INFORME FINAL TRABAJO ARENDIZAJE SUPERVISADO

YULIANNY ÁLVAREZ VILLAMIZAR
ESTEBAN CARO PELAÉZ
JORGE LUIS RODRÍGUEZ LÓPEZ
JUAN PABLO RAMÍREZ RINCÓN

UNIVERSIDAD DE ANTIOQUIA
1 8 0 3

Introducción

En el siguiente informe se muestra la implementación de tres algoritmos de aprendizaje supervisado para la predicción de la deserción estudiantil en la Facultad de Ingeniería. La utilización de métodos predictivos en esta problemática es de gran interés dado que se puede identificar estudiantes con alta probabilidad de deserción para tomar medidas pertinentes desde los programas de bienestar universitario y así evitar los altos costos sociales que acarrea esta problemática mientras se protege la salud mental del estudiantado.

1. DESCRIPCIÓN DE LOS DATASETS

Se tienen dos datasets, cada uno contiene una muestra de estudiantes matriculados a las materias álgebra y cálculo diferencial. En dichos datasets se tienen algunos atributos importantes de los estudiantes, tales como; programa al cual está inscrito, características sociodemográficas, particularidades familiares y desempeño académico del semestre. Estos atributos se ven reflejados en las 33 columnas que tiene cada dataset. Por otro lado, para el dataset de álgebra se tienen 651 registros mientras que para cálculo diferencial se tienen 397 registros. En ninguno de los dos dataset se tienen valores nulos y se identifican dos duplicados dentro de cada tabla, los cuales deciden dejarse ya que existe la probabilidad de que dos estudiantes cuenten con las mismas características sociodemográficas, académicas y familiares.

2. ANÁLISIS EXPLORATORIO

A continuación, se muestra un análisis exploratorio de las variables que se consideran más representativas e importantes.

2.1 Variable objetivo

Para este problema predictivo, la variable a predecir es la nota obtenida por los estudiantes en el último examen de cada una de las materias. A continuación, observamos la distribución de dicha variable para ambos cursos.

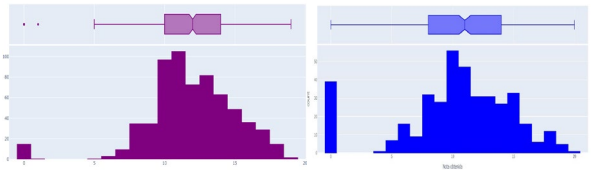

Figura 1. Variables objetivo álgebra

Figura 2. Variable objetico cálculo

En la figura 1 y 2 es posible identificar que en promedio los estudiantes tienen un mejor desempeño en álgebra que en cálculo. Además de esto, la proporción de estudiantes que sacan cero en el tercer examen es mucho mayor en cálculo que en álgebra.

2.2 Variables numéricas importantes

En la figura 3 se observa la distribución de las edades de los estudiantes para el curso de álgebra. Aunque la distribución de la misma variable para cálculo presenta leves diferencias, el comportamiento es muy similar; la mayoría de los estudiantes tienen entre 15 y 18 años.

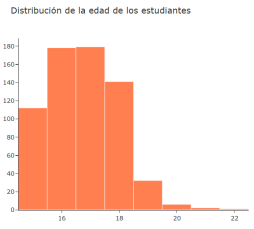

Figura 3. Edades

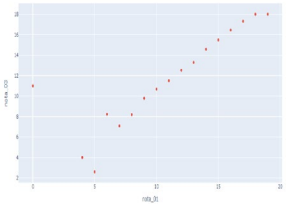

Figura 4. Correlación nota 2 – nota 3

En la figura 4 se observa que existe una alta correlación entre la nota obtenida tanto en el examen 1 y como en el examen 2 con la nota que se obtiene en el tercer parcial. Esto tiene mucho sentido dado a la naturaleza acumulativa del contenido de ambos cursos. El comportamiento de estas variables es muy similar en ambas materias.

2.2 Variables numéricas importantes

En las bases de datos la mayor cantidad de estudiantes están matriculados en el programa de Ingeniería Industrial como se observa en la figura 5.

