Informe sobre los Resultados de la Práctica Titanic

Introducción

En el marco de la práctica realizada por el grupo conformado por Isidro Javier García Fernández y Álvaro Sánchez Hernández, estudiantes del Doble Grado en Matemáticas e Ingeniería Informática, se abordó el desafío de encontrar el mejor clasificador para los datos del Titanic publicados en Kaggle (https://www.kaggle.com/c/titanic).

Preprocesamiento de Datos

Antes de proceder con la implementación de los clasificadores, se llevaron a cabo diversas tareas de preprocesamiento de datos. Inicialmente, se configuró la ruta del archivo CSV que contenía los datos del Titanic. Se eliminaron las observaciones con datos faltantes y columnas consideradas innecesarias para el análisis, como 'Name', 'PassengerId', 'Ticket', 'Cabin' y 'Embarked'. Además, se redujo la cantidad de observaciones para agilizar las pruebas, conservando solo las primeras 300 filas.

Validación Cruzada

Se empleó el método de validación cruzada para dividir el conjunto de datos en conjuntos de entrenamiento y prueba. El 20% de los datos se reservó para prueba, y el resto se utilizó para entrenar los modelos.

Resultados de los Clasificadores

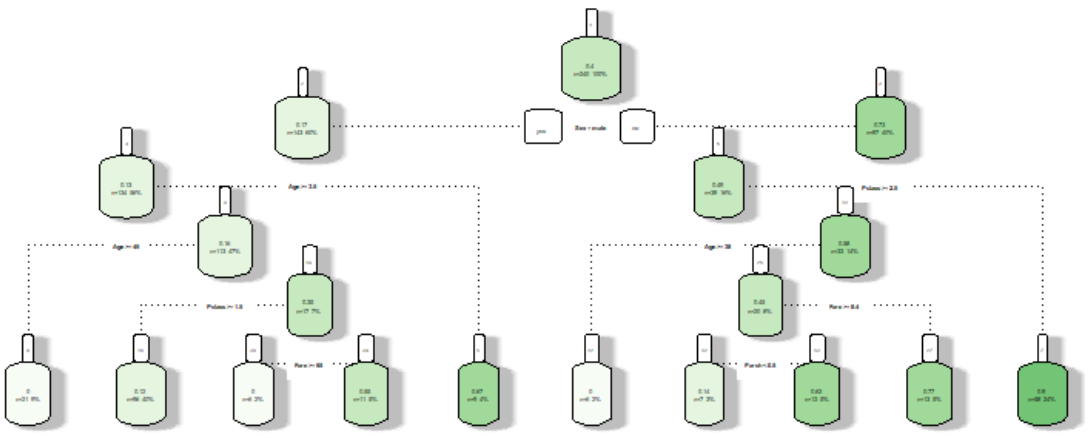
A continuación, se detallan los resultados obtenidos para cada uno de los clasificadores implementados:

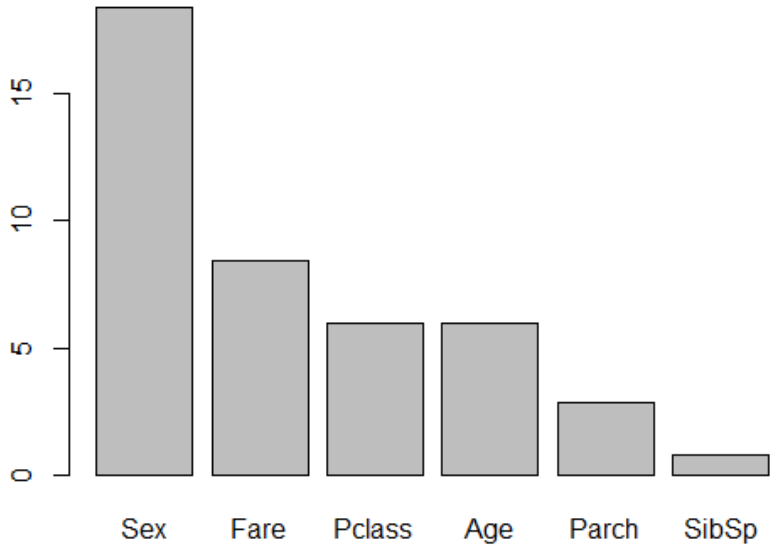
- **Perceptrón**:

Se utilizó un perceptrón con una capa oculta de 2 neuronas.
La precisión del modelo fue de aproximadamente 0.53.

