Informe Brest Cancer Model

Origen de los datos

Nuestro dataset de Breast Cancer Wisconsin está recogido de Kaggle, en el siguiente enlace podrás acceder a él.

Breas Cancer Winsconsin (Diagnostic)

Problemas de negocio

Descripción del problema

El cáncer de mama es una de las principales causas de mortalidad entre las mujeres a nivel mundial. La detección temprana del cáncer es crucial para que las probabilidades de supervivencia del paciente se vean incrementadas y también reducir los altos costos del tratamiento.

Aún detectado el tumor a tiempo, la intervención médica se prolonga hasta determinar si es benigno o maligno, debido principalmente a la interpretación de las imágenes de las mamografías de los radiólogos.

Impacto en el Negocio

Los principales desafíos a los que se enfrentan las instituciones médicas en la detección tardía del cáncer de mama son los siguientes:

- **Costos de tratamiento elevado**
- **Tasa de supervivencia reducida**
- **Errores de diagnóstico**

En nuestro proyecto nos centramos en eliminar los errores de diagnóstico de los médicos que, indirectamente, causan atrasos que conllevan a la detección tardía del cáncer de mama, aumentando los costos del tratamiento y haciendo más difícil la supervivencia de este.

Aplicación de vuestra solución

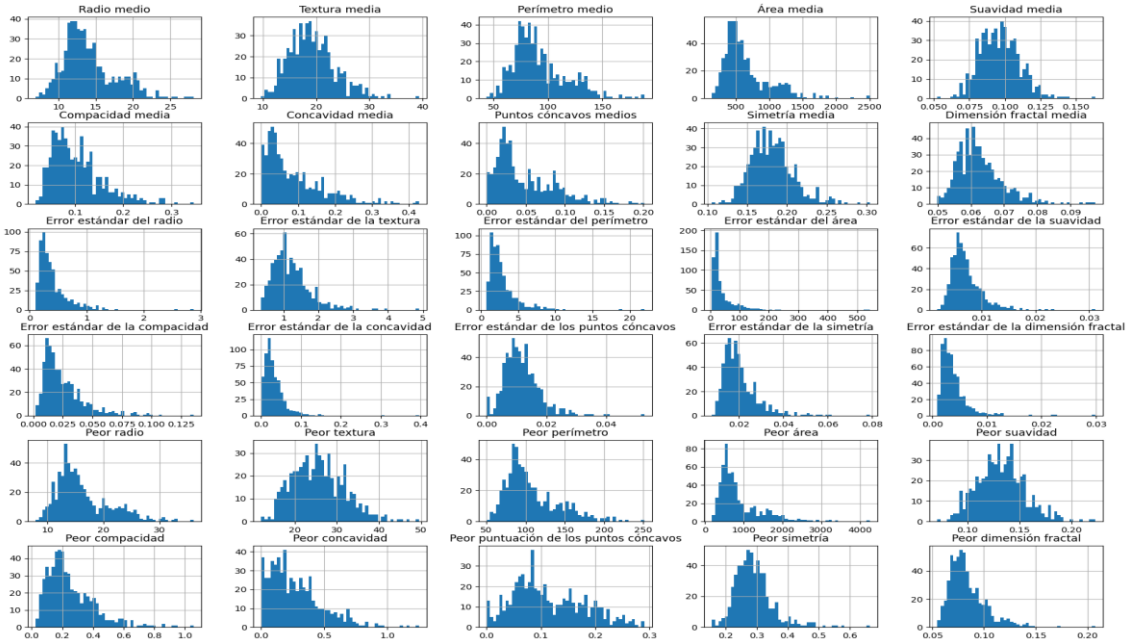
Para lograr nuestro objetivo hemos entrenado modelos de clasificación para ayudar a los médicos a minimizar el error y a una rápida detección.

Descripción de las variables

Nuestro dataset se compone de 30 columnas de datos reales a la hora de aplicarse en la detección de cáncer de mama, las cuales todas ellas contienten valores de tipo Float, y una columna target que representa si el túmor analizado es benigno o maligno, siendo esta la única columna con valores objeto.

