
UNIVERSIDAD DE GRANADA

# PRÁCTICA 5:

## CLASIFICACIÓN CON KNIME

TRATAMIENTO INTELIGENTE DE DATOS

MÁSTER UNIVERSITARIO EN INGENIERÍA INFORMÁTICA

AUTOR

Pablo Valenzuela Álvarez (pvalenzuela@correo.ugr.es)

# CONTENIDO

# EJERCICIO 1. TIPOS DE CRISTAL

En el primer ejercicio trabajaremos con una base de datos usada para la investigación criminológica. Este conjunto contiene ejemplos de cristales de 7 tipos encontrados en escenas del crimen.

## CLASIFICADORES

Empezaremos construyendo varios clasificadores como nos indica el guion de prácticas.

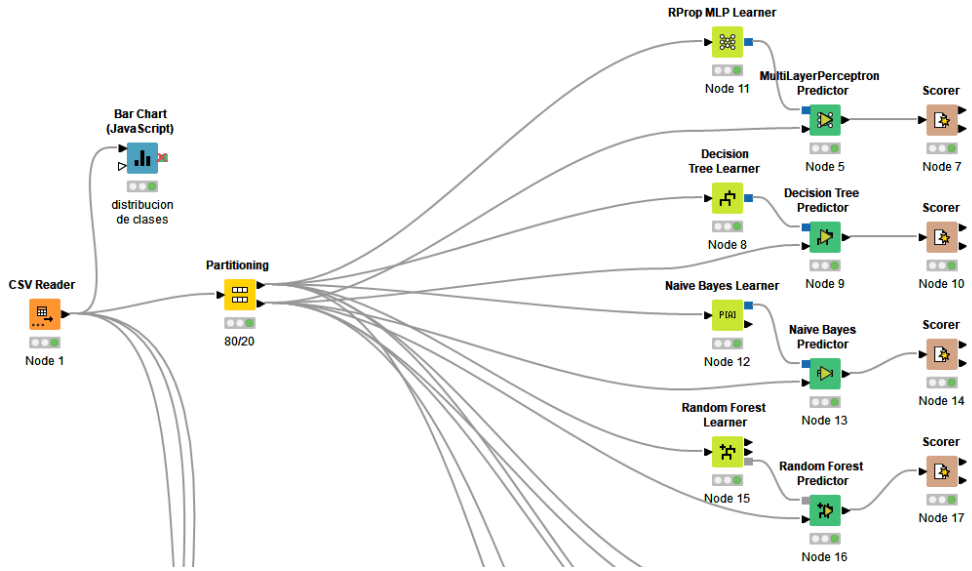


Figura 1: Parte de los clasificadores usados.

## COMPARATIVA DE MÉTRICAS

En la siguiente tabla podemos ver la métrica de precisión (**accuracy**) de cada clasificador usado.

| Clasificador | Precisión (accuracy) |
|---|---|
| **MultiLayerPerceptron** | 58,14% |
| **Decision Tree** | **67,442%** |
| **Naive Bayes** | 37,209% |
| **Random Forest** | **79,07%** |
| **K Nearest Neighbor** | 60,465% |
| **K Nearest Neighbor (distance)** | 55,814% |
| **SVM** | 60,465% |

Tabla 1. Comparativa de métricas de los clasificadores.

El clasificador **Random Forest** es el que destaca seguido de lejos por **Decision Tree**. La ventaja significativa de estos dos clasificadores indica que, al construir y combinar los resultados de sus árboles de decisión, se reduce el sobreajuste y mejora la generalización, lo que ayuda a obtener una mayor precisión.

## APLICANDO VALIDACIÓN CRUZADA

Como última parte del ejercicio 1, aplicaremos validación cruzada sobre el conjunto de datos, para ello usaremos los dos mejores clasificadores obtenidos en la parte anterior: **Random Forest** y **Decision Tree**.

