**Aplicación de la metodología CRISP-DM para un modelo de clasificación según el tipo de tumor de un paciente**

**(maligno o benigno)**

**INTEGRANTES:**

Tomás Cardona Montoya

Camilo Velásquez Hincapié

Andrés Felipe Arias Medina

**DOCENTE:**

Ana Isabel Oviedo Carrascal

**CURSO:**

Analítica de Datos Estructurados (Minería)



**Universidad Pontificia Bolivariana**

Facultad de Ingeniería

Ingeniería en Ciencia de Datos

Medellín, Colombia

Mayo de 2023

1. **ENTENDIMIENTO DEL NEGOCIO**

El propósito de esta etapa es identificar y definir los objetivos y requisitos del proyecto desde una perspectiva empresarial, con el fin de transformarlos posteriormente en metas técnicas y elaborar un plan de proyecto.

* 1. DESCRIPCIÓN DEL NEGOCIO

Este modelo puede servir para ser una herramienta en el campo de la salud, en donde se pueda establecer un mayor control en las enfermedades cancerígenas de las personas, dando un diagnóstico respaldado por un profesional en el tema más la validación del proyecto a realizar, por lo que se automatizan procesos y ayudaría la conservación de la salud en las personas al tener una detección temprana de esta enfermedad.

* 1. DESCRIPCIÓN DEL PROBLEMA

La enfermedad del cáncer en los últimos tiempos es una de las mayores causas de muerte en la sociedad, es por esto que es importante tener un pronto diagnóstico de personas que se le sospecha por múltiples síntomas la presencia de tumores que puedan desencadenar en esta enfermedad, pero en múltiples ocasiones los tiempos de espera para exámenes médicos y dar respuesta a los pacientes son de larga duración y además estos no suelen tener una respuesta certera acerca del tipo de tumor que presentan, por esto es necesaria la realización del proyecto para un respaldo médico a los profesionales encargados del tema y así mejorar el sistema con el cual se maneje a este tipo de enfermedades.

* 1. OBJETIVOS DE LA MINERÍA
* Realizar un modelo que permita predecir qué tipo de pacientes presentan un tumor cancerígeno.
* Investigar cuales son las características con mayor relevancia en los pacientes que presentan un tumor cancerígeno.
  1. DISEÑO DE SOLUCIÓN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PROBLEMA** | **TIPO DE MINERÍA** | **TAREA ANALÍTICA** | **MÉTODOS** | **MEDIDAS DE EVALUACIÓN** |
| Detección de cáncer mediante la clasificación del tumor (benigno o maligno) de acuerdo con las características de este. | Predictiva | Clasificación (si un paciente tiene o no cáncer a partir de si el tumor es benigno o maligno | Árboles de clasificación, KNN, redes neuronales, regresión logística, máquinas de soporte vectorial. | Matriz de confusión, precisión, recall, F1 score |

* 1. RECURSOS HARDWARE Y SOFTWARE

Google Colaboratory

1. **ENTENDIMIENTO DE LOS DATOS**

Este paso requiere el estudio del origen de los datos, es comprender cómo funciona la estructura que propicia esos datos. Todo con la finalidad de dar una correcta interpretación a los datos, entender su calidad y el modo de acción correcto para su tratamiento.

* 1. CICLO DE LOS DATOS (GENERACIÓN, ALMACENAMIENTO, MODIFICACIÓN Y PERIODICIDAD DE LOS DATOS)

En la página de Kaggle la única especificación que se tiene sobre los datos es que la frecuencia que se usa para actualizar los datos es de un año. Por lo que se comprende de los datos es que les hicieron seguimiento a los pacientes durante determinado tiempo, posiblemente entre 6 meses o 1 año, tomando mediciones periódicas de cada mes o semana de las diferentes características de un tumor que en total serían 10, después se sacaron tres medidas de cada una de estas variables que son la medía, el error estándar y el valor máximo, para así determinar la clasificación del tumor como benigno o maligno.

* 1. DICCIONARIO DE DATOS

|  |  |  |
| --- | --- | --- |
| **VARIABLE** | **DESCRIPCIÓN** | **TIPO DE DATO** |
| id | Es el id para cada paciente de los que se registran datos | Numérico |
| diagnosis | Es la clasificación del estado del cáncer del paciente, B para benigno y M para maligno | categórico |
| radius mean | Media del radio que puede ocupar o afectar el tumor cancerígeno. | Numérico |
| texture\_mean | La media de la homogeneidad o heterogeneidad de la distribución de las características dentro del tumor | Numérico |
| perimeter\_mean | Media del perímetro en el cual se puede encontrar o afectar el tumor cancerígeno | Numérico |
| area\_mean | Media del área en el que el tumor cancerígeno se encuentra | Numérico |
| smoothness\_mean | Es la media de suavidad que tiene en la textura el tumor cancerígeno | Numérico |
| compactness\_mean | Media de la compacidad del tumor cancerígeno en esta influyen el perímetro y el área | Numérico |
| concavity\_mean | La media de concavidad define la irregularidad en el contorno del tumor cancerígeno | Numérico |
| concave points\_mean | Media de los puntos en el contorno del tumor que presenta una curvatura hacía adentro | Numérico |
| symmetry\_mean | Media de la simetría del tumor, la cual describe qué tan uniforme o equilibrada es la forma del tumor en relación con un eje o plano determinado. | Numérico |
| fractal\_dimension\_mean | Media de la dimensión fractal del tumor, puede proporcionar información sobre la textura y la forma del tumor. | Numérico |
| radius\_se | Es el error estándar del radio que puede ocupar o afectar el tumor cancerígeno. | Numérico |
| texture\_se | Es el error estándar de homogeneidad o heterogeneidad de la distribución de las características dentro del tumor | Numérico |
| perimeter\_se | Es el error estándar del perímetro en el cual se puede encontrar o afectar el tumor cancerígeno | Numérico |
| area\_se | Es el error estándar del área en el que el tumor cancerígeno se encuentra | Numérico |
| smoothness\_se | Es el error estándar de suavidad que tiene en la textura el tumor cancerígeno | Numérico |
| compactness\_se | Es el error estándar de la compacidad del tumor cancerígeno en esta influyen el perímetro y el área | Numérico |
| concavity\_se | Es el error estándar de concavidad define la irregularidad en el contorno del tumor cancerígeno | Numérico |
| concave points\_se | Es el error estándar de los puntos en el contorno del tumor que presenta una curvatura hacía adentro | Numérico |
| symmetry\_se | Es el error estándar de la simetría del tumor, la cual describe qué tan uniforme o equilibrada es la forma del tumor en relación con un eje o plano determinado. | Numérico |
| fractal\_dimension\_se | Es el error estándar de la dimensión fractal del tumor, puede proporcionar información sobre la textura y la forma del tumor. | Numérico |
| radius\_worst | Se refiere al valor máximo del radio que puede ocupar o afectar el tumor cancerígeno. | Numérico |
| texture\_worst | Se refiere al valor máximo de homogeneidad o heterogeneidad de la distribución de las características dentro del tumor | Numérico |
| perimeter\_worst | Se refiere al valor máximo del perímetro en el cual se puede encontrar o afectar el tumor cancerígeno | Numérico |
| area\_worst | Se refiere al valor máximo del área en el que el tumor cancerígeno se encuentra | Numérico |
| smoothness\_worst | Se refiere al valor máximo de suavidad que tiene en la textura el tumor cancerígeno | Numérico |
| compactness\_worst | Se refiere al valor máximo de la compacidad del tumor cancerígeno en esta influyen el perímetro y el área | Numérico |
| concavity\_worst | Se refiere al valor máximo de concavidad define la irregularidad en el contorno del tumor cancerígeno | Numérico |
| concave points\_worst | Se refiere al valor máximo de los puntos en el contorno del tumor que presenta una curvatura hacía adentro | Numérico |
| symmetry\_worst | Se refiere al valor máximo de la simetría del tumor, la cual describe qué tan uniforme o equilibrada es la forma del tumor en relación con un eje o plano determinado. | Numérico |
| fractal\_dimension\_worst | Se refiere al valor máximo de la dimensión fractal del tumor, puede proporcionar información sobre la textura y la forma del tumor. | Numérico |