- **Árbol de Decisión (RPart):**

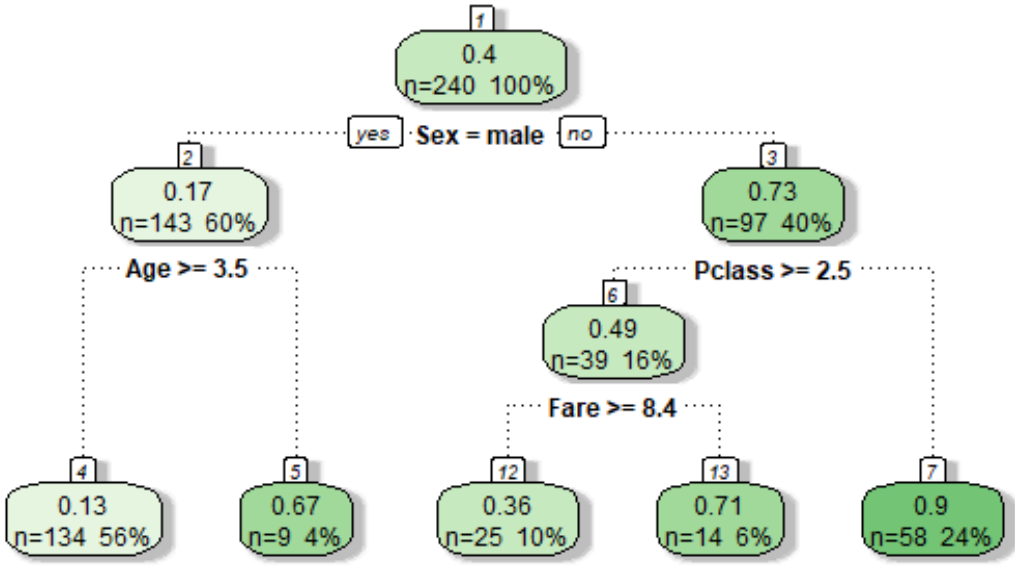
Se entrenó un árbol de decisión sin podar, obteniendo una precisión de aproximadamente 0.2.
Podemos ver una representación del árbol y un gráfico de la importancia de atributos.

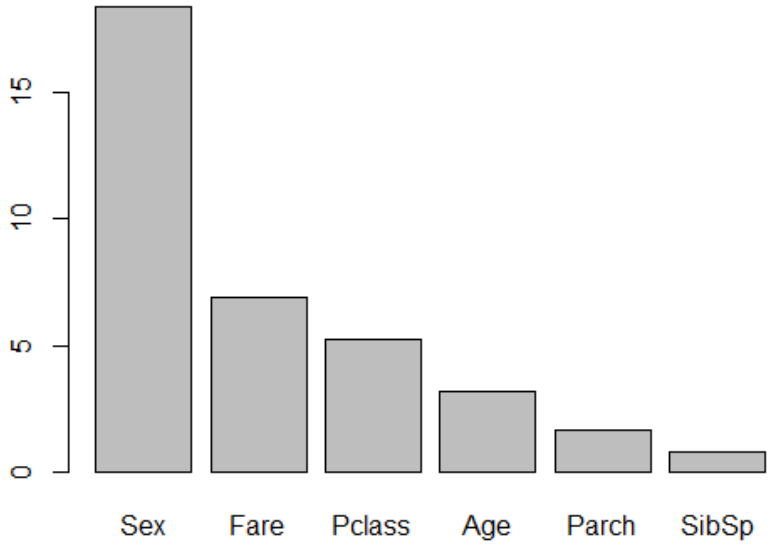




Vemos que el atributo “Sexo” es el atributo con mayor importancia. Esto tiene sentido dentro del problema, puesto que en el acontecimiento sobrevivieron un mayor número de mujeres que de hombres. Otros atributos que también son relevantes a la hora de clasificar son por ejemplo la “Clase”, puesto que sobrevivieron más personas de clases altas que de tercera clase.

Luego, se aplicó una poda manual utilizando el valor mínimo de complejidad (CP), mejorando la precisión a alrededor del 0.7.
Se exploraron diferentes valores de CP mediante la función tune.rpart, confirmando que la poda manual resultó en el mejor modelo.
De nuevo la precisión era de 0.7.
De nuevo se presenta una imagen del árbol podado, que podemos ver que permite una mejor clasificación en nuestro problema. También se presenta otro gráfico con importancia de atributos donde, de nuevo, el atributo “Sexo” es el más relevante. En este caso le otorga aún más importancia a ese atributo que en el caso anterior.





- **SVM (Support Vector Machine):**

Se implementaron SVMs tanto con kernel radial como lineal.
La precisión para ambas SVMs fue de aproximadamente 0.01.

Se utilizó la función tune para encontrar los mejores parámetros en el caso de un kernel polinómico, logrando una precisión de alrededor del 0.016.

Conclusiones

En términos generales, el árbol de decisión con poda manual mostró la mejor precisión entre los clasificadores evaluados, alcanzando un 0.7. Sin embargo, es importante destacar que estos resultados pueden variar dependiendo de la configuración exacta de los conjuntos de entrenamiento y prueba, así como de la semilla utilizada para la generación de números aleatorios.

Cabe mencionar que la implementación de la SVM no alcanzó un rendimiento destacado en este escenario, con una precisión bastante baja. Se recomienda explorar configuraciones adicionales y, posiblemente, considerar técnicas de ingeniería de características para mejorar los resultados.