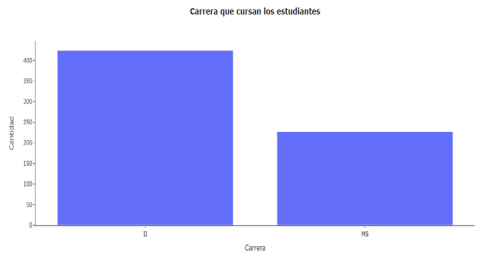

Figura 5. Distribución de matriculados por programa

3. MODELO BASE – REGRESIÓN LINEAL

3.1 Dataset Álgebra

A partir del *método de filtrado* de selección de variables, reducimos el dataset a las variables que se consideraron que tenían importancia dado a que eran significativas con respecto a la variable objetivo, el tercer parcial. Dichas variables se muestran a continuación:

```
array(['departamento', 'sexo', 'direccion', 'madre_edu', 'padre_edu',
       'madre_trab', 'padre_trab', 'razon', 'guardian', 't_estudio',
       'faltas', 'actividades_extra', 'internet', 'relacion_fam',
       'tiempo_libre', 'cons_alcohol_sem', 'cons_alcohol_finde',
       'nota_01', 'nota_02', 'padres_edu'], dtype=object)
```

Figura 6. Variables seleccionadas álgebra método de filtrado

Implementamos un modelo de regresión lineal múltiple que permite predecir las notas del examen 3 de los estudiantes mediante la librería Se obtienen las siguientes métricas de evaluación para el conjunto de datos de testing:

```
MSE: 2.54    MAE: 0.97      R2: 0.75
```

Si se toma como métrica de evaluación principal el R2, es posible concluir que el 75% de la variabilidad de la nota del tercer parcial de álgebra lineal es explicado por las variables observadas en la figura 6.

3.2 Dataset Cálculo

En el caso de los datos para la materia de cálculo diferencial, la reducción de variables también se hizo mediante el método de filtrado, obteniéndose las siguientes:

```
array(['sexo', 'direccion', 'madre_edu', 'padre_edu', 'madre_trab',
       'padre_trab', 'guardian', 't_estudio', 'faltas', 'internet',
       'relacion_sen', 'relacion_fam', 'tiempo_libre', 'salir_amigos',
       'cons_alcohol_finde', 'nota_01', 'nota_02', 'padres_edu'],
      dtype=object)
```

Figura 7. Variables seleccionadas cálculo método de filtrado

A continuación, se muestran las métricas de desempeño obtenidas mediante este modelo:

```
MSE: 5.39    MAE: 1.56    R2: 0.76
```

Si se toma como métrica de evaluación principal el R2, es posible concluir que el 76% de la variabilidad de la nota del tercer parcial de cálculo es explicado por las variables observadas en la figura 7.

4. SELECCIÓN DE VARIABLES POR OTROS MÉTODOS

Es de vital importancia tener certeza que las variables seleccionadas para explicar la variable objetivo sean las adecuadas, dado que la selección de características es vital en el aprendizaje automático. Ayuda a reducir la dimensionalidad de los datos, minimizar el riesgo de sobreajuste y mejorar el rendimiento general del modelo [1]. En este caso, se decidió implementar dos métodos de selección de variables: wrapper e integrados. Los algoritmos implementados posteriormente se evaluarán con las variables resultantes por ambos métodos.

4.1 Método Wrapper

Para la técnica wrapper utilizamos dos métodos; forward y backward. Se decidió dejar las variables en común arrojadas en ambos resultados.