* 1. REGLAS DE CALIDAD

|  |  |
| --- | --- |
| **VARIABLE** | **REGLA DE CALIDAD** |
| id | No tiene, sólo son id de las personas |
| diagnosis | Solo pueden tomar la clasificación B para benigno o M para maligno |
| radius\_mean | Si el valor es inferior a 10, se considera de bajo riesgo con una alta.  Si el valor está entre 10 y 15, se considera de riesgo moderado.  Si el valor es mayor que 15, se considera de alto riesgo |
| texture\_mean | Si el valor es inferior a 12, se considera baja textura.  Si el valor está entre 12 y 18, se considera textura moderada.  Si el valor es mayor que 18, se considera alta textura |
| perimeter\_mean | Si el valor es inferior a 80, se considera un perímetro pequeño.  Si el valor está entre 80 y 120, se considera un perímetro mediano.  Si el valor es mayor que 120, se considera un perímetro grande |
| area\_mean | Si el valor es inferior a 500, se considera un área pequeña.  Si el valor está entre 500 y 1000, se considera un área mediana.  Si el valor es mayor que 1000, se considera un área grande |
| smoothness\_mean | Si el valor es inferior a 0.05, se considera una textura suave.  Si el valor está entre 0.05 y 0.1, se considera una textura moderadamente suave.  Si el valor es mayor que 0.1, se considera una textura rugosa |
| compactness\_mean | Si el valor es inferior a 0.05, se considera una compactación baja.  Si el valor está entre 0.05 y 0.1, se considera una compactación moderada.  Si el valor es mayor que 0.1, se considera una compactación alta |
| concavity\_mean | Si el valor es igual a 0, se considera una concavidad nula.  Si el valor está entre 0 y 0.1, se considera una concavidad baja.  Si el valor es mayor que 0.1, se considera una concavidad alta |
| concave points\_mean | Si el valor es igual a 0, se considera un número nulo de puntos cóncavos.  Si el valor está entre 0 y 0.03, se considera un número bajo de puntos cóncavos.  Si el valor es mayor que 0.03, se considera un número alto de puntos cóncavos |
| symmetry\_mean | Si el valor es inferior a 0.1, se considera una simetría baja.  Si el valor está entre 0.1 y 0.2, se considera una simetría moderada.  Si el valor es mayor que 0.2, se considera una simetría alta |
| fractal\_dimension\_mean | Si el valor es inferior a 0.05, se considera una dimensión fractal baja.  Si el valor está entre 0.05 y 0.08, se considera una dimensión fractal moderada.  Si el valor es mayor que 0.08, se considera una dimensión fractal alta |
| radius\_se | Si el valor es inferior a 0.3, se considera un radio estándar bajo  Si el valor está entre 0.3 y 0.6, se considera un radio estándar moderado.  Si el valor es mayor que 0.6, se considera un radio estándar alto |
| texture\_se | Si el valor es inferior a 1, se considera una textura uniforme.  Si el valor está entre 1 y 2, se considera una textura moderadamente variada.  Si el valor es mayor que 2, se considera una textura altamente variada. |
| perimeter\_se | Si el valor es inferior a 4, se considera un perímetro estándar bajo.  Si el valor está entre 4 y 6, se considera un perímetro estándar moderado.  Si el valor es mayor que 6, se considera un perímetro estándar alto. |
| area\_se | Si el valor es inferior a 50, se considera un área estándar muy baja.  Si el valor está entre 50 y 100, se considera un área estándar baja.  Si el valor está entre 100 y 200, se considera un área estándar moderada.  Si el valor está entre 200 y 300, se considera un área estándar alta.  Si el valor está entre 300 y 400, se considera un área estándar alta.  Si el valor está entre 400 y 500, se considera un área estándar muy alta.  Si el valor es mayor que 500, se considera un outlier. |
| smoothness\_se | Si el valor es inferior a 0.01, se considera una suavidad.  Si el valor está entre 0.01 y 0.02, se considera una suavidad estándar moderada.  Si el valor" es mayor que 0.02, se considera una suavidad estándar alta |
| compactness\_se | Si el valor es inferior a 0.02, se considera una compactación estándar baja.  Si el valor está entre 0.02 y 0.05, se considera una compactación estándar moderada.  Si el valor es mayor que 0.05, se considera una compactación estándar alta. |
| concavity\_se | Si el valor es de cero, se considera que la concavidad es nula.  Si el valor es inferior a 0.01, se considera una concavidad estándar baja.  Si el valor está entre 0.01 y 0.03, se considera una concavidad estándar moderada.  Si el valor es mayor que 0.03, se considera una concavidad estándar alta |
| concave points\_se | Si el valor es de cero, se considera que los puntos de concavidad son nulos.  Si el valor es inferior a 0.003, se considera un número de puntos cóncavos estándar bajo.  Si el valor de está entre 0.003 y 0.008, se considera un número de puntos cóncavos estándar moderado.  Si el valor es mayor que 0.008, se considera un número de puntos cóncavos estándar alto |
| symmetry\_se | Si el valor es inferior a 0.015, se considera una simetría estándar baja.  Si el valor está entre 0.015 y 0.03, se considera una simetría estándar moderada.  Si el valor es mayor que 0.03, se considera una simetría estándar alta. |
| fractal\_dimension\_se | Si el valor es inferior a 0.0035, se considera una dimensión fractal estándar baja.  Si el valor está entre 0.0035 y 0.006, se considera una dimensión fractal estándar moderada.  Si el valor es mayor que 0.006, se considera una dimensión fractal estándar alta. |
| radius\_worst | Si el valor es inferior a 10, se considera un radio máximo estándar bajo.  Si el valor está entre 10 y 20, se considera un radio máximo estándar moderado.  Si el valor es mayor que 20, se considera un radio máximo estándar alto |
| texture\_worst | Si el valor es inferior a 15, se considera una textura máxima estándar baja.  Si el valor está entre 15 y 25, se considera una textura máxima estándar moderada  Si el valor es mayor que 25, se considera una textura máxima estándar alta |
| perimeter\_worst | Si el valor es inferior a 90, se considera un perímetro máximo estándar bajo.  Si el valor está entre 90 y 150, se considera un perímetro máximo estándar moderado.  Si el valor es mayor que 150, se considera un perímetro máximo estándar alto. |
| area\_worst | Si el valor es inferior a 400, se considera un área máxima estándar muy baja .  Si el valor está entre 400 y 800, se considera un área máxima estándar baja .  Si el valor está entre 800 y 1200, se considera un área máxima estándar moderadamente baja.  Si el valor está entre 1200 y 1600, se considera un área máxima estándar moderada.  Si el valor está entre 1600 y 2000, se considera un área máxima estándar moderadamente alta.  Si el valor está entre 2000 y 2400, se considera un área máxima estándar alta  Si el valor está entre 2400 y 2800, se considera un área máxima estándar muy alta  Si el valor es mayor que 2800, se considera un área máxima estándar extremadamente alta |
| smoothness\_worst | Si el valor es inferior a 0.08, se considera una suavidad máxima estándar baja.  Si el valor está entre 0.08 y 0.12, se considera una suavidad máxima estándar moderada.  Si el valor es mayor que 0.12, se considera una suavidad máxima estándar alta. |
| compactness\_worst | Si el valor es inferior a 0.15, se considera una compacidad máxima estándar baja.  Si el valor está entre 0.15 y 0.3, se considera una compacidad máxima estándar moderada  Si el valor es mayor que 0.3, se considera una compacidad máxima estándar alta |
| concavity\_worst | Si el valor es de cero, se considera que la concavidad es nula.  Si el valor es inferior a 0.05, se considera una concavidad máxima estándar baja.  Si el valor está entre 0.05 y 0.2, se considera una concavidad máxima estándar moderada.  Si el valor es mayor que 0.2, se considera una concavidad máxima estándar alta. |
| concave points\_worst | Si el valor es de cero, se considera que los puntos de concavidad son nulos.  Si el valor es inferior a 0.025, se considera un número de puntos cóncavos máximo estándar bajo.  Si el valor está entre 0.025 y 0.05, se considera un número de puntos cóncavos máximo estándar moderado.  Si el valor es mayor que 0.05, se considera un número de puntos cóncavos máximo estándar alto. |
| symmetry\_worst | Si el valor es inferior a 0.1, se considera una simetría máxima estándar muy baja con alta .  Si el valor está entre 0.1 y 0.2, se considera una simetría máxima estándar baja.  Si el valor está entre 0.2 y 0.4, se considera una simetría máxima estándar moderada.  Si el valor está entre 0.4 y 0.5, se considera una simetría máxima estándar moderadamente alta.  Si el valor es mayor que 0.5, se considera una simetría máxima estándar muy alta. |
| fractal\_dimension\_worst | Si el valor de es inferior a 0.05, se considera una dimensión fractal máxima estándar baja.  Si el valor está entre 0.05 y 0.1, se considera una dimensión fractal máxima estándar moderado.  Si el valor es mayor que 0.1, se considera una dimensión fractal máxima estándar alta. |

