

Informe del Proyecto: Predicción de Calificaciones de Estudiantes

Elaborado por: María José Rojas Mosquera

1. Objetivo del Proyecto

El propósito de este proyecto es predecir la calificación final de los estudiantes utilizando dos modelos de aprendizaje automático: Regresión Lineal y Random Forest. A lo largo del proceso, comparamos los resultados de ambos modelos para ver cuál se ajusta mejor a los datos y ofrece las predicciones más precisas.

2. Estrategia de Solución

La estrategia que seguimos se estructuró en varias fases, las cuales incluyen desde el procesamiento de los datos hasta la comparación de los modelos:

2.1 Preparación de los Datos

Empezamos cargando los datos desde un archivo CSV y luego transformamos las variables categóricas en variables dummy. Esto se hizo para que el modelo pudiera entender estas variables de manera adecuada. Después, se separaron las características (X) de la variable objetivo (G3, la calificación final) y se dividieron en dos conjuntos: entrenamiento y prueba.

2.2 Implementación de los Modelos

A continuación, entrenamos dos modelos:

- **Modelo 1: Regresión Lineal:** Utilizamos este modelo como punto de partida para ver qué tan bien podía predecir las calificaciones de los estudiantes.
- **Modelo 2: Random Forest:** Este modelo más avanzado fue utilizado para hacer una comparación con el modelo de regresión lineal. Además, ajustamos los parámetros del modelo utilizando GridSearchCV, lo que ayudó a mejorar su rendimiento.

2.3 Evaluación de los Modelos

Para evaluar cómo funcionaron los modelos, medimos su desempeño con tres métricas clave:

- RMSE (Error cuadrático medio)
- R2 (Coeficiente de determinación)
- MAE (Error absoluto medio)

Estas métricas nos ayudaron a ver cuán cerca estaban las predicciones de los modelos de los valores reales.

2.4 Comparación de los Modelos

Una vez que ambos modelos fueron entrenados y evaluados, los comparamos utilizando las métricas mencionadas. Esto nos permitió ver cuál de los dos modelos tenía un mejor desempeño en términos de precisión y capacidad de ajuste a los datos.

3. Resultados Obtenidos

- **Regresión Lineal:**
Este modelo logró un R^2 de aproximadamente 0.40, lo que indica que pudo explicar solo una parte moderada de la variabilidad de las calificaciones. Aunque el modelo fue útil, los errores (medidos por RMSE y MAE) fueron bastante altos, lo que sugiere que hay espacio para mejorar.
- **Random Forest:**
Este modelo mostró un R^2 superior a 0.80, lo que significa que pudo capturar la mayor parte de la variabilidad de las calificaciones. Además, los errores fueron significativamente menores que los del modelo de regresión lineal, lo que demuestra una mejor capacidad de predicción.

4. Conclusiones

- Random Forest resultó ser el modelo más eficaz para predecir las calificaciones finales de los estudiantes, con un rendimiento mucho mejor en términos de precisión y ajuste a los datos en comparación con Regresión Lineal.
- La optimización de Random Forest con GridSearchCV permitió obtener un modelo más preciso al ajustar sus parámetros, lo que fue clave para mejorar los resultados.
- Este proyecto demuestra cómo los modelos más complejos como Random Forest pueden ofrecer mejores resultados, especialmente cuando las relaciones entre las variables son complejas y no lineales, como es el caso de las calificaciones de los estudiantes.