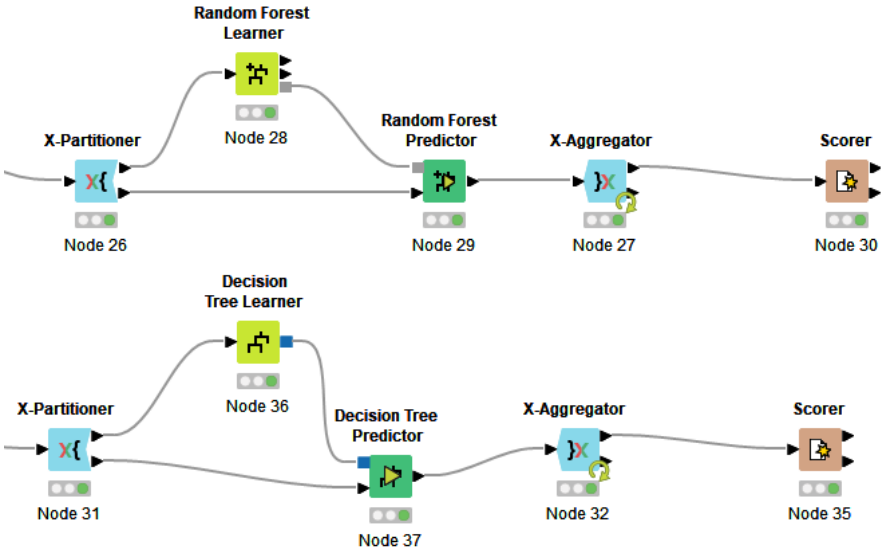


Figura 2: Validación cruzada realizada sobre los clasificadores.

El objetivo de realizar esta técnica es, medir el error promedio y calcular la desviación típica respecto a ese error.

Para medir el error promedio usaremos esta fórmula:

$$Error = \frac{1}{10}\sum_{i=1}^{10} Error_i$$

Y para la desviación típica:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n}}$$

(Siendo $e_i$ el valor del error de cada partición y $\bar{e}$ el valor medio del error).

File Hilite

| Type \Pre... | 1 | 2 | 3 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| 1 | 61 | 7 | 1 | 0 | 0 | 1 |
| 2 | 10 | 60 | 1 | 3 | 1 | 1 |
| 3 | 7 | 5 | 5 | 0 | 0 | 0 |
| 5 | 0 | 2 | 0 | 10 | 0 | 1 |
| 6 | 0 | 1 | 0 | 0 | 8 | 0 |
| 7 | 1 | 1 | 0 | 1 | 0 | 26 |

Correct classified: 170 Wrong classified: 44

Accuracy: 79,439% Error: 20,561%

Cohen's kappa (κ): 0,717%

Figura 3: Random Forest: precisión y error tras la validación cruzada.

| Row ID | Error in % | Size of Test Set | Error Count |
|---|---|---|---|
| fold 0 | 18.182 | 22 | 4 |
| fold 1 | 23.81 | 21 | 5 |
| fold 2 | 18.182 | 22 | 4 |
| fold 3 | 23.81 | 21 | 5 |
| fold 4 | 14.286 | 21 | 3 |
| fold 5 | 18.182 | 22 | 4 |
| fold 6 | 14.286 | 21 | 3 |
| fold 7 | 31.818 | 22 | 7 |
| fold 8 | 19.048 | 21 | 4 |
| fold 9 | 23.81 | 21 | 5 |

Figura 4: Random Forest: tasa de error de las porciones de la validación cruzada.

| Type \Pre... | 1 | 2 | 3 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| 1 | 53 | 12 | 4 | 0 | 0 | 1 |
| 2 | 12 | 54 | 5 | 3 | 1 | 1 |
| 3 | 8 | 3 | 6 | 0 | 0 | 0 |
| 5 | 0 | 2 | 0 | 10 | 0 | 1 |
| 6 | 2 | 2 | 0 | 0 | 5 | 0 |
| 7 | 1 | 3 | 1 | 0 | 1 | 23 |

Correct classified: 151 Wrong classified: 63

Accuracy: 70,561% Error: 29,439%

Cohen's kappa (κ): 0,597%

Figura 5: Decision Tree: precisión y error tras la validación cruzada.

