Laura Cristina López Bedoya

CASO PRÁCTICO TAXI FLOTA AYTO. MADRID

Dando respuesta al requerimiento del Ayuntamiento de Madrid a la empresa IASA, se ha construido un algoritmo que clasifica medioambientalmente dicha flota según su grado de contaminación.

El Ayuntamiento de Madrid nos ha proporcionado los datos de 2018 referentes a la flota de taxis de la ciudad, y cuenta con más 15.000 vehículos. Cada licencia puede tener adscrito un único vehículo para la prestación del servicio. Este conjunto de datos se ofrece en un fichero junto con los datos históricos de la flota de taxi, indicando para cada vehículo, su matrícula, marca, modelo, combustible, potencia, número de plazas, etc.

Datos del ayuntamiento de Madrid

La contaminación fue medida a partir del tipo de combustible, y la Fecha de Matriculación (Año). En el desarrollo se implementaron cuatro algoritmos para determinar cual de ellos presentaba un mejor resultado para el caso de estudio. Los algoritmos usados fueron:

- SVM (Support Vector Machine)
- Árboles de Decisión
- Bosques aleatorios (Random Forest)
- Cuantificador Bayesiano Ingenuo

Para nuestro caso la variable a predecir es Clasificación medioambiental. En el dataset cada tipo de clasificación cuanta con el siguiente número de datos:

0	B	C	ECO
46	14927	11116	20829

De acuerdo con lo anterior se puede evidenciar un desbalance de los datos, debido a que el tipo 0 cuenta con un numero mucho menor a los demás tipos y esto puede generar un sesgo al aplicar los algoritmos.

Adjunto se encuentra el archivo del desarrollo realizado, en este se puede evidenciar el paso a paso del proceso ejecutado donde se realizó la carga de datos, con la exploración y preparación de los mismos para posteriormente aplicar los algoritmos presentados anteriormente. El dataset se divide en datos de entrenamiento del modelo con un 75% y el 25% restante serán los datos de prueba para evaluar el modelo.

Los valores de los datos de prueba con los que cuenta cada tipo de clasificación medioambiental son:

0	B	C	ECO
12	3732	2779	5207

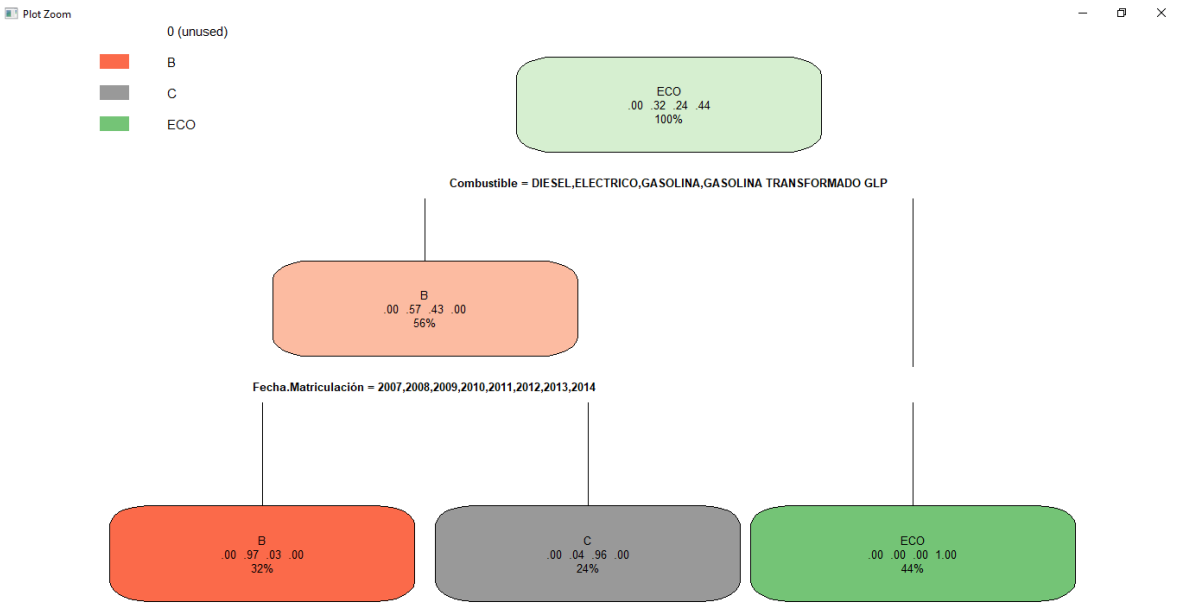
Como resultado de cada algoritmo se obtuvieron los siguientes resultados:

- SVM (Support Vector Machine) : En este algoritmo la predicción de la clasificación medioambiental del tipo 0 y ECO es correcta, sin embargo se presentan fallos en la predicción del tipo B y C.

```
confusionMatrix_svm
   prediccion
        0    B    C  ECO
0      12    0    0    0
B       0 3631  101    0
C       0   96 2683    0
ECO     0    0    1 5206
```

- Árboles de Decisión: En este algoritmo la predicción de la clasificación medioambiental del tipo ECO es correcta, sin embargo se presentan fallos en la predicción del tipo B y C. Para el tipo 0 no predice ningún valor, esto puede resultar del desbalance de los datos mencionado al inicio.

```
confusionMatrix_tree
   prediccion
        0    B    C  ECO
0       0    1   11    0
B       0 3631  101    0
C       0  128 2651    0
ECO     0    0    1 5206
```



- Bosques aleatorios (Random Forest): : En este algoritmo la predicción de la clasificación medioambiental del tipo 0 y ECO es correcta, sin embargo se presentan fallos en la predicción del tipo B y C.

```
confusionMatrix_rf
   prediccion
        0    B    C  ECO
0      12    0    0    0
B       0 3631  101    0
C       0   96 2683    0
ECO     0    0    0 5207
```

- Cuantificador Bayesiano Ingenuo: : En este algoritmo la predicción de la clasificación medioambiental del tipo 0 es correcta, sin embargo se presentan fallos en la predicción de los demás tipos.

```
confusionMatrix_nb
   prediccion
        0    B    C  ECO
0      10    1    0    1
B       0 3631  101    0
C       0  105 2674    0
ECO     0    0    1 5206
```

De cada matriz de confusión se calcula el porcentaje de acierto del algoritmo, obteniendo el siguiente resultado:

```
            modelos        correctos
1       Naive Bayes 98.2182438192668
2     Random Forest 98.3205456095482
3 Arbol de decisión 97.9369138959932
4               SVM 98.3120204603581
```

Se elige como mejor algoritmo para clasificación medioambiental Random Forest con un porcentaje de acierto de 98.320. En la matriz de confusión podemos evidenciar que este algoritmo presenta el menor numero de falsos positivos y negativos, comparado con el numero de valores que tiene cada tipo de clasificación en los datos utilizados como prueba.