1. **PREPARACIÓN DE DATOS**

En esta fase se opta por adecuar el conjunto de datos para ponerse a disposición de los diferentes métodos de modelado. Para ello, se realizará una limpieza de los datos para mejorar su calidad, se añadirán nuevos datos en caso de ser necesario para lograr un balance general y se aplicarán ciertas técnicas en los datos para una adaptación previa al modelado.

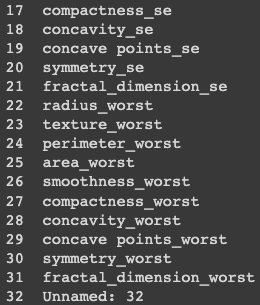
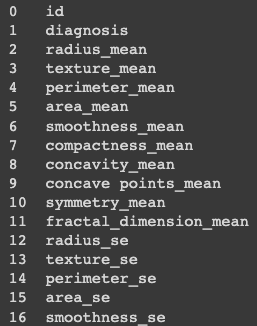
* 1. INTEGRACIÓN

No ha sido necesario la integración de los datos en este caso puesto que los datos están estructurados por columnas en un único archivo csv. La integración de datos suele ser necesaria cuando se tienen múltiples fuentes de datos que deben combinarse o cuando los datos se encuentran en diferentes formatos o estructuras. En esta ocasión, los datos ya se encuentran consolidados y categorizados por características propias de cada tumor en formato numérico.

* 1. SELECCIÓN DE VARIABLES

Para este proceso se seleccionaron aquellas variables que aportan información técnica expresada en formato numérico sobre características específicas de los tumores, de resto se suprimieron variables del formato original de los datos que no proporcionaban información alguna para determinar el tipo de tumor en cuestión.

Las variables originales entregadas por el conjunto de datos inicial son las siguientes:



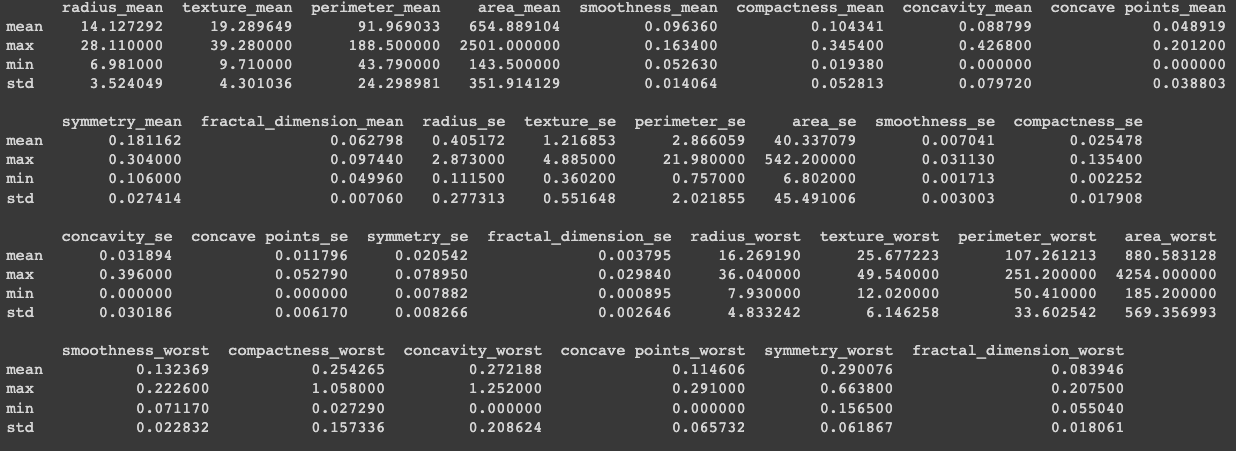
Durante la interpretación de estas columnas se decretó la eliminación de la columna “id” por tratarse de un identificador de registro sin aportación en la predicción y de la columna “Unnamed” por ser una columna de nulos.

