Asignatura	Datos del alumno	Fecha
Aprendizaje Automático	Apellidos: Proaño Guevara	03-ene-2020
	Nombre: Daniel	

Trabajo: Clasificación con Naive Bayes

- Análisis descriptivo de los datos.

  Los datos corresponden a votos de cada uno de los miembros de la cámara de representantes en 16 cuestiones diferentes, como votos para la ayuda a niños discapacitados, costos compartidos sobre un proyecto de agua, adopción de presupuestos, congelación de honorarios médicos, ayudas a el salvador, sobre grupos religiosos en las escuelas, prohibición de pruebas anti-satelitales, ayudas a Nicaragua, temas de misiles, inmigración, reducciones a la compañía synfuel, gastos en educación, derechos de demanda de fondos, crimen, exportaciones libre de impuestos, política administrativa sobre Sudáfrica, y en base a la forma que los representantes votaron se conoce si son miembros del partido Republicano o Demócrata

- Determinar el conjunto de modelización y el de validación.

  El conjunto de validación se separa en una proporción de 75-25 aleatoriamente, en el programa se establece una semilla de aleatoriedad para que el e que el experimento pueda ser repetido y se obtengan los mismos resultados, se utilizó la librería scikit-learn, la clase train_test_split.

- Tratamiento de missing.

  ¿Para el tratamiento de missing se utilizó el imputador de Scikit-Learn que analiza el dataset buscando valores correspondientes en este caso a <<?>>, y utiliza una estrategia de buscar los valores más frecuentes en las columnas. No se eliminaron las filas que contienen estos valores desconocidos ya que esto generaría un sesgo importante consecuente de la pérdida de información y por lo tanto no se podría generalizar el clasificador para nuevos datos.

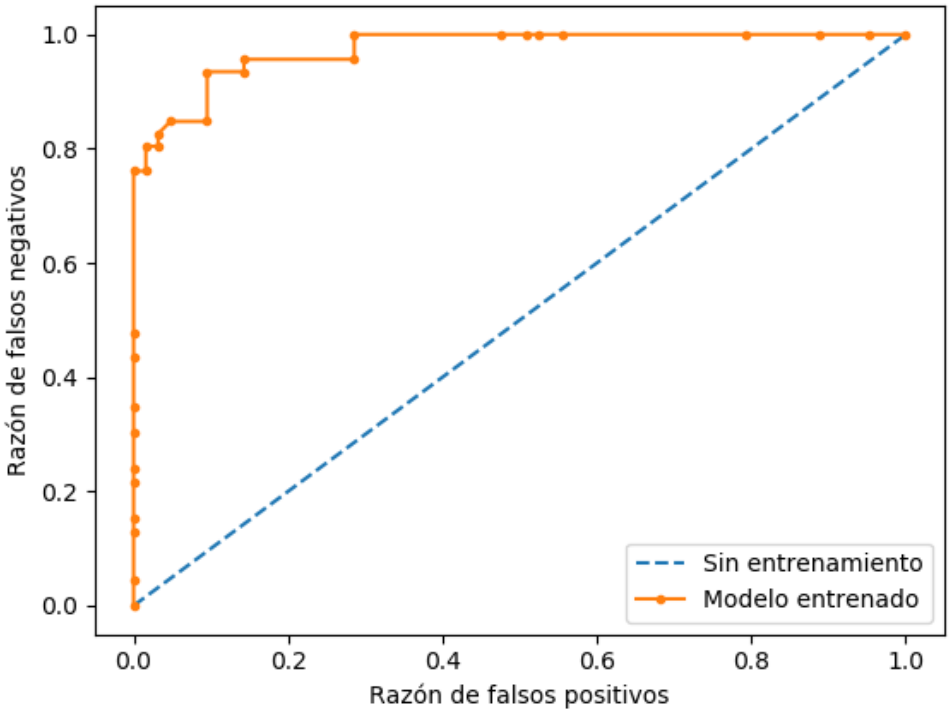
- Calcular las métricas de evaluación de ajuste adecuadas.

  Las métricas utilizadas fueron la matriz de confusión, con el siguiente resultado:

```
Matriz de confusión
Predicho   0   1
Real
0         55   8
1          3  43
1 --> Republicano, 0 --> Demócrata
```

Con estos datos se evalúa la precisión del modelo, resultando en un 0.8990825688073395 de precisión.

Para visualizar mejor la capacidad de clasificación, se construyó también una ROC:



Y con esta curva se calculó el área bajo la curva con un resultado de 0.969 en relación a datos positivos.

- Comentar los resultados obtenidos.

  El modelo, a pesar que presenta un porcentaje relativamente bajo de precisión (0.89) generaliza adecuadamente el problema ya que, si se considera el tratamiento de missing, puede llevar a que el algoritmo cometa errores frecuentemente al no conocer con exactitud los datos con los que debe ser

entrenado. El área bajo la curva muestra una gran habilidad del clasificador para discriminar entre ambas clases.

- Otros comentarios que parezcan adecuados.

  Se utilizó el clasificador Naïve Bayes de tipo multinomial, por encima del gaussiano, dado que se cuenta con una gran cantidad de categorías diferentes y sus interdependencias van a ser mejor tratadas por el algoritmo multinomial.

A continuación, se presenta una captura de los resultados del código y se adjunta el código en un archivo .py