Trabajo práctico No.4

María Lucia Quintanilla Álvarez - Mariana Ospina Mejía

Este informe presenta un resumen detallado de la estrategia de solución propuesta, los resultados obtenidos y las conclusiones derivadas de la aplicación de dos modelos de aprendizaje automático supervisado en un conjunto de datos seleccionado del UCI Machine Learning Repository para el proyecto. Para abordar nuestro problema específico, encontramos un dataset adecuado que cuenta con la información necesaria, para obtener un conjunto de datos que consta de 5875 instancias y 19 atributos.

Antes de preprocesar los datos, realizamos un análisis exploratorio para familiarizarnos con el dataset. Examinamos la distribución de las variables numéricas, identificamos valores atípicos o anomalías, y calculamos estadísticas descriptivas como la media, mediana y desviación estándar. Además, utilizamos histogramas para visualizar las distribuciones, diagramas de caja para detectar valores atípicos y un mapa de calor de correlación para explorar las relaciones entre las variables.

Para el preprocesado y limpieza de datos se eliminaron los datos faltantes, duplicados o irrelevantes, seleccionando las variables con menor correlación con la variable de respuesta "Target", así como las columnas "sex" y "NHR".

Se selecciono dos algoritmos de aprendizaje automático supervisado para nuestro problema de regresión. En el primer modelo realizo un análisis de regresión utilizando el modelo Ridge, donde se divide los datos en conjuntos de entrenamiento y prueba, define el modelo y se realiza una búsqueda de hiperparámetros para encontrar los mejores parámetros.

Después de ajustar el modelo, evalúa su desempeño en el conjunto de prueba utilizando métricas como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2). Finalmente, guarda el modelo entrenado y los resultados, y visualiza la curva de aprendizaje junto con las predicciones versus valores reales.

En el segundo modelo, se realiza un análisis de regresión utilizando el modelo de Random Forest. Se lleva a cabo una búsqueda de hiperparámetros para encontrar los mejores valores, luego se ajusta el modelo y se evalúa su desempeño en el conjunto de prueba utilizando métricas similares al primer modelo. Además, se visualiza la curva de aprendizaje y las predicciones versus valores reales.

El análisis de concordancia entre las predicciones de los dos modelos reveló un coeficiente kappa de Cohen de -0.0038, indicando una concordancia muy baja o prácticamente nula entre los clasificadores evaluados. Este resultado subraya la necesidad de investigar las razones detrás de esta falta de concordancia y considerar posibles mejoras tanto en los modelos como en el proceso de clasificación. Además, se llevó a cabo una comparación visual de los gráficos de predicciones versus valores reales de cada modelo para una comprensión más detallada de su desempeño.