* 1. DESCRIPCIÓN ESTADÍSTICA

El conjunto de datos inicial consta de 33 columnas con un total de 568 registros.

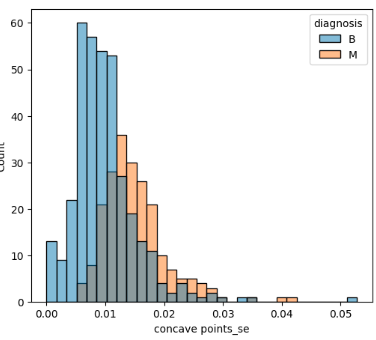
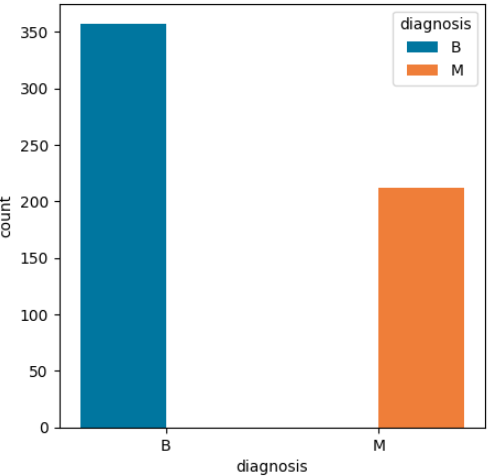
Al desplegar las funciones estadísticas se muestran las medidas más relevantes de las columnas, que al analizarlas sin mucho detalle arrojan valores que no entran en la escala normal en términos médicos. Por lo tanto, desde este punto comienzan las sospechas de posibles datos atípicos en las diferentes columnas, pero a partir del entendimiento del negocio se puede asumir su posibilidad por la variabilidad de casos médicos posibles.

Los datos descritos son los siguientes:



Para este caso, sirvieron principalmente medidas como la media aritmética, la desviación estándar, el máximo y el mínimo. Con ellas se armó un panorama inicial para determinar un tumor “común” y con ello verificar posibles registros referentes a tumores particulares o fuera de lo común, verificando los datos con las tablas iniciales del entendimiento del negocio.

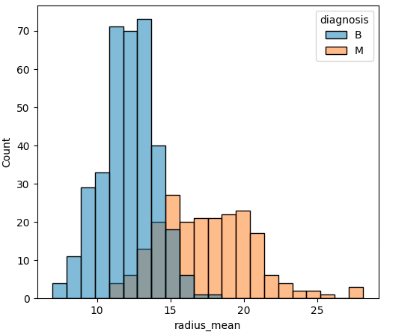
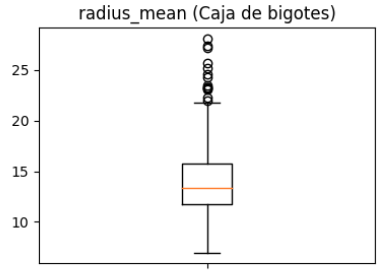
Adicionalmente se graficaron las diferentes columnas de acuerdo con el tipo de dato que la compone. Para las columnas numéricas se empleó el histograma y la caja de bigotes (boxplot) y en el caso de la única columna categórica que usó el diagrama de barras tradicional.

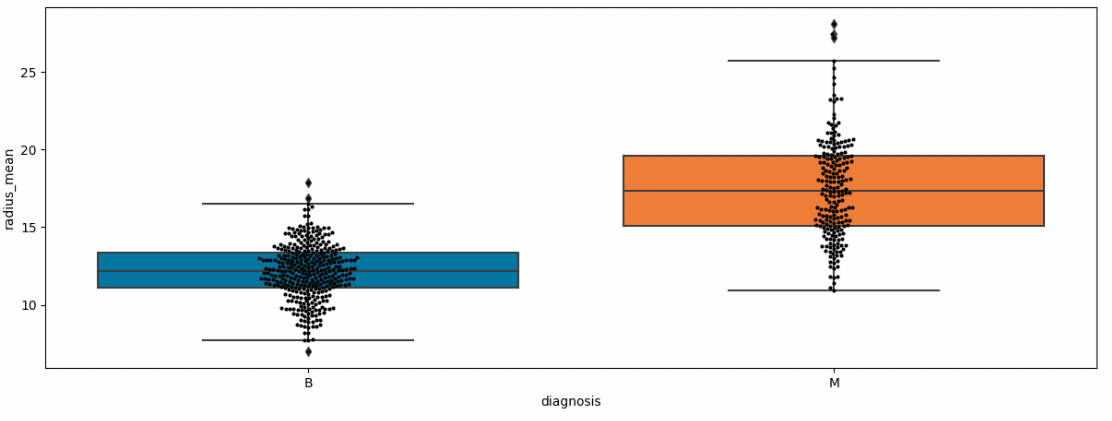


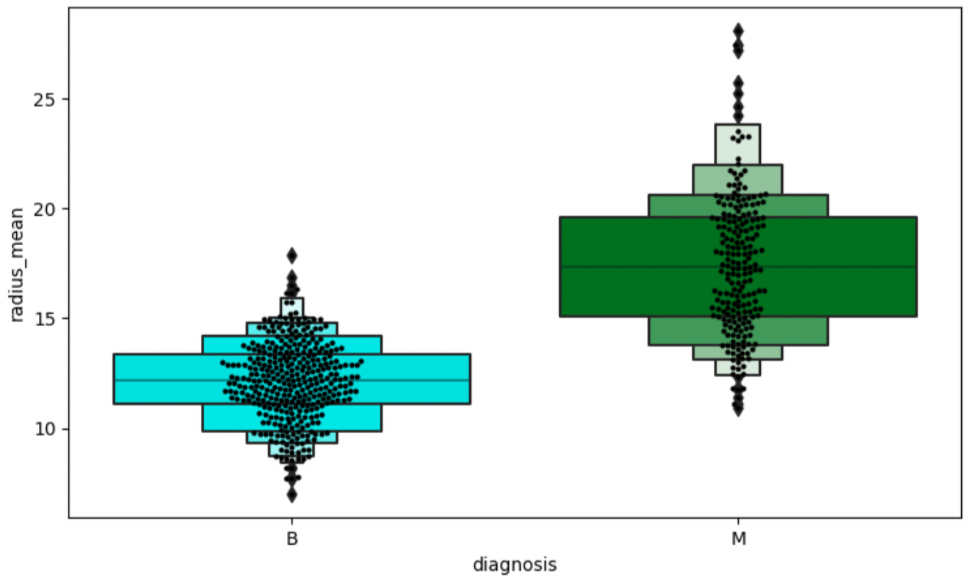
**Nota:** Cada columna tiene sus propias gráficas dependiendo del tipo de dato.

