UNIVERSIDAD DE GRANADA

Práctica 3. Clasificación de datos.

Tratamiento Inteligente de Datos.

Máster Profesional en Ingeniería Informática

Curso académico 2022/2023

Autor

José Alberto Gómez García

Índice

1. Introducción.

En esta práctica se buscará preprocesar los datos para poder realizar un estudio comparativo entre distintos algoritmos de clasificación. Se buscará obtener la precisión en predicciones sobre un conjunto de prueba lo más alta posible.

La base de datos a emplear en esta práctica contiene información sobre accidentes de tráfico. Nótese que es una base de datos bastante más sencilla que la utilizada en la anterior práctica.

Dado que el fichero "KNIME Workflow" sin resetear es demasiado pesado para ser subido a la entrega de Prado, este podrá encontrarse en este enlace a Google Drive. En la entrega de Prado se podrá encontrar el "KNIME Workflow" reseteado.

2. Preparación de los datos.

Como siempre que debemos analizar una base de datos, conviene echarle un vistazo para ver si hay valores nulos, errores o incoherencias que debamos tratar antes de comenzar el análisis en sí de los datos. En esta práctica, dado que la gran mayoría de los clasificadores debe hacer uso de valores numéricos, deberemos realizar la conversión de cadenas de texto a identificadores numéricos (Label Encoding).

Así pues, se realizan las siguientes conversiones.

- Para la columna "Fatality", los accidentes no mortales se codifican como 0, mientras que los mortales se codifican como 1.
- En la columna "Gender" los conductores se codifican como 0, mientras que las conductoras lo hacen como 1.
- Para la columna "Drug_Involvement", la ausencia de drogas se codifica como 0, la presencia de drogas como 1, y el "not_reported" se clasifica como 2. Este último bien podría corresponder a valores perdidos, aunque no lo consideraré así, ya que además había valores "unknown" que son más candidatos a ser auténtico valores perdidos. "Not_reported" podría hacer referencia a que la prueba se realizó, pero no trascendió su resultado, o alguna interpretación similar. Además, la mayoría de las observaciones (más de 14.000) corresponden a "Not_reported", por lo que considerarlo valor perdido e imputar podría provocar una gran pérdida de información y/o la generación de patrones que no se corresponden con la realidad.
- Para la columna "Atmospheric_condiction" se usa las siguientes correspondencias:
    - Clear = 0
    - Snow = 1
    - Cloudy = 2
    - Rain = 3
    - Sleet, Hail (Freezing Rain or Drizzle) = 4
    - Fog, Smog, Smoke = 5

    - Severe Crosswinds = 6
    - Blowing Snow = 7
    - Other = 8
    - Blowing Sand, Soil, Dirt = 9

  Los valores del 6 al 9 apenas tienen 100 ocurrencias entre todos ellos, por lo que podríamos darlos por perdidos e imputarlos en un intento de simplificar el modelo generado, sin embargo, los mantendremos. Nótese que la gran mayoría de los valores son “Clear” (21.286), “Cloudy” (3994), “Rain” (19749) y “Snow” (417).

- Para la columna “Roadway” se utilizan las siguientes correspondencias:
    - Urban-Principal Arterial-Other Freeways or Expressways = 0
    - Rural-Minor Collector = 1
    - Rural-Principal Arterial-Interstate = 2
    - Rural-Local Road or Street = 3
    - Urban-Minor Arterial = 4
    - Rural-Major Collector = 5
    - Urban-Other Principal Arterial = 6
    - Rural-Principal Arterial-Other = 7
    - Rural-Minor Arterial = 8
    - Urban-Local Road or Street = 9
    - Rural-Unknown Rural = 10
    - Urban-Principal Arterial-Interstate = 11
    - Urban-Collector = 12
    - Urban-Unknown Urban=13

  En este caso, los datos están distribuidos de una manera más uniforme, y no parece que dar por perdido algún valor pueda simplificar los modelos generados. A lo sumo, “Rural-Unknown Urban”, con 20 ocurrencias, podría darse por perdido, pero no se hará.

