures, absences, goout, Medu, schoolsup

Precisión máxima: 72.91% con 9 variables

4.2 Conjunto Final de Variables Seleccionadas Basado en el consenso de los tres métodos, se seleccionaron 7 variables clave:

Table 8: Conjunto Final de 7 Variables Seleccionadas

Variable	Descripcion	Metodos_Que_La_Seleccionan	Justificacion
failures	Número de fallas académicas previas	3/3	Predictor más fuerte en t
absences	Número total de ausencias	3/3	Alta correlación con rend
higher	Intención de cursar educación superior	2/3	Fuertemente relacionada
age	Edad del estudiante	2/3	Indicador potencial de re
Medu	Nivel educativo de la madre	2/3	Factor socioeconómico im
goout	Frecuencia de salidas sociales	2/3	Indicador de balance vida
guardian	Persona responsable del estudiante	2/3	Indicador de estructura d

4.3 Beneficios de la Selección de Variables Reducción de dimensionalidad: 30 → 7 variables (76.7% reducción)

Mejor interpretabilidad: Variables más comprensibles para stakeholders

Reducción de sobreajuste: Menor riesgo de modelar ruido

Eficiencia computacional: Entrenamiento y predicción más rápidos

5. Caso 3: Variables Seleccionadas sin Balanceo 5.1 Configuración del Caso 3 Variables: 7 variables seleccionadas (failures, absences, higher, age, Medu, goout, guardian)

Balanceo: Sin balanceo de clases

Objetivo: Evaluar rendimiento con variables optimizadas manteniendo distribución natural

5.2 Resultados del Caso 3

Table 9: Resultados del Caso 3: Variables Seleccionadas sin Balanceo

Modelo	AUC	Precision	Sensibilidad	Especificidad	Comparacion_Caso1
Regresión Logística	**0.6461**	**69.39%**	**28.12%**	**89.39%**	**+2.0% AUC**
Gradient Boosting	0.6409	68.37%	28.12%	87.88%	-0.2% AUC
Random Forest	0.6274	68.37%	31.25%	86.36%	+2.4% AUC
KNN	0.6054	72.45%	34.38%	90.91%	+12.4% AUC
SVM Radial	0.6016	68.37%	31.25%	86.36%	+1.8% AUC
Árbol de Decisión	0.5881	68.37%	31.25%	86.36%	-4.0% AUC

5.3 Análisis del Caso 3 Mejoras Observadas vs. Caso 1: Regresión Logística mejora su AUC de 0.6335 a 0.6461

KNN muestra la mayor mejora en precisión (de 63.27% a 72.45%)

Especificidad general más alta: Promedio de 87.71% vs. 81.07% en Caso 1

Características Persistentes: Sensibilidad aún baja: Promedio de 30.56%

Trade-off sensibilidad-especificidad sigue presente

Modelos más simples mejoran relativamente más que modelos complejos

Conclusiones del Caso 3: La selección de variables mejora el rendimiento de modelos más simples

Se mantiene alta especificidad pero baja sensibilidad

7 variables son suficientes para lograr rendimiento competitivo

6. Caso 4: Variables Seleccionadas con Balanceo 6.1 Configuración del Caso 4 Variables: 7 variables seleccionadas

Balanceo: Con sobremuestreo (up-sampling)

Objetivo: Combinar beneficios de selección de variables y balanceo de clases

6.2 Resultados del Caso 4

Table 10: Resultados del Caso 4: Variables Seleccionadas con Balanceo

Modelo	AUC	Precision	Sensibilidad	Especificidad	AUC_vs_Caso3	Sensibilidad_vs_Ca
KNN	**0.6536**	**63.27%**	**65.62%**	**62.12%**	**+4.82%**	**+91.2%**
Random Forest	0.6383	65.31%	37.50%	78.79%	+1.09%	+20.0%
SVM Radial	0.6316	69.39%	46.88%	80.30%	+3.00%	+50.0%

Gradient Boosting	0.6283	62.24%	53.12%	66.67%	-1.26%	+88.9%
Regresión Logística	0.6278	61.22%	46.88%	68.18%	-1.83%	+66.7%
Árbol de Decisión	0.5743	61.22%	31.25%	75.76%	-1.38%	0.0%