* 1. LIMPIEZA DE ATÍPICOS

Al realizarse una esquematización general de cada columna se observó presencia de atípicos en prácticamente todas las columnas. Se hizo uso de varias librerías para observar sustancialmente cada columna en el proceso de selección de atípicos a suprimir. Se interpretó un histograma inicial para visualizar la distribución de los datos y posteriormente se graficó un caja de bigotes (boxplot) con algunas variedades visuales proporcionadas por la librería seaborn.



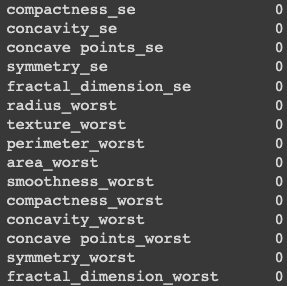
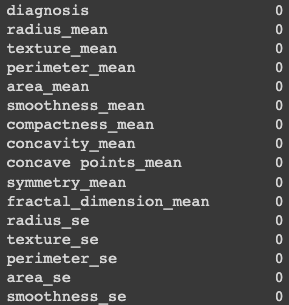




Finalmente, no se realizó ninguna eliminación de datos por conclusiones sobre el entendimiento de los datos. Se precisó la posibilidad de datos fuera del estándar al tratarse de temas médicos impredecibles, ya que no existe un límite en el desarrollo de estos tumores en las personas.

* 1. LIMPIEZA DE NULOS

El archivo por defecto no traía campos nulos en ninguna columna, por ende no fue necesario la eliminación de registros o aplicación de imputaciones con la mediana. Esto fue verificado con la función isna().sum() de Python.

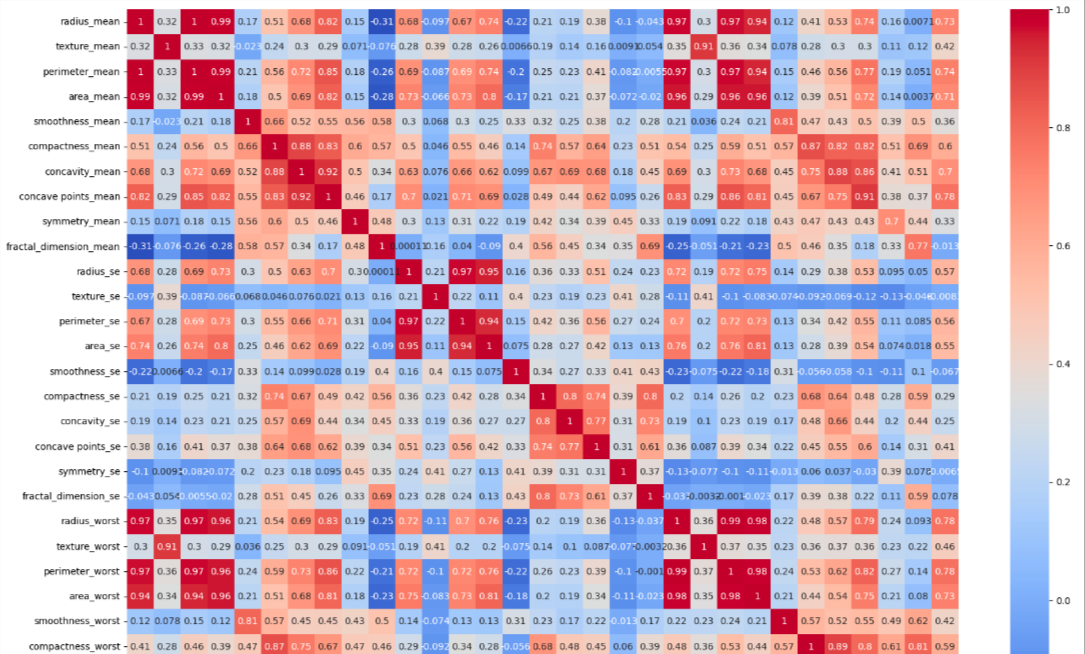


* 1. CREACIÓN DE NUEVAS VARIABLES

No se contempló la posibilidad de incluir nuevas variables dado que el entendimiento de los datos arrojó un significado médico para cada una de las columnas, por lo tanto, era un riesgo innecesario la unificación o creación de variables por no poseer los suficientes conocimientos en el área médica. Esto probablemente hubiese conducido a la pérdida de información.

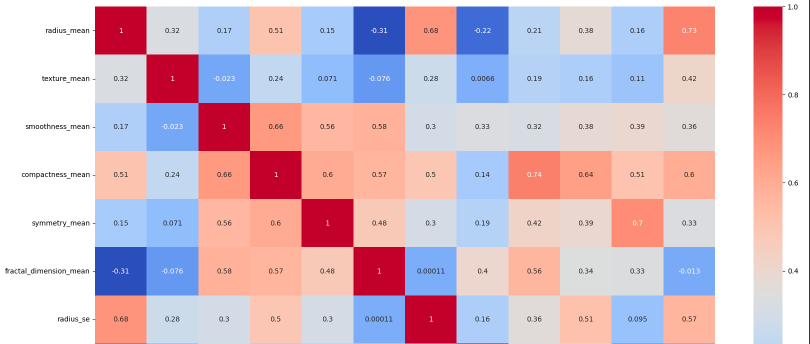
* 1. ANÁLISIS DE CORRELACIONES PARA REDUNDANCIA

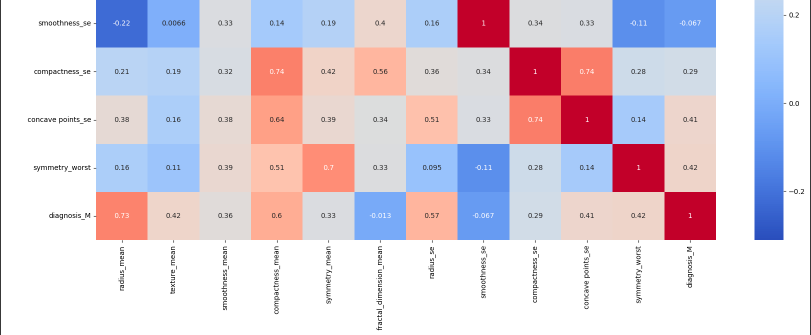
Para esta sección se procedió con la creación de dummies para las variables categóricas, esta vez solo se tenía una variable con estas características y precisamente era la variable objetivo.