Además, deberemos tener en cuenta la gestión de valores imposibles y erróneos. En varias de las columnas existen valores erróneos, como “nan” o “\N”, por lo que eliminaremos las celdas con dicho contenido.

En la columna “Age” hay valores menores a 16, cuando en EE. UU. es ilegal conducir con menos de dicha edad. Pudiera ser que en realidad hubiera accidentes en los que el conductor tuviera 14 o 15 años (un niño que coge prestado el coche a su padre sin que este se dé cuenta, por ejemplo), pero un niño de 2 años no creo que provoque un accidente al volante. Por tanto, eliminaremos las celdas con valores menores a 16. Relativo a la columna de la edad, se observa que la mayoría de los datos se corresponden a edades entre los 20 y 40 años, aunque hay bastantes datos entre los 40 y 60 años. Por encima de los 60 años hay alguna ocurrencia, pero no se observa ningún outlier claro pues parece haber gente que conduce incluso con 90 años.

Relativo de la columna del alcohol, la gran mayoría de las observaciones no nulas corresponden a una tasa de alcoholemia de 0.0. Existen ocurrencias en el rango [0.0; 0.4] y no parece observarse ningún outlier, por lo menos teniendo en cuenta las tasas de alcoholemia permitidas en España. Cerca del 60% de los valores de esta columna están perdidos, por lo que uno podría plantearse su eliminación. Esto no se hará dado que la lógica nos dice que la tasa de alcoholemia es muy importante para determinar la gravedad de un accidente de tráfico, como posteriormente se verá en los experimentos, por lo que perderíamos una gran cantidad de información útil.

Una vez hemos limpiado y organizado la base de datos, deberemos imputar los valores nulos. Para ello, hemos seguido varios enfoques y hemos evaluado qué método era mejor a partir de la precisión de las predicciones en un conjunto de prueba haciendo uso de un árbol de clasificación C4.5 (al estilo de la práctica anterior). Los resultados obtenidos en función del método son los siguientes:

- Imputar con la media. Precisión del **87%**.
- Imputar con la moda. Precisión del 86%.
- Eliminar filas que tuvieran algún valor nulo. Precisión del 74.9%.
- Eliminar las columnas “Age”, “Alcohol_Results” y “Drug_Involvment” por tener muchos valores nulos, especialmente las dos últimas. No se eliminan las columnas “Roadway” y “Atmospheric_Condition” por tener 105 y 176 valores nulos solamente, lo cual podría considerarse asumible. Precisión del 69%.
- No imputar ningún valor. Precisión del 69.8%.

Parece ser que la mejor opción es imputar los valores haciendo uso de la media. Eliminar las filas que tienen algún valor nulo reduce la base de datos a un 30% de la original, por lo que perdemos mucha información. Por otra parte, eliminar columnas con valores nulos es sin duda la peor opción, ya que perdemos información que resulta clave para determinar la gravedad de un accidente, como la edad o si estuvieron implicados alcohol y/o drogas.

Tras analizar la base de datos, también nos damos cuenta de que la clase que buscamos clasificar, “Fatality”, no está balanceada (70% no_fatal, 30% fatal). Por tanto, y aunque el desbalanceo no es muy excesivo, durante la experimentación veremos qué resultados obtenemos antes y después de aplicar SMOTE, el cual balanceará las clases al 50/50 aproximadamente.

Para terminar este apartado, cabe mencionar que todos los valores serán normalizados antes de ser pasados a los diferentes modelos de clasificación. Hay algunos modelos, como los árboles de decisión, en los que la normalización de los datos puede ayudar a mejorar los resultados, pero no es determinante. Sin embargo, utilizaremos modelos basados en la idea de los “K vecinos más cercanos”, los cuales dependen mucho de que los datos estén normalizados para ofrecer buenos resultados.

3. Modelos utilizados.

Durante la experimentación utilizaremos los siguientes modelos de clasificación. Dado que no todos ellos han sido abordados con profundidad en clase de teoría, se ofrece una breve descripción de cada uno de ellos.

