Documento de informe

Introducción

El problema planteado fue predecir el nivel de inglés que obtendrá un estudiante en el examen Saber pro a partir de diferentes características del mismo. El conjunto de datos usados durante el entrenamiento fue el compendio de los resultados individuales en las pruebas Saber Pro entre los años 2018 y 2021 de 1,12 millones de evaluados; a su vez, de cada examinado se contaba con 50 características (casi todas ellas cualitativas) y sus resultados en los 7 componentes de la prueba.

Se hicieron varias pruebas para determinar qué características eran las más apropiadas; en el planteamiento del proyecto se consideró usar el núcleo asociado al pregrado, pero debido a la gran cantidad de valores individuales que podía tomar (alrededor de 50) se tuvo que descartar, dado que la codificación usada para los valores categóricos (One-hot) no tiende a funcionar bien con este tipo de características y a que los tiempos de entrenamiento tendían a extenderse mucho. (Se intento usar otras formas de codificación pero tendían a dar muchos errores con otras secciones del código)

Se usaron finalmente las siguientes características (en el planteamiento del proyecto se consideró solo el núcleo asociado al pregrado y las primeras dos de la lista, pero se tuvo que incluir a las demás para mejorar el rendimiento):

- Metodología del programa
- Tipo de institución
- Valor matricula (Rangos, no valores exactos)
- Horas trabajadas por semana
- Si la matricula es pagada o no por los padres
- Estrato socioeconómico
- Si se tiene acceso a un computador en casa
- Si se tiene acceso a internet en casa
- Si está estudiando en este momento
- Si tiene crédito de estudios
- Si la familia tiene automóvil

Además, se eliminaron todas las muestras a las que les faltaba alguna de estas características y se tomaron 100.000 aleatoriamente para el entrenamiento.

Desarrollo

Se uso inicialmente codificación one-hot y PCA, aunque este ultimo no afecto significativamente ninguno de los tres modelos (debido a que no tiende a funcionar muy bien en variables categóricas) y no se implementó otra técnica de reducción dimensional dado que todas las características eran de baja cardinalidad; se entrenó un modelo SVM con kernel polinómico, un random forest de 300 árboles y una red neuronal con 6 capas (usando varias funciones de activación); todos los hiper parámetros fueron determinados por prueba y error, debido a que los tiempos de entrenamiento eran demasiado grandes para grid-search.

Resultados

Método	**F1 Score**	**MCC**
SVM	0,319	0,168

Random-forest	0,320	0,153
Red neuronal	0,373	0,173

Conclusiones

- Existen varias posibles razones por las cuales no se pudo conseguir el objetivo, puede ser que no exista una correlación directa entre las características disponibles y la etiqueta de interés, que sea necesario incluir alguna de las características de mayor cardinalidad (como el departamento de origen o el núcleo al que pertenece el pregrado del estudiante) o que la combinación de hiper parámetros usados no sea la adecuada (puede que exista un kernel muy específico para SVM o una arquitectura diferente para la red neuronal que de un resultado mejor).

- En caso de querer mejorar el resultado, podría intentarse usar una codificación diferente para las variables (especialmente importante si se incluyen características de alta cardinalidad) o una técnica alternativa para reducir la dimensionalidad del sistema; a su vez, podría ser favorable usar validación cruzada, a pesar del aumento que esto significaría en los tiempos de entrenamiento.