Posteriormente, se graficó con funciones de Python la matriz de correlaciones: 

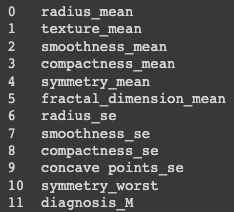
Al ser tantas columnas en el conjunto de datos, la matriz quedó particularmente grande y con dificultades de lectura e interpretación.

Ahora, bajo el criterio del 0.80, se procede a eliminar aquellas variables que estén correlacionadas entre sí con una correlación igual o mayor a 0.80. Para ello, se creó una función que buscara en la matriz de correlación anterior aquellas variables con esta característica para luego eliminarla. La matriz de correlación quedó así tras haber aplicado la función de supresión:



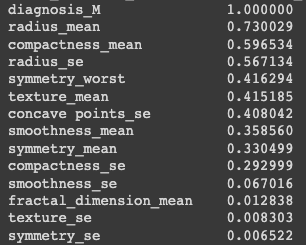


Adicionalmente, se observa una reducción considerable de variables a un total de 12.



* 1. ANÁLISIS DE CORRELACIONES PARA IRRELEVANCIA (PREDICCIONES)

Para esta instancia se estableció no eliminar esas correlaciones menores a 0.1, omitiendo la recomendación decretada para los análisis predictivos, dado la cantidad de variables restantes hasta este punto, En caso de eliminar las variables con dicha correlación se estaría reduciendo en gran magnitud el conjunto de datos. Aunque para subsanar la situación si se suprimieron aquellas variables con una correlación menor a 0.01, es decir, las últimas dos de esta lista:

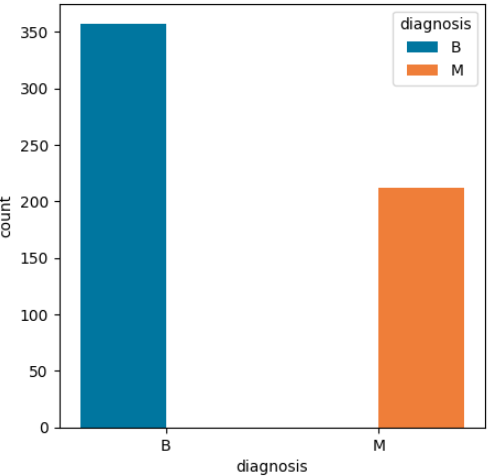


* 1. REDUCCIÓN DE DIMENSIÓN (OPCIONAL EN PREDICCIONES)

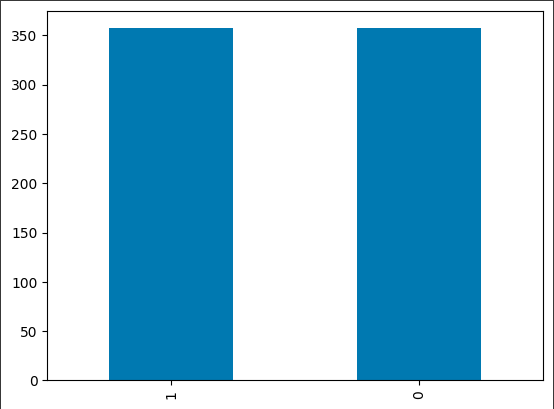
En este punto ya se había tenido una reducción de variables satisfactorio, pasando de 33 variables iniciales a sólo 12 contando la variable objetivo, es decir, ya no era necesario reducirlas más por el mismo riesgo de suprimir información valiosa para el modelado. Esta reducción de dimensionalidad prácticamente fue llevada a cabo con el manejo de correlaciones.

* 1. BALANCEO (CLASIFICACIÓN)

Este paso era más que necesario por las frecuencias absolutas halladas en la variable objetivo, existiendo un desbalance claro:

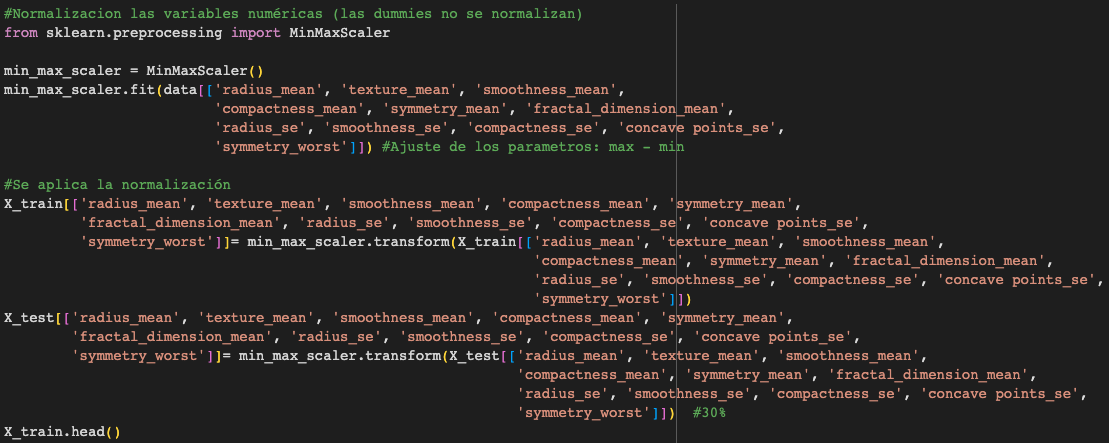


Para realizar el respectivo balanceo se hizo primero una división con el criterio 70-30 usando sklearn, el 70% de los datos para entrenamiento y el 30% restante para prueba. El balanceo se llevó a cabo con SMOTE y destinada a esa variable objetivo “diagnosis”. La particularidad del caso fue la aplicación del balanceo para el conjunto de datos de entrenamiento, es decir, ese 70% de la división inicial.



* 1. TRANSFORMACIONES

La transformación a realizar consistió en normalizar las variables numéricas, sin embargo, fue después de la creación del árbol de decisión dada a su interpretación y fue realizada tanto en el conjunto de entrenamiento con el de prueba, es decir, en el 100% de los datos.