| Row ID | Error in % | Size of Test Set | Error Count |
|---|---|---|---|
| fold 0 | 27.273 | 22 | 6 |
| fold 1 | 23.81 | 21 | 5 |
| fold 2 | 31.818 | 22 | 7 |
| fold 3 | 28.571 | 21 | 6 |
| fold 4 | 23.81 | 21 | 5 |
| fold 5 | 27.273 | 22 | 6 |
| fold 6 | 23.81 | 21 | 5 |
| fold 7 | 36.364 | 22 | 8 |
| fold 8 | 38.095 | 21 | 8 |
| fold 9 | 33.333 | 21 | 7 |

Figura 6: Decision Tree: precisión y error tras la validación cruzada.

| Resultados tras aplicar validación cruzada | | |
|---|---|---|
| **Random Forest** | Precisión | 79,439% |
| | Media error | 20,5414% |
| | Desviación típica | **5,063630638** |
| **Decision Tree** | Precisión | 70, 561% |
| | Media error | 29,4157% |
| | Desviación típica | **4,988916396** |

Tabla 2: Resultados tras aplicar validación cruzada.

Analizando las desviaciones de ambos clasificadores vistas en la tabla 2, podemos decir que son relativamente similares e indican una consistencia equilibrada en los resultados. En este caso, la desviación del clasificador Decision Tree es ligeramente menor, lo que sugiere una coherencia mayor en las predicciones en relación con su media. La elección de un clasificador u otro dependerá de otros factores como: la complejidad del problema, interpretabilidad del modelo, etc.

# EJERCICIO 2: ACCIDENTES DE TRÁFICO

En el segundo ejercicio aplicaremos distintas técnicas sobre un conjunto de datos relacionado con accidentes de tráfico. Estas técnicas estás destinadas mejorar la toma de decisiones al predecir si los accidentes tendrán víctimas mortales o no.

## PREPROCESADO

La primera medida que vamos a tomar, es arreglar los valores perdidos. La característica ***Alcohol_Results*** (marcada en rojo en la figura 7) contiene muchos de ellos y la decisión que se tomó fue la de tratar estos valores como **0,0**. Si los hubiese eliminado habríamos perdido casi de 18000 casos y haberlos imputado hubiese "falseado" el resultado de las predicciones. Los demás valores perdidos serán eliminados ya que sólo representan 300 casos de los 28300 totales.

| Row ID | I weekday | I Age | S Gender | S Alcohol... | S Drug_I... | S Atmosp... | S Roadway | S Fatality |
|---|---|---|---|---|---|---|---|---|
| Row0 | 6 | 18 | Female | ? | Not Reported | Clear | Urban-Principal Arterial-Other Freeways or Expressways | no_fatal |
| Row1 | 2 | 54 | Female | ? | No | Rain | Urban-Principal Arterial-Other Freeways or Expressways | no_fatal |
| Row2 | 7 | 29 | Male | 0.21 | Not Reported | Cloudy | Rural-Minor Collector | fatal |
| Row3 | 5 | 55 | Male | ? | Not Reported | Clear | Rural-Principal Arterial-Interstate | no_fatal |
| Row4 | 2 | 54 | Male | 0.34 | Unknown | Clear | Rural-Local Road or Street | fatal |
| Row5 | 6 | 29 | Female | ? | Not Reported | Clear | Rural-Local Road or Street | no_fatal |
| Row6 | 5 | 22 | Male | ? | Not Reported | Clear | Urban-Minor Arterial | no_fatal |
| Row7 | 6 | 24 | Male | 0.0 | Not Reported | Cloudy | Rural-Principal Arterial-Interstate | no_fatal |
| Row8 | 7 | 21 | Female | ? | No | Clear | Rural-Principal Arterial-Interstate | no_fatal |
| Row9 | 7 | 29 | Male | 0.1 | Unknown | Cloudy | Rural-Major Collector | no_fatal |
| Row10 | 6 | 48 | Male | 0.16 | Not Reported | Clear | Urban-Other Principal Arterial | no_fatal |
| Row11 | 3 | 65 | Male | ? | Not Reported | Cloudy | Rural-Principal Arterial-Interstate | no_fatal |
| Row12 | 4 | 11 | Male | ? | Not Reported | Clear | Rural-Major Collector | no_fatal |
| Row13 | 4 | 57 | Male | ? | Not Reported | Clear | Rural-Minor Collector | fatal |
| Row14 | 7 | 50 | Male | 0.17 | Unknown | Clear | Rural-Major Collector | fatal |
| Row15 | 5 | 53 | Female | 0.0 | Not Reported | Clear | Urban-Other Principal Arterial | fatal |
| Row16 | 7 | 26 | Female | ? | Not Reported | Clear | Rural-Minor Arterial | fatal |
| Row17 | 5 | 16 | Male | 0.0 | No | Clear | Rural-Local Road or Street | no_fatal |
| Row18 | 6 | 12 | Male | ? | Not Reported | Clear | Rural-Principal Arterial-Other | no_fatal |