- **KNN** (K-Nearest Neighbors). Método de aprendizaje supervisado en el que se asigna a una nueva observación la etiqueta de la clase a la que pertenece la mayoría de sus vecinos más cercanos en el espacio de características. Presenta el inconveniente de que los datos deben estar balanceados, sino se tenderá a clasificar una nueva entrada al valor de la clase con más ocurrencias.
- **IB1.** Variante del anterior, dado que en lugar de asignar la etiqueta de la clase a la que pertenece la mayoría de sus vecinos, se utiliza una combinación lineal de las etiquetas de sus vecinos más cercanos para asignar la etiqueta a la observación. Esto permite al algoritmo ser más robusto y preciso que KNN en algunas situaciones.
- **Naive Bayes.** Método de aprendizaje supervisado basado en el teorema de Bayes y la suposición de independencia de las características. Se utiliza para asignar una nueva observación a una de varias posibles clases, basándose en la información proporcionada por las características de la observación y el conocimiento previo sobre la distribución de cada clase en el espacio de características.
- **Arboles de decisión.** Método de aprendizaje supervisado que se representa gráficamente como un árbol con nodos y ramas. Cada nodo del árbol representa una decisión, y las ramas representan las posibles opciones que se pueden tomar en cada decisión
- **Arboles de decisión orientados por gradiente.** Variante del modelo anterior en el que en lugar de seguir un enfoque “divide y vencerás” se utiliza el cálculo del gradiente para encontrar el mejor corte en cada nodo del árbol, lo cual mejora la precisión.
- **RandomForest.** Se hace uso de un conjunto de árboles de decisión entrenados de forma independiente. Cada árbol en el bosque se construye a partir de un subconjunto aleatorio de las características y muestras del conjunto de datos original. Cuando se realiza una predicción, cada árbol en el bosque emite una predicción y el resultado final se obtiene mediante el agregado de todas las predicciones individuales
- **AdaBoost**. Es un método de ensamblado que consta de varios clasificadores básicos entrenados de forma secuencial y agregados de forma que se obtenga un modelo final más preciso. Cada clasificador básico es entrenado utilizando un subconjunto de las muestras del conjunto de datos original, y el conjunto de muestras utilizado en cada etapa del entrenamiento se selecciona de manera que se ponga un mayor énfasis en las muestras que fueron clasificadas de forma incorrecta por los clasificadores anteriores.

- **Redes neuronales** (mediante los nodos RProp MLP Learner y MultiLayer Perceptron Predictor). La red neuronal se entrena utilizando un conjunto de datos etiquetados, ajustando los pesos de las conexiones entre las neuronas de forma que se minimice el error en las predicciones de la red. Una vez entrenada, la red neuronal puede utilizarse para clasificar nuevas observaciones basándose en sus características.

Se intentó hacer uso del clasificador SVM y K-Star (Weka), pero sus altos tiempos de ejecución (superiores a la hora y media) eran inasumibles, por lo que se tuvieron que descartar. En particular, sabemos que SVM no suelen ser apropiados para este tipo de problemas, dada la gran cantidad de datos que hay.

4. Resultados obtenidos.

Una vez hemos procesado los datos, pasemos a realizar los experimentos con los diferentes tipos de algoritmos de clasificación. Dado que no tenemos "muchos" datos (28.391 filas), haremos uso de la técnica de la validación cruzada. Así, entrenaremos el modelo varias veces (7), cada una de ellas con un conjunto de datos de prueba distinto. Esto permite evaluar el rendimiento del modelo de manera más precisa

Los resultados obtenidos utilizando la base de datos "original", a la que se le han imputado por media lo valores perdidos, se muestra a continuación.

Algoritmo	**Precisión (Aciertos/Fallos)**	**ROC**
KNN	83.058% (3368/687)	0.833
IB1	78.150% (3169/886)	0.734
Naive Bayes	81.233% (3294/761)	0.856
Decision Tree	82.121% (33300/725)	0.818
Gradient Boosted Decision Tree	**86.683% (3515/540)**	**0.906**
RandomForest	83.502% (3386/669)	0.869
AdaBoost	85.869% (3482/573)	0.884
Red neuronal	85.031% (3448/607)	0.869

En líneas generales, los clasificadores son capaces de realizar predicciones bastante acertadas. En lo relativo a precisión, el clasificador que peor se comporta es la variación del KNN con nombre IB1, implementado en Weka; mientras que el que mejores resultados proporciona es el árbol de decisión orientado por gradiente, con una precisión del **86.683%**.