1. **MODELAMIENTO, EVALUACIÓN E INTERPRETACIÓN**
   1. CONFIGURACIÓN MÉTODOS DE APRENDIZAJE SUPERVISADOS

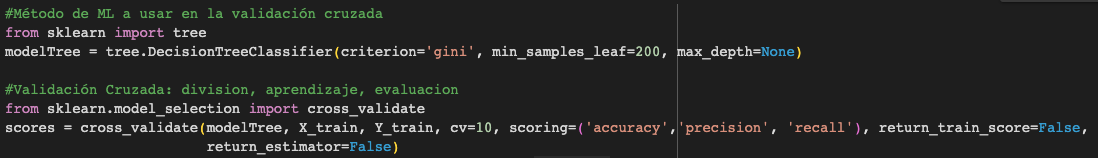
Para iniciar el modelado se eligieron aleatoriamente 4 modelos supervisados para aplicar al set de entrenamiento y se escogieron 4 medidas de calidad para categorizarlos, los métodos elegidos fueron:

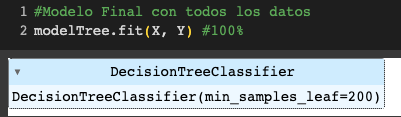
* Árbol de decisión
* K-Nearest-Neighbor (KNN)
* Red Neuronal
* Regresión Logística

**Medidas de evaluación:** precisión, recall, accuracy y curva ROC.

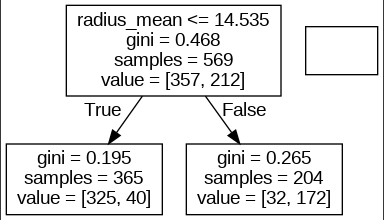
**Nota:** Para cada uno de los métodos se aplicó una validación cruzada previa a la ejecución del método.

***Árbol de decisión:***



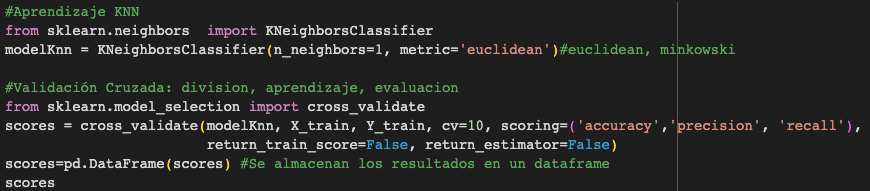


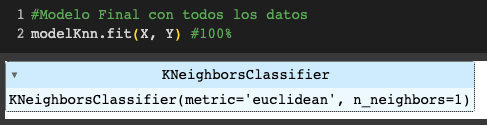
Una vez ejecutado el algoritmo del método con parámetros por defecto se hizo la graficación sencilla del árbol para validarlo.



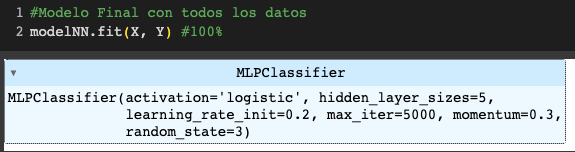
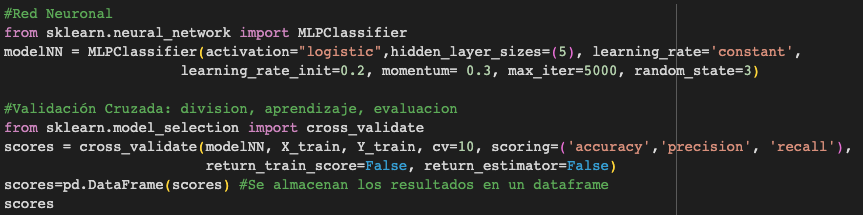
Indicando nuevamente que la variable más relevante entre las predictoras es el radio.

***KNN:***

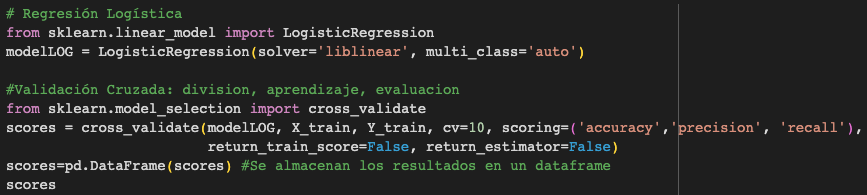


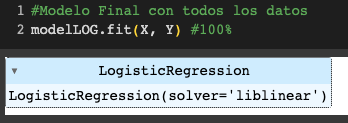


***Red Neuronal:***



***Regresión Logística:***





En general se aplicaron parámetros estándar en todos los modelos como forma indagatoria para las medidas de calidad. A partir de este experimento se seleccionaron más tarde los modelos a los cuales se les haría la hiperparametrización.

* 1. CONFIGURACIÓN MÉTODOS DE ENSAMBLE

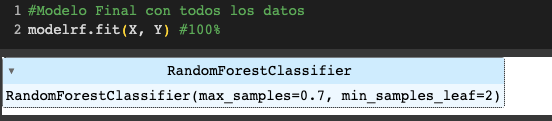
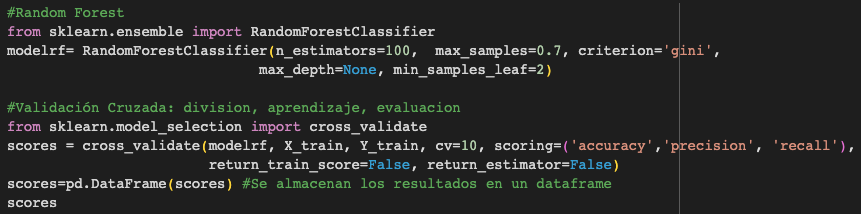
Continuando con el modelado, se seleccionaron aleatoriamente tres métodos de ensamble para compararlos bajos los mismos estándares con los modelos supervisados, los seleccionados fueron:

* Random Forest
* Boosting
* Bagging

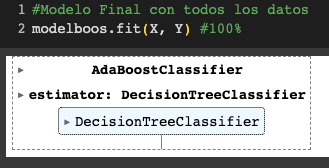
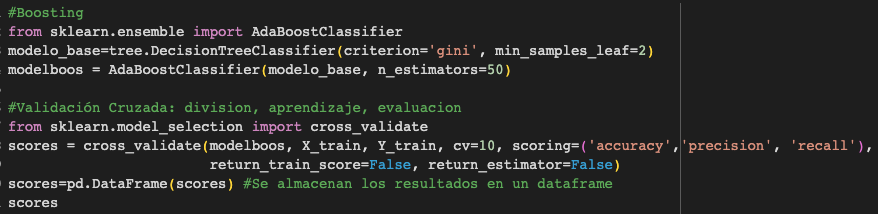
**Medidas de evaluación:** precisión, recall, accuracy y curva ROC.

De igual manera se emplearon parámetros estándar y se buscaron las mismas medidas de calidad.