	MÉTODO WRAPPER	
	Dataset Álgebra	**Dataset Cálculo**
Variables seleccionadas	'madre_trab_en_casa', 'faltas', 'razon_otro', 'razon_reputacion', 'madre_trab_profesor', 'razon_recomendacion', 'padre_trab_otro', 'madre_trab_salud', 'razon_habilidad', 'madre_trab_servicios', 'cons_alcohol_finde', 'edad', 'padre_trab_servicios', 'padre_edu', 'tiempo_libre', 'promedio_acumulado', 'soporte_edu_fam-encoded'	't_examen','razon_recomendacion', 'actividades_extra-encoded', 't_estudio', 'promedio_acumulado', 'cons_alcohol_sem', 'padre_trab_servicios', 'ausencias', 'razon_reputacion', 'soporte_edu_extra-encoded', 'edad', 'guardian_padre', 'relacion_fam', 'padre_trab_otro'

Figura 8. Variables seleccionadas por wrapper

4.2 Métodos integrados

Implementamos la técnica se lección de variables tanto por la regularización Lasso y Ridge con el objetivo de comprar el resultado de ambas características seleccionadas mediante la penalización de coeficientes.

	MÉTODOS INTEGRADOS (Lasso)	
	Dataset Álgebra	**Dataset Cálculo**
Variables seleccionadas	'faltas', 'cons_alcohol_finde', 'salud', 'promedio_acumulado', 'departamento-encoded', 'sexo-encoded', 'direccion-encoded', 'postgrado-encoded', 'madre_trab_otro', 'padre_trab_servicios', 'razon_otro'	'edad', 'madre_edu', 'faltas', 'relacion_fam', 'cons_alcohol_finde', 'ausencias', 'nota_02', 'promedio_acumulado', 'sexo-encoded', 'soporte_edu_extra-encoded', 'soporte_edu_fam-encoded', 'actividades_extra-encoded', 'relacion_sen-encoded', 'padre_trab_servicios', 'razon_recomendacion', 'razon_reputacion', 'guardian_madre'

Figura 9. Variables seleccionadas por integrados

En este caso se eligió dejar el dataset resultante de aplicar el método LASSO, ya que el método RIDGE con una *alpha* muy pequeño, nos arrojaba muy pocas variables significativas, lo que nos dejaba el modelo muy corto en columnas, por lo que decidimos omitir.

5. OTROS ALGORITMOS DE APRENDIZAJE SUPERVISDO

A continuación, se implementarán otros dos algoritmos de aprendizaje para posteriormente evaluar las métricas de evaluación de todos los algoritmos propuestos. Cabe resaltar que los algoritmos que se mostrarán a continuación se alimentan con las variables seleccionadas tanto con los métodos wrapper como integrados, con el objetivo de analizar cuál modelo y cuáles variables ayudan a dar predicciones más confiables.

5.1 Random forest

5.1.1 Dataset Álgebra

Métricas de evaluación – Conjunto de datos testing.

Variables seleccionadas mediante método wrapper

```
MSE:1.9965    MAE:0.8924       R2: 0.7751
```

Variables seleccionadas mediante Lasso

```
MSE:1.8011    MAE:0.8039       R2: 0.8053
```

Según las métricas, y tomando como referencia el R2 podemos observar que hay una mejoría de aproximadamente del 3% al alimentar el modelo con las variables seleccionadas mediante Lasso con respecto a este mismo modelo alimentado con las variables seleccionadas mediante el método wrapper.

Búsqueda de los mejores hiperparámetros

Anteriormente se observó que las variables que arrojan un mejor desempeño en las métricas de evaluación del Random Forest fueron aquellas que quedaron seleccionadas mediante la regularización Lasso. Por tanto, se muestra la optimización de hiperparámetros mediante GridSearch y Random Search con sus respectivas métricas de desempeño.

Grid Search

```
{'criterion':'absolute_error',
'max_depth': 7, 'max_leaf_nodes': 23,
'min_samples_split': 5}

R²:0.8376 MSE: 2.00  MAE: 0.81
```

Random Search

```
{'criterion': 'squared_error',
'max_depth': 7, 'max_leaf_nodes': 16,
'min_samples_split': 79.0}

R²: 0.735 MSE: 3.26 MAE: 1.06
```

Se observa que los hiperparámetros que generan las mejores métricas de evaluación son los encontrados mediante Grid Search,

5.1.2 Dataset Cálculo

Métricas de evaluación – Conjunto de datos testing.

