**Tema 5**

**Task 4: Complete a Data Science Capstone Project**

**Francisco Paniagua Barrantes**

**Contexto de la situación**

La unión médica de Wisconsin a recopilado un set de datos de imágenes de tumores en los senos de las mujeres para estudiar y buscar detectar si esos tumores son benignos o malignas, usualmente esto se realiza con el análisis de varios médicos. Dada la situación actual del mundo y de Wisconsin con respecto al Coronavirus COVID-19, se requiere todo el personal médico que sea posible, por lo que esta tarea de diferenciación de si las masas son positivas o negativas haría mas eficiente el proceso y se podrían utilizar recursos para la emergencia. La unión médica necesita conocer cuales variables son importantes, si es posible realizar pronósticos y con qué exactitud.

**Preguntas**

1. ¿Cuáles de las variables proporcionadas son realmente útiles?
2. ¿Se puede realizar un pronóstico?
3. ¿Con que exactitud se puede confiar en la respuesta?

**Hallazgos**

Al revisar el set de datos y sus variables se determinó, que si es posible realizar un pronóstico. En el set de datos se observan una gran cantidad de variables, pero para generar un pronóstico más acertado se descartaron algunas de ellas, específicamente : id, fractal\_dimension\_mean, fractal\_dimension\_se, fractal\_dimension\_worst, area\_mean, area\_se, area\_worst, smoothness\_mean, symmetry\_mean, texture\_se, smoothness\_se, compactness\_se,concave points\_se, symmetry\_se y compactness\_worst.

Es importante aclarar que las variables area\_mean, area\_se, area\_worst se descartaron porque, al normalizar los datos, lejos de aportar al modelo, iban a generar overfitting o resultados erróneos.

Las variables seleccionadas como importantes para el pronóstico de tumores malignos o benignos en el pecho de las mujeres se muestran en el Gráfico 1 Variables de importancia. Cabe recalcar que la cantidad de variables se redujo en un 50%, pasando a definir 16 las variables en importancia, por lo que el tiempo de toma de datos por la parte médica se puede reducir en valores cercanos a la mitad.

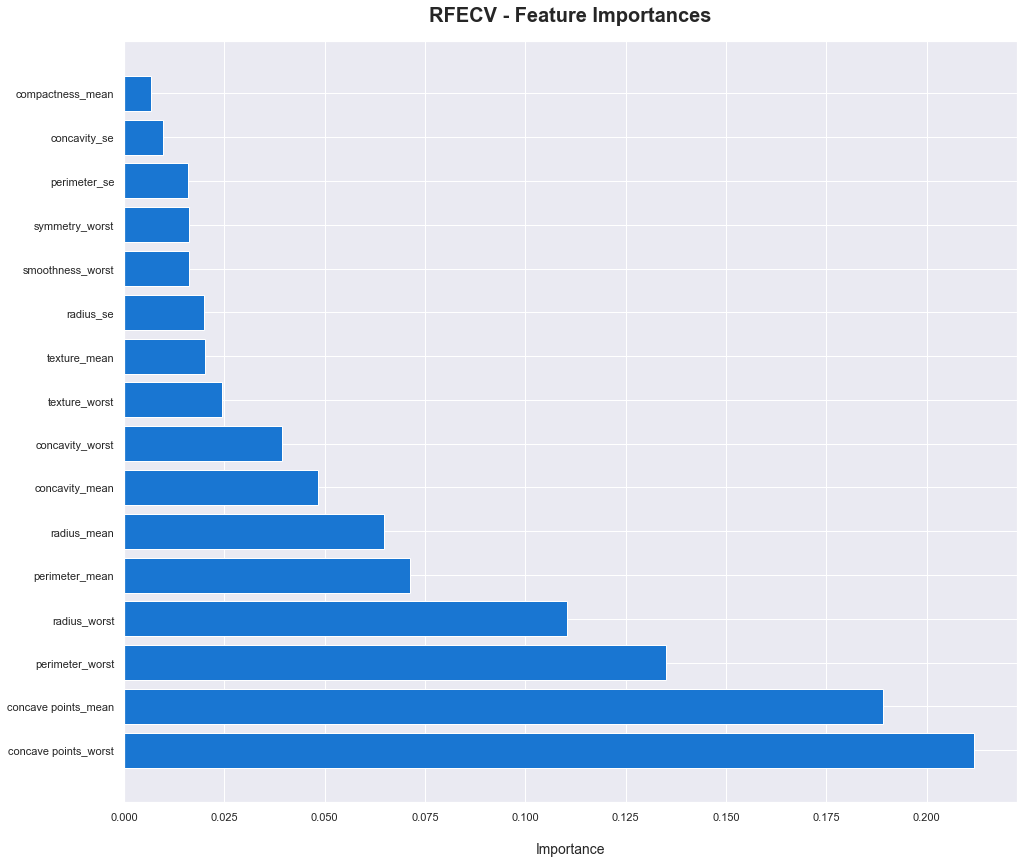


Gráfico 1. Variables de importancia

Al aplicar diferentes modelos de aprendizaje automático se obtuvieron resultados muy satisfactorios, esto por que se obtuvo una exactitud en todos los modelos, mayor al 96% como se observa en el Cuadro 1.

Es importante tener en cuenta que el indicador más importante es recall, por el hecho de que es preferible tener un falso positivo que un falso negativo cuando se trata de un pronóstico de un tumor maligno.

Por lo anterior, se buscó que recall sea lo más cercano a 1 posible, después de este indicador el segundo en importancia es accuracy, que también se buscó que este lo más cercano a 1 que sea posible.

Por lo anterior, pese a que Gradient Boosting Classifier presentó un mejor accuracy y un recall general con un valor más alto, el recall del Random Forest para los tumores malignos verdaderos fue de 0.97, siendo el mayor de los cuatro modelos, por este motivo se selecciona RF como el modelo seleccionado.



Cuadro 1. Métricas obtenidas

**Recomendaciones**

Se recomienda desarrollar un poco mas la normalización de los datos, con el fin de verificar si manteniendo las variables de area\_mean, area\_se, area\_worst se logra obtener un mejor resultado.

Se recomienda no realizar la normalización de los datos con la función logaritmo o con BoxCox, ya que se indefinen valores y se presentan resultados erróneos e inconsistentes.