PREDICCIÓN DE INGRESO A LA UNI

Sebastian Alejandro Cajo Peña

1. Preguntas planteadas

- ¿Cuántos postulantes al concurso de admisión ingresan? Se busca determinar la cantidad total de postulantes que logran ingresar de acuerdo al conjunto de datos disponible.
- ¿Qué tasa de ingreso tenemos? Se calcula la tasa de ingreso para conocer el porcentaje de estudiantes que superan el examen de admisión y poder usarlo como referencia para comparar la predicción.
- ¿Existe un perfil de alumno que ingresa? ¿Es posible identificar patrones en los datos geográficos de origen, educación o residencia que sugieran un perfil típico de alumnos con mayor probabilidad de ingreso?
- ¿Podemos predecir qué alumno que no haya ingresado tiene más probabilidad de ingresar en el siguiente examen de admisión? Utilizando modelos predictivos, se busca anticipar la probabilidad de ingreso en futuros exámenes.

2. Aproximación a la solución

Después de evaluar varios enfoques, se decidió utilizar Random Forest debido a su robustez, capacidad de manejar tanto variables categóricas como numéricas, y su alto rendimiento en términos de precisión. También se evaluó el algoritmo de Árbol de Decisión, pero se descartó en favor de Random Forest por su mejor desempeño en datos complejos y con un mayor número de variables.

Manejo de Datos Faltantes

- **Variables Categóricas:** Se realizaron gráficos de distribución para evaluar las cantidades de datos categóricos faltantes. Para reducir errores, se imputaron los valores faltantes con las categorías más frecuentes, dado que esta técnica asegura que los patrones de los datos no se vean alterados significativamente.
- **Variables Numéricas:** En el caso de las variables numéricas, se utilizó la mediana para imputar los valores faltantes. Este enfoque es menos sensible a los valores atípicos en comparación con el uso de la media, garantizando una mejor estabilidad en el modelo.

En ciertos casos, se consideraron relaciones entre columnas para inferir valores más precisos para los datos faltantes.

División del Conjunto de Datos

El conjunto de datos se dividió en dos partes: un 70% para entrenamiento y un 30% para prueba. Esta separación permitió evaluar adecuadamente el desempeño del modelo sobre datos no vistos.

Hiperparámetros del Modelo

Se configuraron los siguientes parámetros para el Random Forest:

- **n_estimators:** Se utilizaron 1500 árboles en el bosque para maximizar la robustez del modelo.
- **max_depth:** Se estableció una profundidad máxima de 25 niveles para prevenir el sobreajuste sin sacrificar demasiado la precisión.
- **random_state:** Se fijó una semilla aleatoria de 50 para garantizar la reproducibilidad de los resultados.

Se hicieron distintas pruebas variando los hiperparámetros para lograr una mejor aproximación de la probabilidad de ingreso sin que afecte significativamente a la Evaluación del modelo; se visualizó que a mayor número en los hiperparámetros, el modelo se optimizaba, pero consumía una cantidad significativa del procesamiento del entorno; para hiperparámetros con valores superiores el entorno se desconectaba debido a la falta de memoria RAM.

3. Hallazgos y Conclusiones

Precisión del modelo

El modelo de Random Forest alcanzó un 97.08% de precisión en las predicciones. Este alto nivel de acierto demuestra que el modelo fue capaz de aprender patrones relevantes a partir de los datos suministrados, lo que sugiere que las variables seleccionadas tienen una relación significativa con la probabilidad de ingreso. A medida que se modificaban los hiperparámetros se notaba un aumento en la precisión; sin embargo, no era demasiado, y los valores predichos por el modelo no diferían significativamente entre sí.

Importancia de las Variables

Algunas de las variables que más influyeron en la predicción incluyen:

- **Puntajes anteriores:** Los postulantes con mejores resultados en pruebas previas tuvieron una mayor probabilidad de ingreso.
- **Modalidad:** Los postulantes de la modalidad Ingreso Directo CEPRE presentaron una mayor tasa de éxito a comparación de las demás modalidades, incluyendo la principal, Ordinario.

Posibilidad de predecir futuros ingresos

El modelo puede ser utilizado para identificar a postulantes que, aunque no hayan ingresado en un cierto intento, presentan una alta probabilidad de hacerlo en el examen siguiente, basándose en las características geográficas e históricas; sin embargo, se recomienda alimentar con una mayor cantidad de datos que sirvan como factores externos al modelo.

Consideraciones

- Durante la implementación en Google Colab, se evidenció que el entorno fue superado en capacidad debido al tamaño del conjunto de datos y la complejidad del modelo.
- Para futuros análisis se debe utilizar un entorno de mayor capacidad de procesamiento o plataformas más robustas.
- Ampliar el conjunto de variables considerando factores externos no contemplados en esta implementación.
- Realizar pruebas adicionales con algoritmos más robustos para evaluar la efectividad del modelo usado.