El área bajo la curva mide la capacidad de un modelo de clasificación binaria para diferenciar entre dos clases. Se mide como la curva trazada por la tasa de verdaderos positivos (TPR) en función de la tasa de falsos positivos (FPR) mientras se varía el umbral de clasificación. Un modelo perfecto tendría un coeficiente ROC de 1, mientras que un modelo aleatorio tendría un coeficiente ROC de 0,5.

De los utilizados, el modelo que mayor ROC es el árbol orientado por gradiente, con un valor de **0.906,** mientras que el peor vuelve a ser la variación IB1 del algoritmo KNN.

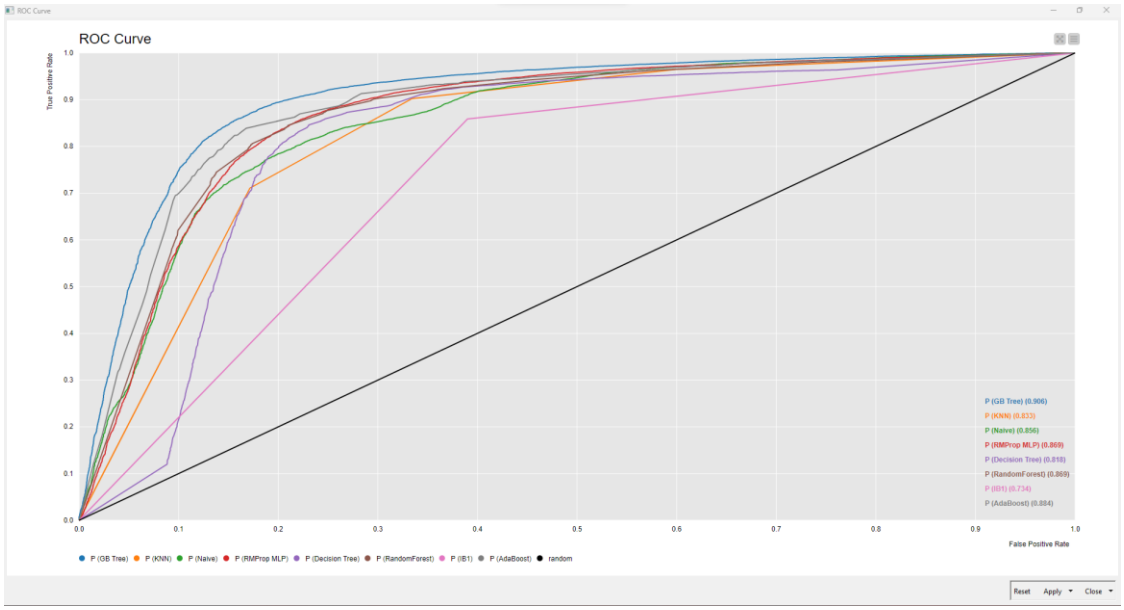

Imagen 1. Curva ROC de los distintos modelos de clasificación.

Por otra parte, si sobremuestreamos la clase minoritaria (accidentes con víctimas mortales), de manera que este balanceada, obtenemos estos otros resultados.

Algoritmo	**Precisión (Aciertos/Fallos)**	**ROC**
KNN	89.570% (10898/1269)	0.927
IB1	89.537% (10894/1273)	0.875
Naive Bayes	82.757% (10069/2098)	0.877
Decision Tree	88.428% (10759/1408)	0.883
Gradient Boosted Decision Tree	89.233% (10857/1310)	0.944
RandomForest	**90.984% (11070/1097)**	**0.951**
AdaBoost	88.083% (10717/1450)	0.917
Red neuronal	86.504% (10525/1642)	0.897

El peor algoritmo ha sido en esta ocasión Naive Bayes, con una respetable precisión del **82.757%**. Al sobremuestrear los datos, el modelo "Random Forest" permite obtener mayores precisión y coeficiente ROC que el árbol de decisión orientado por gradiente. "Random Forest" alcanza un **90.894%** de precisión y un coeficiente ROC de 0.951

En la siguiente imagen se recoge el conjunto de curvas ROC de los distintos modelos de clasificación evaluados en la práctica.