Figura 7. Valores perdidos de la característica Alcohol_Results.

Una exploración de los valores anómalos nos muestra que no hay valores que se salgan de la normalidad. El valor detectado (ver figura 8) en la dimensión *edad* corresponde a una persona de 99 años, y los observados en *Alcohol_Results* equivalen a todos los que no son "0,0". Podríamos determinar que no hay valores anómalos.

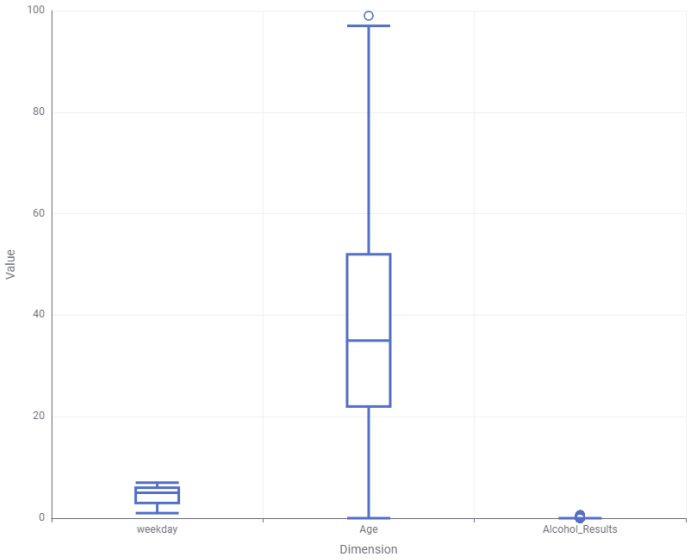


Figura 8. Exploración de los valores anómalos.

Por último, vamos a aplicar una técnica de sobre muestreo ya que el enunciado de la práctica nos dice que las clases no están balanceadas. Para ello, con el nodo **SMOTE** solucionamos este problema (ver figura 9).

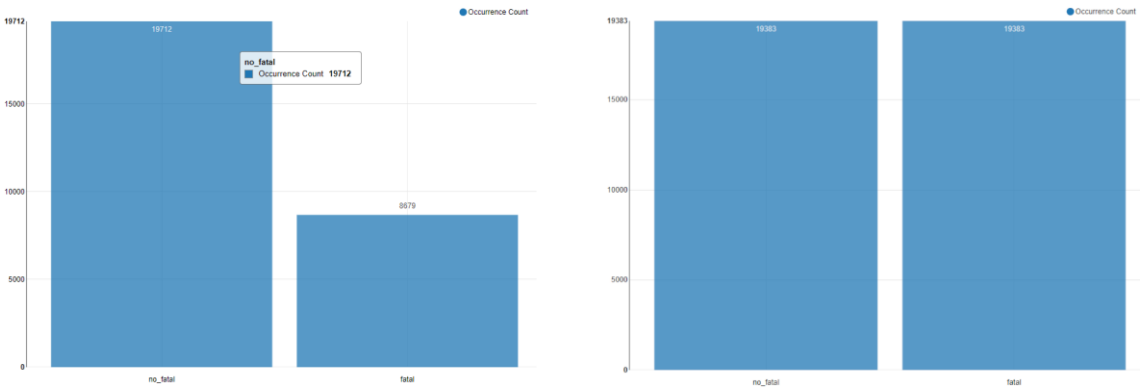


Figura 9. Antes y después de aplicar SMOTE.