Variables seleccionadas mediante método wrapper

```
MSE: 2.9326 MAE: 1.0558 R2: 0.8514
```

Según las métricas mostradas, podemos deducir un buen desempeño del modelo, según su R^2 el modelo es capaz de explicar la variabilidad de la nota 03 en aproximadamente un 85%.

Variables seleccionadas mediante Lasso

```
MSE: 2.9123  MAE:1.0653  R2: 0.8542
```

Según los resultados obtenidos, podemos concluir que el modelo tiene un desempeño bastante similar al alimentado con las variables escogidas con el método wrapper.

Búsqueda de los mejores hiperparámetros

Dado que los métodos de selección de variables dan resultados bastante simulares, la búsqueda de los mejores hiperparámetros se hará basados en el modelo que contempla las variables seleccionadas por el método wrapper.

Grid Search

```
{'criterion': 'squared_error',
'max_depth': 7, 'max_leaf_nodes': 25,
'min_samples_split': 6}
```

Con base a estos parámetros se corrió el modelo nuevamente, y se obtuvieron los siguientes resultados en sus métricas:

```
R² Score: 0.846 MSE: 2.94 MAE: 1.00
```

Random Search

```
{'criterion': 'squared_error',
'max_depth': 7, 'max_leaf_nodes': 17,
'min_samples_split': 25.0}
```

Con base a estos resultados se corrió nuevamente el modelo, que arrojó las siguientes métricas de desempeño:

```
R²: 0.839 MSE: 3.06 MAE: 1.02
```

Como podemos notar en las métricas de desempeño con los hiperparámetros optimizados, el modelo tuvo un desempeño peor que el realizado sin la optimización de hiper parámetros.

5.2 XGBoost - Extreme Gradient Boosting

5.2.1 Dataset Álgebra

Métricas de evaluación – Conjunto de datos testing.

Variables seleccionadas mediante método wrapper

```
R²: 0.728  MAE:1.168   MSE :3.354
```

Según las métricas de desempeño del modelo, podemos decir que tiene un desempeño aceptable. Analizando su R² podemos decir, que el modelo logra explicar en un 73% los datos de la variable objetivo.

Variables seleccionadas mediante Lasso

```
R²: 0.81   MAE:1.023     MSE :2.342
```

Como podemos observar en los resultados, tiene un mejor desempeño el dataset realizado con las variables del método integrado, ya que logra explicar aproximadamente en un 81% la variable objetivo.

Búsqueda de los mejores hiperparámetros

Anteriormente se observó que las variables que arrojan un mejor desempeño en las métricas de evaluación del XGBoost fueron aquellas que quedaron seleccionadas mediante la regularización Lasso. Por tanto, se muestra la optimización de hiperparámetros.

El número de árboles óptimos para este modelo que contempla las variables seleccionadas mediante Lasso es 50, generando las siguientes métricas de desempeño:

```
R²: 0.848   MAE:0.929   MSE :1.868
```

5.2.2 Dataset Cálculo

Métricas de evaluación – Conjunto de datos testing.

Variables seleccionadas mediante método wrapper

```
R²: 0.8527 MAE:1.275   MSE:2.808
```

Según las métricas, es posible concluir que tiene un muy buen desempeño. Analizando su R² podemos decir, que el modelo logra explicar en un 85% los datos de la variable objetivo.

Variables seleccionadas mediante Lasso

```
R²: 0.853   MAE: 1.275   MSE: 2.808
```

Como podemos observar en los resultados, las métricas arrojadas por este método son bastante similares. Ahora bien, debemos esperar a observar qué pasa al momento de realizar la optimización de hiperparámetros.

Búsqueda de los mejores hiperparámetros

En el caso de los hiperparámetros para este modelo se buscó el número de árboles para hacer el ensamble que fuera óptimo, al realizar el método nos arrojó que el óptimo serían *50* por lo que, se corrió nuevamente con base a estos resultados y las métricas obtenidas fueron:

Variables seleccionadas mediante método wrapper

```
R²: 0.896 MAE: 0.886  MSE: 1.974
```

Como se puede evidenciar en las métricas de desempeño, hay una mejoría considerable luego de la aplicación de hiper parámetro lo que nos deja un R² de casi el 90%, una métrica bastante buena, pero se debe tener cuidado con temas de sobreajuste del modelo.