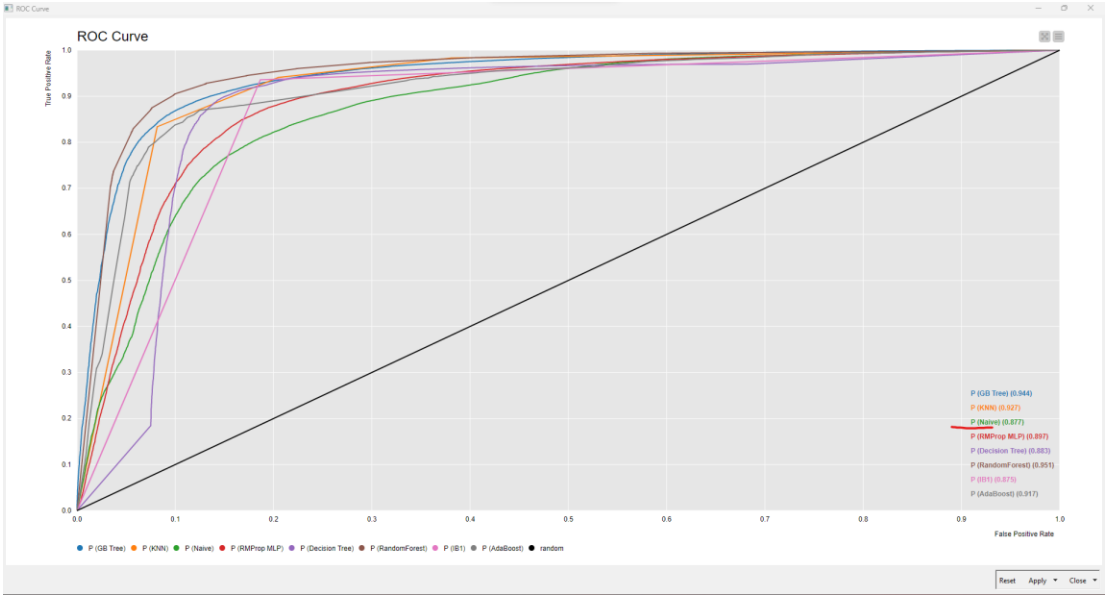

Imagen 2. Curva ROC de los distintos modelos de clasificación. Sobremuestreo aplicado.

A la hora de clasificar la gravedad de un accidente (con víctimas mortales o no), los errores pueden ocasionar grandes pérdidas. De entre los errores, consideramos que predecir como no mortal un accidente que sí es mortal es el peor de los errores. Así pues, comparamos modelos que tengan precisiones similares, ya que puede ser que un modelo con mayor precisión cometa menores errores, pero que estos nos supongan una mayor pérdida.

Si echamos un vistazo a las matrices de confusión de los dos mejores modelos para cada situación podemos obtener algunos datos curiosos.

File Hilite

Fatality \ Prediction (Fatality)	0	1
0	2596	220
1	320	919

Correct classified: 3,515 Wrong classified: 540

Accuracy: 86.683% Error: 13.317%

Cohen's kappa (κ): 0.679%

Imagen 3. Matriz de confusión "Gradient Boosted Tree"

File Hilite

Fatality \ P...	0	1
0	2603	213
1	360	879

Correct classified: 3,482 Wrong classified: 573

Accuracy: 85.869% Error: 14.131%

Cohen's kappa (κ): 0.656%

Imagen 4. Matriz de confusión "AdaBoost"

El modelo del árbol de decisión orientado por gradiente obtiene mejor precisión y menor número de errores a la hora de predecir errores mortales (esquina inferior izquierda); por lo que concluimos que es el mejor clasificador para el caso en que usemos la base de datos sin modificar.