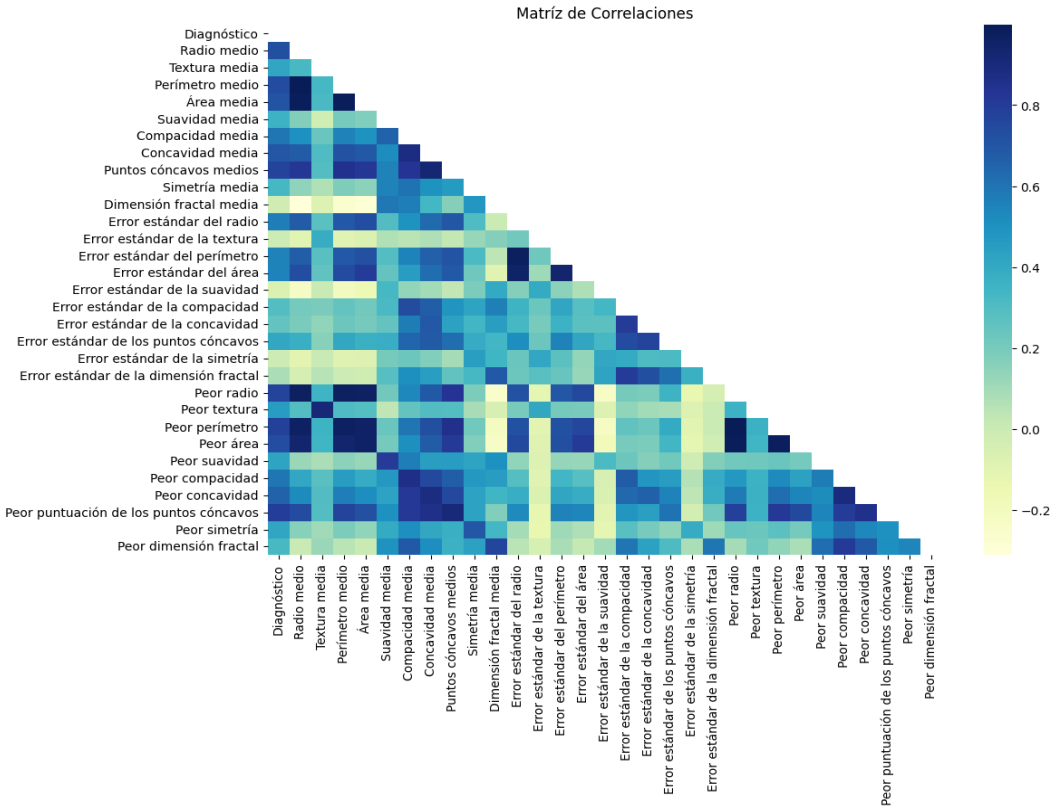
Feature engineering

Distribución de variables



Como se puede observar en la imagen, aunque algunos datos si se encuentran balanceados, existe cierta tendencia a la asimetría Negativa.

Correlaciones



Estudiando la gráfica, podemos apreciar como existe un amplio número de correlaciones entre las diferentes columnas, no solo se centra respecto a la columna Diagnóstico, sino que también entre las otras columnas existe diversas correlaciones.

Reducción de dimensionalidad

Tras hacer diversas pruebas, hemos observado que la reducción de dimensionalidad no es viable en nuestro caso, puesto que al no tener demasiadas dimensiones con las que trabaja el modelo, realizar, comprometa la información, maximizando la perdida de información y aumentando el bias del modelo, lo que puede desembocar en un sobreajuste del modelo.

Tras esa valoración, decidimos no utilizar ninguna técnia de reducción de dimensionalidad.

Balanceo de datos

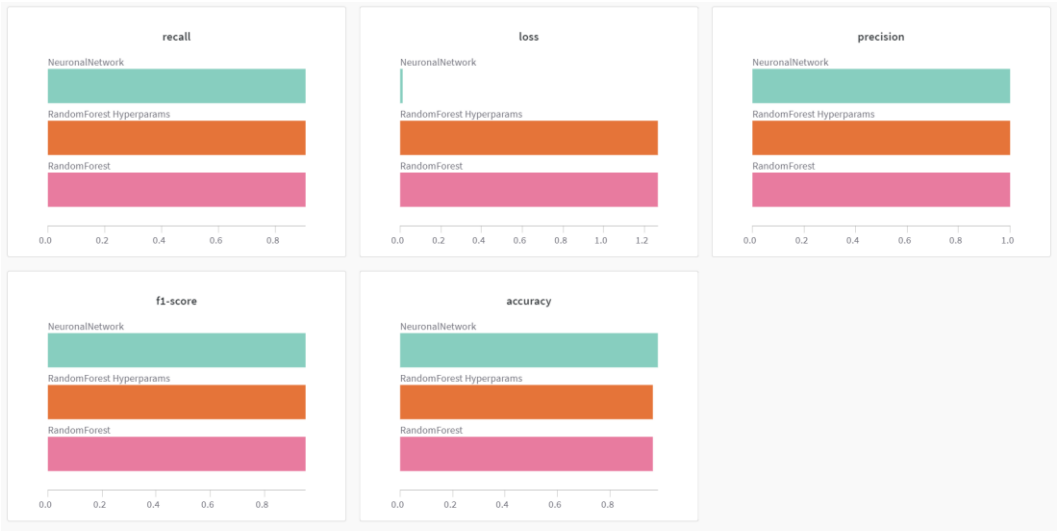
Como los datos que manejamos son sensibles y críticos, hemos decido no comprometer la información y evitar agregar ruido que pueda perjudicar el desempeño del modelo.

Transformación

El dataset sobre el que estamos trabajando trae ya datos limpios y dispuestos a trabajar con ellos directamente, salvo por la columna Diagnóstico, por lo que la única transformación que hemos hecho a sido mappear esta columna sustituyendo los valores por 0 o 1, siendo 1 si el tumor es Maligno.

Modelos

Como podemos ver en la tabla imagen inferior, obtenida desde Weights & Biases, las diferencias entre los modelos son mínimas, lo que mejor se puede apreciar es la diferencia de *loss* entre los modelos.



RandomForest		RandomForest Optim		NeuronalNetwork	
Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
0.9766	0.9531	1	0.9063	0.9839	0.9531
Accuracy	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**
0.9649	0.9508	0.9766	0.9508	0.9825	0.96825

El modelo que mejor resultados nos ha arrojado es la Red Neuronal, por lo que será el que utilicemos para las predicciones.