## MODELOS DE CLASIFICACIÓN

En esta sección probaremos distintos parámetros en modelos de clasificación sobre los datos ya preprocesados, también se aplicará validación cruzada de 5 splits.

| Random Forest | | | | |
|---|---|---|---|---|
| **Split criterion** | **Limit number of levels** | **Minimun node size** | **Number of models** | **Accuracy** |
| Information gain ratio | 10 | 1 | 100 | 82,42% |
| Information gain | 10 | 1 | 100 | 83,019% |
| Gini index | 10 | 1 | 100 | 83,024% |
| Gini index | 100 | 1 | 100 | 83,669% |
| Gini index | 100 | 5 | 100 | 83,282% |
| Gini index | 100 | 1 | 200 | 83,607% |
| Gini index | 100 | 1 | 50 | **83,679%** |

Tabla 3: Prueba de parámetros en Random Forest

| Naive Bayes | | | | |
|---|---|---|---|---|
| **Default probability** | **Minimun standard deviation** | **Threshold standard deviation** | **Maximun nominal number per atribute** | **Accuracy** |
| 0,0001 | 0,0001 | 0,0 | 20 | **76,82%** |
| 0,001 | 0,0001 | 0,0 | 20 | 76,732% |
| 0,00001 | 0,0001 | 0,0 | 20 | 76,815% |
| 0,00001 | 0,0001 | 1,0 | 20 | 70,18% |
| 0,00001 | 0,0001 | 0,0 | 5 | 76,688% |

Tabla 4: Prueba de parámetros en Naive Bayes

| KNN | | |
|---|---|---|
| **Number of neigbours** | **Weight by distance** | **Accuracy** |
| 3 | No | **73,779%** |
| 2 | No | 73,711% |
| 5 | No | 73,438% |
| 3 | Si | 57,011% |

Tabla 5: Prueba de parámetros en KNN

| KNN (Mahalanobis Distance) | | |
|---|---|---|
| **Number of neigbours** | **Weight by distance** | **Accuracy** |
| 3 | No | **71,4%** |
| 3 | Si | 66,894% |

Tabla 6: Prueba de parámetros en KNN (Mahalanobis Distance)