Variables seleccionadas mediante Lasso

```
R²: 0.871 MAE:0.926  MSE :2.467634
```

Se puede evidenciar una mejoría leve en el R², ya que explica en un 2% más la variabilidad de los datos en la variable nota 03 con respecto al modelo sin optimización de hiperparámetros.

6. MEJORES MODELOS SELECCIONADOS

6.1 Dataset Álgebra

A continuación, se muestra un cuadro comparativo que resume las métricas de desempeño. En este cuadro se muestra únicamente las métricas de la regresión lineal y los algoritmos planteados con los diferentes métodos de selección de variables ya con sus hiperparámetros optimizados mediante el método de la cuadrilla y el aleatorizado.

ÁLGEBRA			
Algoritmo / Métricas	MAE	MSE	R2
Regresión lineal	0.97	2.54	0.75
Variables seleccionadas mediante método wrapper + GridSearch	0.96	2.58	0.79
Random Forest Variables seleccionadas mediante regularización L1 + GridSearch	0.83	2.23	0.82
Random Forest Variables seleccionadas mediante método wrapper + Random Search	1.06	3.37	0.73
Random Forest Variables seleccionadas mediante regularización L1 + Random Search	1.06	3.26	0.73
XGBoost Variables seleccionadas mediante método wrapper + Ensamble	0.89	2.01	0.84
XGBoost Variables seleccionadas mediante regularización L1 + Ensamble	0.93	1.87	0.85

Figura 10. Resumen de métricas de desempeño.

Observamos que el modelo que genera mejores métricas de desempeño es un XGBoost, empleando un método de selección de características mediante regularización Lasso y búsqueda de mejores hiperparámetros mediante un ensamble.

6.2 Dataset Cálculo

Igualmente, se muestra un cuadro comparativo que resume las métricas de desempeño. En este cuadro se muestra únicamente las métricas de la regresión lineal y los algoritmos planteados con los diferentes métodos de selección de variables ya con sus hiperparámetros optimizados mediante el método de la cuadrilla y el aleatorizado.

CÁLCULO			
Algoritmo / Métricas	MAE	MSE	R2
Regresión lineal	1.56	5.39	0.76
Random Forest Variables seleccionadas mediante método wrapper + GridSearch	1.09	3.61	0.81
Random Forest Variables seleccionadas mediante regularización L1 + GridSearch	0.97	2.99	0.84
Random Forest Variables seleccionadas mediante wrapper + Random Search	1.11	0.64	0.80
Random Forest Variables seleccionadas mediante regularización L1 + Random Search	1.11	3.23	0.83
XGBoost Variables seleccionadas mediante método wrapper + Ensamble	1.09	2.61	0.86
XGBoost Variables seleccionadas mediante regularización L1 + Ensamble	1.16	3.05	0.84

Figura 11. Resumen de métricas de desempeño 2

Aunque a primera vista pareciera que todos los modelos presentan un comportamiento similar, el XGBoost con una selección de características con el método wrapper y su optimización de hiperparámetros genera métricas de desempeño más deseable de manera global.

7. CONCLUSIONES

7.1 El análisis exploratorio de los datos constituye una de las etapas que más importantes y que más consume tiempo dentro del marco de un proyecto de machine learning, dado que permite un correcto entendimiento del problema.

7.2 La selección de características y las estrategias utilizadas para la misma afectan de manera significativa el rendimiento de cualquier algoritmo de machine learning.

7.3 La optimización de los hiperparámetos afecta de manera significativa las métricas de desempeño de los algoritmos.

6. REFERENCIAS

[1] Rosidi, N. (s/f). Advanced feature selection techniques for machine learning models. KDnuggets. Recuperado el 26 de septiembre de 2023, de https://www.kdnuggets.com/2023/06/advanced-feature-selection-techniques-machine-learning-models.html