6.3 Análisis del Caso 4 Mejoras Significativas: KNN logra el mejor AUC global (0.6536) y la mayor sensibilidad (65.62%)

Sensibilidad promedio incrementada de 30.56% (Caso 3) a 48.44%

Balance mejorado entre métricas de rendimiento

Trade-offs Observados: Especificidad disminuye como compensación por mayor sensibilidad

Algunos modelos (GLM, GBM) muestran ligera reducción en AUC

Precisión general se mantiene en niveles similares

Hallazgo Clave: KNN emerge como el mejor modelo en esta configuración, demostrando que con las variables correctas y balanceo adecuado, algoritmos simples pueden superar a modelos más complejos

6.4 Comparativa Entre los 4 Casos

Table 11: Comparativa General de los 4 Casos Experimentale

Caso	Descripcion	Mejor_Modelo	Mejor_AUC	Sensibilidad_Promedio	Especificidad_
Caso 1	30 vars, sin balanceo	Gradient Boosting	0.6425	31.25%	81.07
Caso 2	30 vars, con balanceo	Regresión Logística	0.6316	42.19%	72.53
Caso 3	7 vars, sin balanceo	Regresión Logística	0.6461	30.56%	87.71
Caso 4	**7 vars, con balanceo**	**KNN**	**0.6536**	**48.44%**	**71.97**

7. Conclusiones y Recomendaciones Finales 7.1 Hallazgos Clave del Estudio Selección de Variables es Efectiva: Reducir de 30 a 7 variables mejora la eficiencia sin sacrificar rendimiento

Balanceo Mejora Detección de Riesgo: El sobremuestreo incrementa significativamente la sensibilidad

No Hay un Modelo Único Óptimo: Diferentes configuraciones optimizan diferentes métricas

Variables Críticas Identificadas: Fallas previas, ausencias y aspiración educativa son predictores clave

7.2 Recomendaciones de Implementación Para Prioridad: Detección Temprana de Riesgo Configuración recomendada: Caso 4 con modelo KNN

Ventaja: 65.62% sensibilidad para detectar estudiantes reprobados

Uso: Sistema de alerta temprana

Para Prioridad: Optimización de Recursos Configuración recomendada: Caso 3 con Regresión Logística

Ventaja: 89.39% especificidad para minimizar falsos positivos

Uso: Asignación dirigida de recursos de apoyo

Para Prioridad: Balance General Configuración recomendada: Caso 2 con Random Forest

Ventaja: Buen balance entre todas las métricas

Uso: Monitoreo general institucional

7.3 Recomendación Final Basado en el análisis completo, recomendamos una implementación por fases:

Fase 1 (Semanas 1-6): Implementar Caso 4 con KNN para maximizar detección temprana de estudiantes en riesgo.

Fase 2 (Meses 2-6): Añadir Caso 3 con Regresión Logística como verificación secundaria para optimización de recursos.

Fase 3 (Meses 6-12): Desarrollar un sistema ensemble que combine ambas configuraciones, ajustando automáticamente según prioridades institucionales.

7.4 Limitaciones y Trabajo Futuro Limitaciones Reconocidas: Rendimiento Predictivo Moderado: AUC máximo de 0.6536 indica margen de mejora

Contexto Específico: Datos de sistema educativo portugués

Datos Auto-reportados: Posible sesgo en respuestas

Direcciones Futuras: Integración de Datos Temporales: Seguimiento longitudinal de estudiantes

Validación Cruzada Institucional: Probar en diferentes contextos educativos

Modelos de Explicabilidad: Mejorar interpretabilidad para stakeholders no técnicos

Sistemas de Recomendación: Sugerir intervenciones específicas basadas en perfiles de riesgo

Nota Final: Este estudio demuestra que, aunque la predicción perfecta del rendimiento estudiantil sigue siendo un desafío, es posible identificar factores clave y desarrollar herramientas útiles para la toma de decisiones educativas basadas en datos. La combinación de selección inteligente de variables y técnicas apropiadas de balanceo puede proporcionar insights valiosos para mejorar los resultados estudiantiles.