## VISUALIZACIÓN

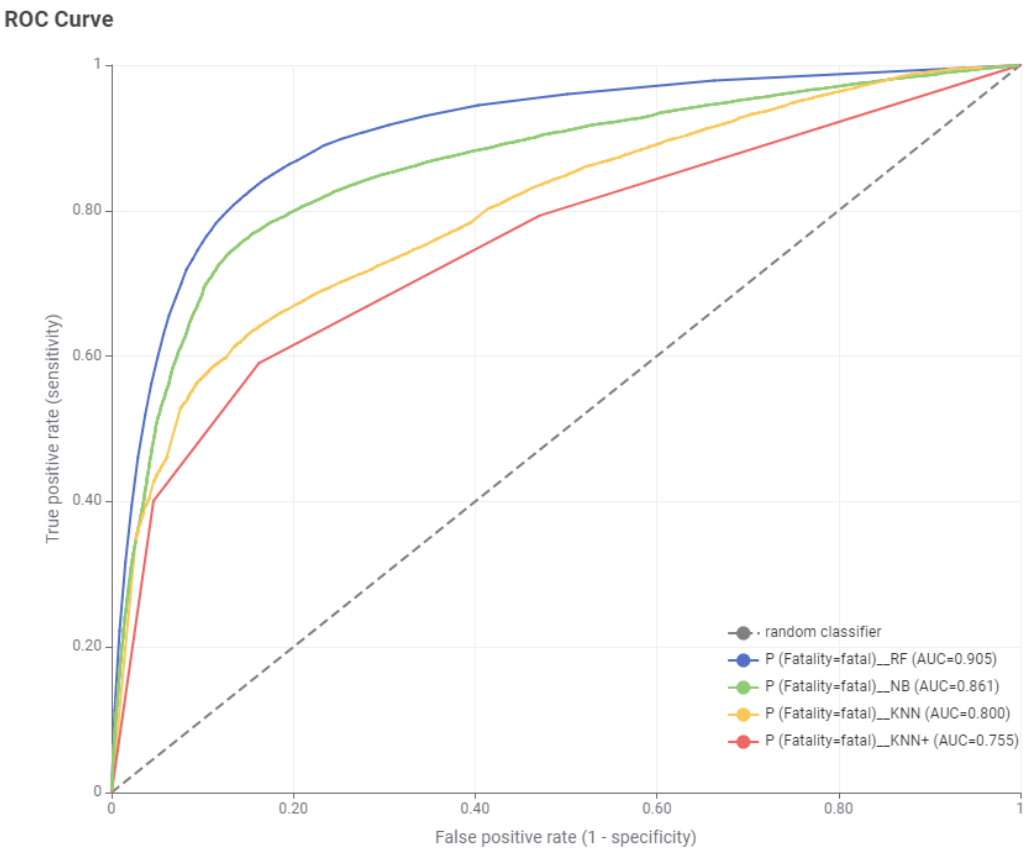


Figura 10. Curva ROC de los modelos de clasificación

## CONCLUSIÓN

Al ver los resultados obtenidos por los modelos en las tablas del 3 al 6 y analizando la gráfica de la curva ROC de la figura 10, podemos llegar a la conclusión de que el modelo **Random Forest** es el mejor. En precisión supera al resto de modelos usados en el ejercicio y también presenta un área bajo la curva superior, lo que lo lleva a ser una elección correcta para este tipo de problemas y la toma de decisiones informada.

## RECOMENDACIONES

Para finalizar el ejercicio, el guion nos insta a dar recomendaciones para evitar futuros incidentes. En la exploración de variables realizada, hemos distinguido las siguientes:

**Género:** se puede apreciar en la figura 11 una clara discriminación de esta variable hacia el género masculino. La única forma de influir en esta variable es la formación en educación vial de las personas, identificando situaciones en el tráfico y que se puede hacer en cada ocasión.

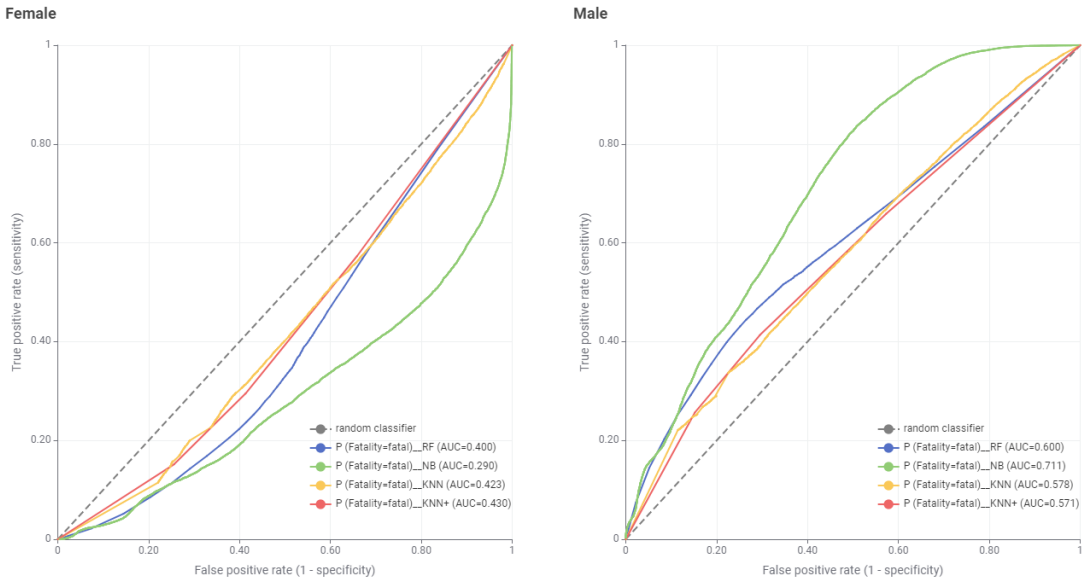


Figura 11. Curva ROC perteneciente al género

**Influencia de drogas:** Las drogas provocan situaciones peligrosas al volante, aparte de ser (la mayoría) ilegales, pueden causar disminución de la coordinación motora, aumento del tiempo de reacción, distracciones o somnolencia. Queda demostrado en la gráfica de la figura 12.

