Tao Izzo Elvira

Aprendizaje supervisado 1

El dataset que se usara en este proyecto es un dataset médico, que contiene datos para saber el estado del feto, este estado será la variable objetivo, en otras palabras, la variable a predecir con nuestro modelo. Para entrar un poco en contexto a continuación se describirán las variables de nuestros 2016 casos registrados:

- Id: identificador único del feto
- b: Hora inicial
- e: Hora final
- LBE: Frecuencia cardiaca del feto
- AC: Numero de aceleraciones por segundo
- FM: Numero de movimientos fetales por segundo
- UC: Numero de contracciones uterinas por segundo
- ASTV: Porcentaje de tiempo con variabilidad anormal a corto plazo
- MSTV: Valor medio de la variabilidad a corto plazo
- ALTV: Porcentaje de tiempo con variabilidad anormal a largo plazo
- MLTV: Valor medio de la variabilidad a largo plazo
- DL: Número de desaceleraciones de la luz por segundo
- DS: Número de desaceleraciones severas por segundo
- DP: Número de desaceleraciones prolongadas por segundo
- DR: Número de desaceleraciones repetitivas por segundo
- Anchura: Anchura del histograma FHR
- Min: Mínimo (baja frecuencia) del histograma de FCF
- Max: Máximo (frecuencia alta) del histograma de FCF
- Nmax: Número de picos del histograma
- Nzeros: Número de ceros del histograma ,
- Mode: Histograma moda
- Mean: Histograma media
- Median; Histograma mediana

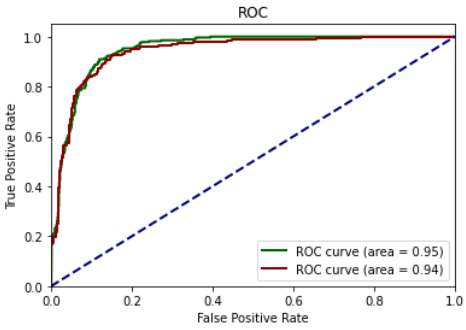
- Variance: Histograma varianza
- Tendency: Histograma tendencia
- Target: Código de clase de estado fetal

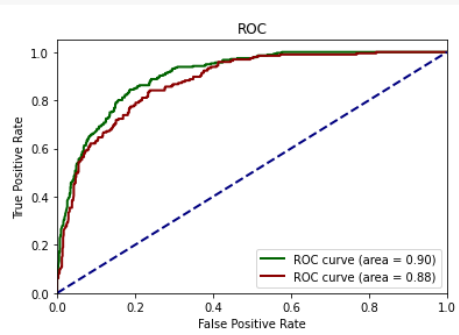
Para la creación del modelo, los datos se dividirán en conjuntos de entrenamiento y test. Al crear estos conjuntos lo que se hace es dividir nuestros datos en dos, el conjunto de entrenamiento se usa para ver si con nuestra elección de variables pueden predecir el conjunto de sí mismo, es decir, si con esos datos pueden saber el estado del feto, por otro lado, el conjunto de test sirve para ver cómo funciona nuestro modelo con datos que no se hayan sido usados anteriormente y es aquí donde sabremos realmente la eficiencia de nuestro modelo.

En los modelos hechos, la división de hace de manera aleatoria con el fin de que no haya un sesgo al elegir variables por ciertas características, a lo que se le conoce como muestra aleatoria.

Se seleccionaron dos tipos de algoritmos, Naive bayes y Support vector machine (SVM) para la creación de modelos, no obstante, con naive bayes se usaron dos subconjuntos el Gaussiano que corresponde a usar variables cuantitativas continuas en su mayoría y el Multinomial, el cual, por otro lado, usa variables discretas en su mayoría, teniendo así 3 diferentes modelos.

Cada uno de los modelos se entrenó con un test del 40% de nuestros datos totales, con respecto a que variables se usaron en cada uno esto varia debido a su tipo, con naive bayes gaussiano se hicieron 3 tipos de prueba y la que mejor funciono fue con ASTV, AC, LBE, UC, Min, Width, DP, DS es decir, usando cuantitativas continuas en su mayoría y 2 discretas dando un test de 0.95 y un train de 0.94. Combinar las variables igualmente funciono con la multinomial, aunque, los resultados fueron más bajos, test de 0.88 y un train de 0.90.

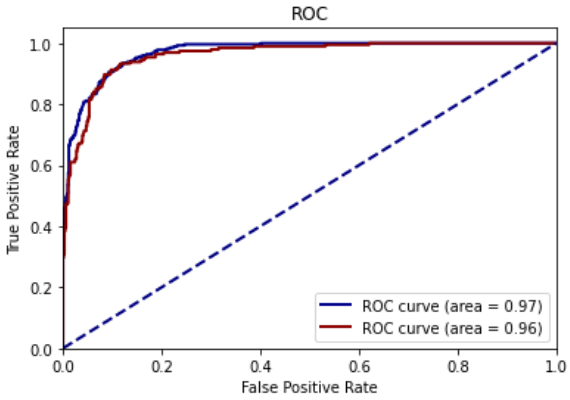




Con respecto al SVM aquí se hizo una combinación de hiperparámetros para obtener el más adecuado usando:

```
param_grid = [
    {"kernel": ["rbf"], "gamma": [1,0.1,0.5,0.01,0.005], "C": [0.1,0.5,1,5,10,50,100]},
    {"kernel": ["linear"], "C": [0.1,0.5,1,5,10,50,100]},
    {"kernel": ["poly"], "C": [0.1,0.5,1,5,10,50,100], "degree": [2,3]},
]
```

Los hiper parámetros con mejor resultado fueron C: 1, con kernel linear dando un test de 0.96 y un train de 0.97, es decir, el mejor modelo de todos los hechos en el proyecto



https://colab.research.google.com/drive/1DfKhrP2INvlOHaOdEL2ken1dYBQoZQaB?usp=sharing