Pasemos ahora a mirar los tres mejores modelos de clasificación al ser evaluados sobre un conjunto de prueba derivado del dataset sobremuestreado.

File Hilite

Fatality \ Prediction (Fatality)	0	1
0	7983	465
1	632	3087

Correct classified: 11,070 Wrong classified: 1,097

Accuracy: 90.984% Error: 9.016%

Cohen's kappa (κ): 0.785%

Imagen 5. Matriz de confusión "Random Forest". Dataset sobremuestreado.

File Hilite

Fatality \ Class [kNN]	0	1
0	7954	494
1	775	2944

Correct classified: 10,898 Wrong classified: 1,269

Accuracy: 89.57% Error: 10.43%

Cohen's kappa (κ): 0.749%

Imagen 5. Matriz de confusión "KNN". Dataset sobremuestreado.

File Hilite

Fatality \ Prediction (Fatality)	0	1
0	7927	521
1	752	2967

Correct classified: 10,894 Wrong classified: 1,273

Accuracy: 89.537% Error: 10.463%

Cohen's kappa (κ): 0.749%

Imagen 5. Matriz de confusión "IB1". Dataset sobremuestreado.

En esta ocasión, el mejor modelo sigue siendo aquel con mejor precisión ("Random Forest) al ser el que menos falla al predecir accidentes mortales. Sin embargo, entre los modelos que hacen uso de la técnica de los "K vecinos más cercanos" podría sernos más útil IB1, a pesar de tener una precisión ligeramente inferior que KNN, pues se equivoca menos al predecir accidentes mortales como no mortales.

Una vez hemos tratado qué modelo es mejor en cada situación, hablemos brevemente sobre que características son las que más información nos proporcionan. En primer lugar, comprobaremos el árbol de decisión (C4.5) que obtenemos al aplicar validación cruzada sobre el conjunto de datos original (con datos perdidos imputados por media).

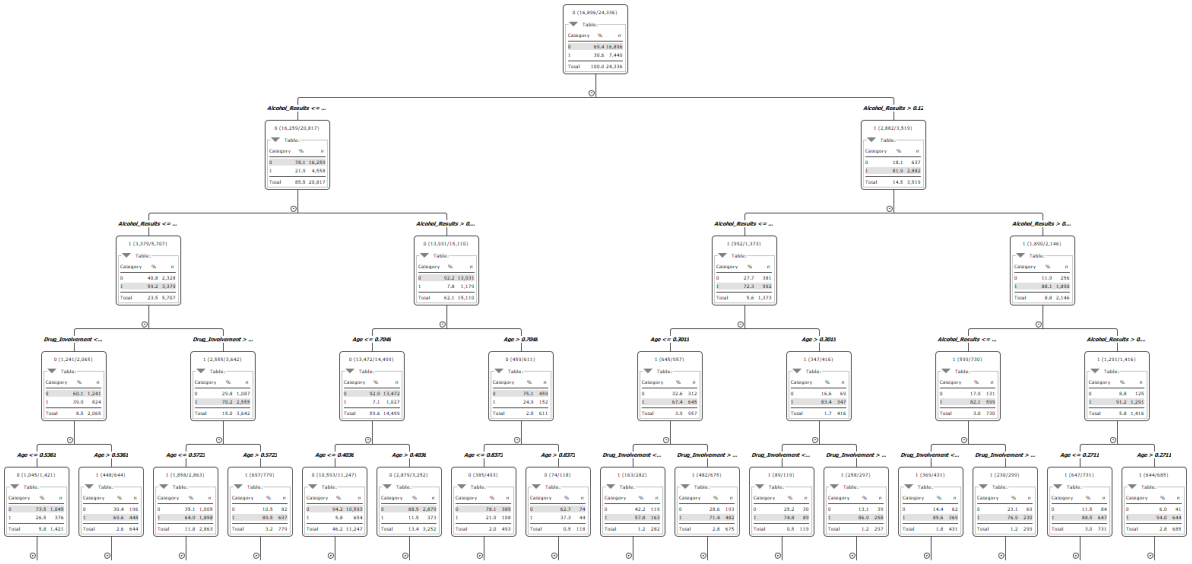

Imagen 6. Árbol de decisión sobre conjunto de datos original.