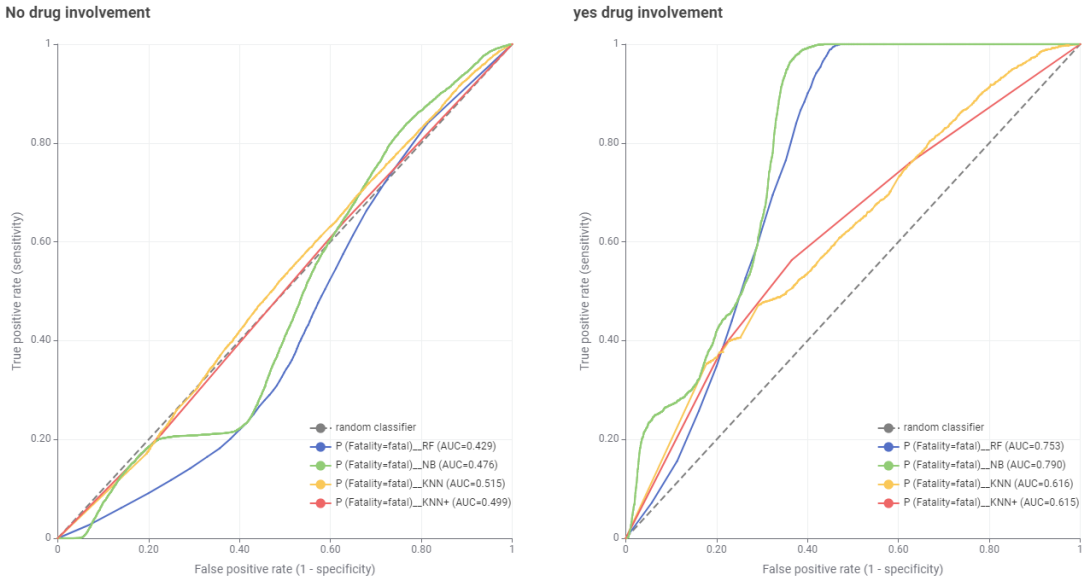


Figura 12. Curva ROC perteneciente a la influencia de drogas

**Tipo de carretera:** Como muestra la figura 13, una parte importante de accidentes mortales se produce en zonas residenciales tanto urbanas como rurales. Una solución a este problema puede ser la de la reducción de velocidad en carretera, mejor señalización de pasos de cebra, mejorar la iluminación en la noche, etc.

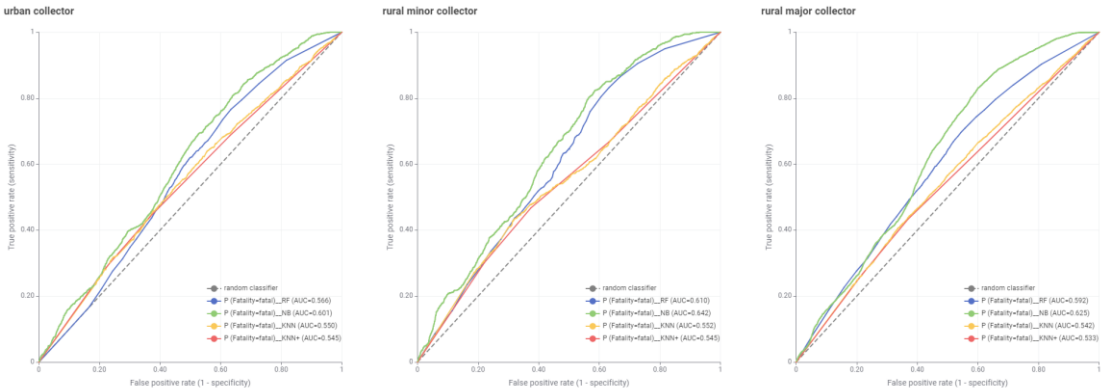


Figura 13. Curva ROC perteneciente al tipo de carretera