Si uno mira con un poco de detalle, pues es difícil encajar en un folio esta imagen, verá como en los dos primeros niveles se utiliza el nivel de alcohol, mientras que en los dos siguientes niveles se usa la edad y la implicación de drogas, en función de la rama.

Para ver la importancia de cada una de las variables, se han generado nuevos árboles de decisión, variando las variables que hacían uso para construirse. Como punto de partida, tengamos en cuenta que el árbol mostrado anteriormente permite obtener una precisión del **82.233%.** Las combinaciones de características probadas y precisiones obtenidas son las siguientes:

- Solo alcohol: 82.713%
- Solo alcohol y drogas: 84.439%
- Todo a excepción de alcohol y drogas: 63.576%
- **Edad, género y alcohol: 85.006%**

Podemos ver que, en ocasiones, es mejor prescindir de algunas variables, aunque intuitivamente parezcan que puedan dar información, como en este caso la condición climática, el tipo de carretera o el día de la semana. A raíz de los resultados anteriores, podemos deducir que la gran mayoría de los accidentes pueden ser clasificados como mortales o no en función del nivel de alcohol del conductor y su consumo de drogas; prescindir de dicha información hace que la precisión del clasificador se desplome.

5. Conclusiones y sugerencias.

En el marco de la asignatura, podemos destacar algunas conclusiones alcanzadas.

- La preparación de los datos es un aspecto importantísimo que no podemos descuidar antes de comenzar el análisis como tal de una base de datos.

- El sobremuestreo es necesario cuando las clases están desbalanceadas, aunque no sea mucho, pudiendo ayudarnos a construir mejores clasificadores.
- No existe un clasificador mejor absoluto, sino clasificadores que funcionan mejor en distintas situaciones. Destacamos el gran rendimiento de “Gradient Boosted Tree”, “AdaBoost” y “Random Forest”.
- Aunque la precisión en las predicciones es importante, no es el único factor a tener en cuenta a la hora de evaluar un clasificador. Puede que en ocasiones nos rente más un clasificador con mayor índice ROC, o uno ligeramente menos preciso pero que los errores que comenta sean más propensos a ser de un determinado tipo.

Respondiendo a la pregunta original de la práctica de “¿qué se recomendaría a la administración para evitar accidentes de tráfico con víctimas mortales?”, podemos esbozar algunas sugerencias, por obvias que parezcan.

- Aunque se dé negativo en el control de alcoholemia (de acuerdo con la legislación española) el consumo de drogas hace que los accidentes sean más propensos a ser mortales en todas las edades. Por tanto, convendría reforzar los controles de alcohol y drogas, y seguramente más los segundos.
- Una vez se ha dado positivo en el control de alcoholemia, los accidentes mortales son más frecuentes entre varones menores de 27 años los fines de semana. Por tanto, convendría llevar a cabo campañas de concienciación orientadas a dicho sector de la población, y aumentar los controles de alcohol y drogas en fin de semana.
- Una cantidad importante de los accidentes no mortales se producen dentro del límite legal de alcohol y cuando el conductor es menor de 40 años. Por lo que podría ser interesante abordar campañas de concienciación orientadas a este grupo de edad, recordando que en un accidente no tiene por qué morir siempre alguien, sino que el daño puede ser material, por ejemplo.
- Parece ser que la condición climática y el tipo de carretera no influyen demasiado en la gravedad de los accidentes, por lo que los controles probablemente deban situarse en lugares con mayor concentración de accidentes (dato que desconocemos) o cercanos a lugares de ocio donde se puedan consumir alcohol y/o drogas, especialmente los fines de semana. El que los controles se realicen en días soleados, lluviosos o nublados no parece influir demasiado.