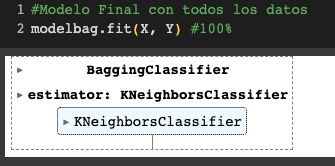
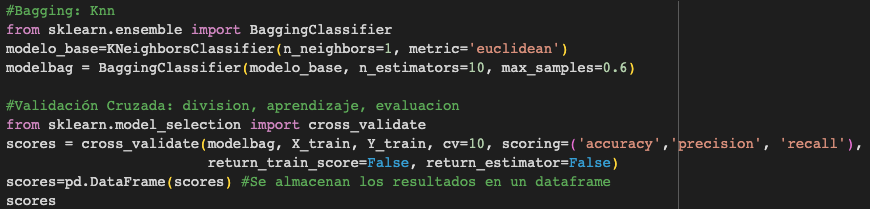
***Random Forest:***



***Boosting:***

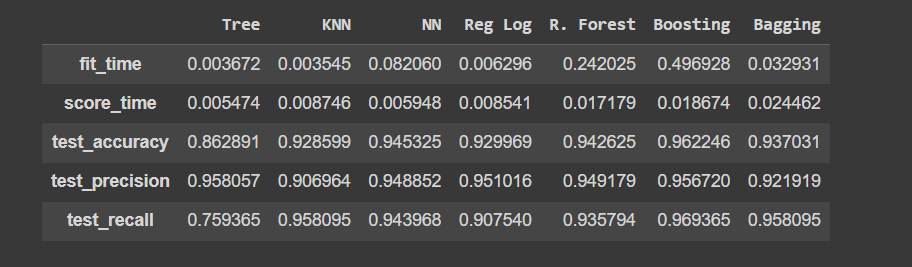


***Bagging:***



* 1. ANÁLISIS DE MEDIDAS DE CALIDAD

Para el análisis de medidas de calidad se realizó una tabla comparativa a partir del entrenamiento de cada modelo, con la aplicación de una validación cruzada en 10 secciones, cuya finalidad era obtener el promedio de las métricas de: fit time (tiempo necesario para entrenar el modelo), score time (tiempo requerido para calcular las puntuaciones o predicciones del modelo), test accuracy (la precisión o tasa de aciertos del modelo en los datos de prueba), test precisión (la precisión del modelo en la predicción de la clase positiva en los datos de prueba), test recall (la exhaustividad o tasa de verdaderos positivos del modelo en los datos de prueba).



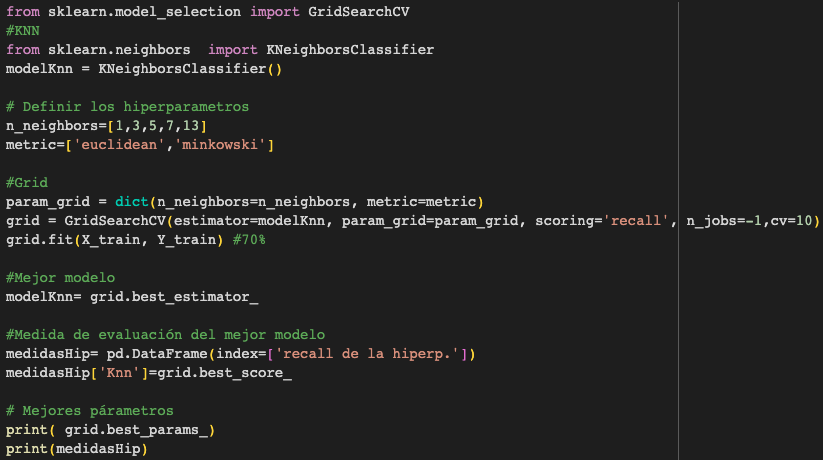
A partir de esta tabla comparativa, se escogieron los 3 modelos para el siguiente paso de hiperparametrización, esta decisión fue con base en la medida de test recall debido a que en este proyecto es fundamental que la tasa de verdaderos positivos sea más eficiente al momento de dar un diagnóstico médico, además, las otras métricas también fueron analizadas para tener una mayor certeza al momento de la elección de los 3 métodos.

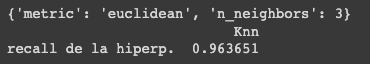
Por esto los modelos que presentan un alto test recall como: Knn con 0.95, NN con 0.94, Boosting 0.96 y Bagging 0.95, son los que se consideran mejor para la hiperparametrización. Se descartó el Bagging por encima de NN debido a que en las otras métricas la red neuronal tiene una ligera ventaja, además ya se tenía previamente un método de ensamble con mejor calidad en el entrenamiento que el Bagging.

* 1. HIPERPARAMETRIZACIÓN

Para este proceso se probaron muchos parámetros, de poco en poco por temas de tiempos de ejecución, a continuación, los códigos con los parámetros que mejor resultado obtuvieron y la selección por GridSearch entre ellos.

***KNN****:*



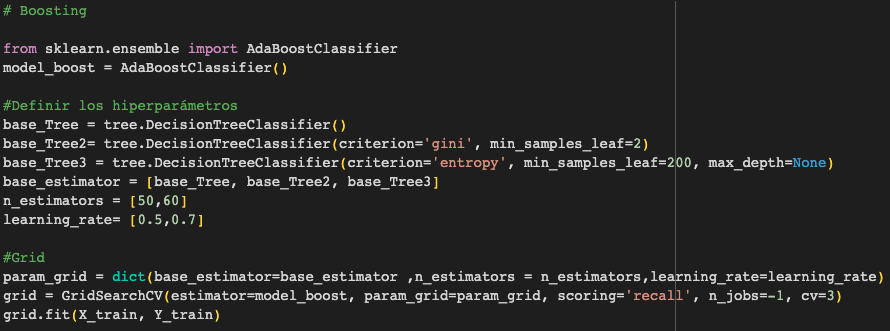


***Red Neuronal:***





***Boosting:***

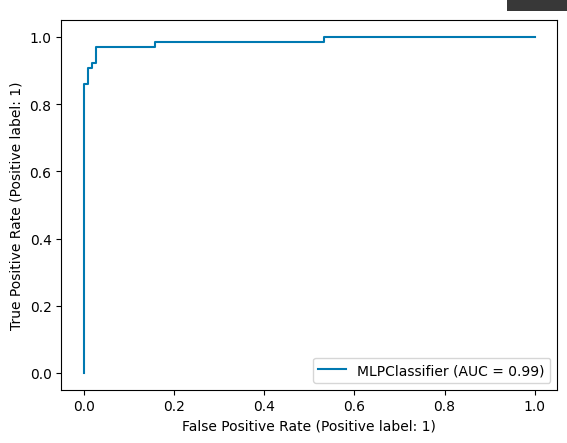


* 1. SELECCIÓN DEL MEJOR MODELO

Al evaluar el modelo se obtuvo lo siguiente:

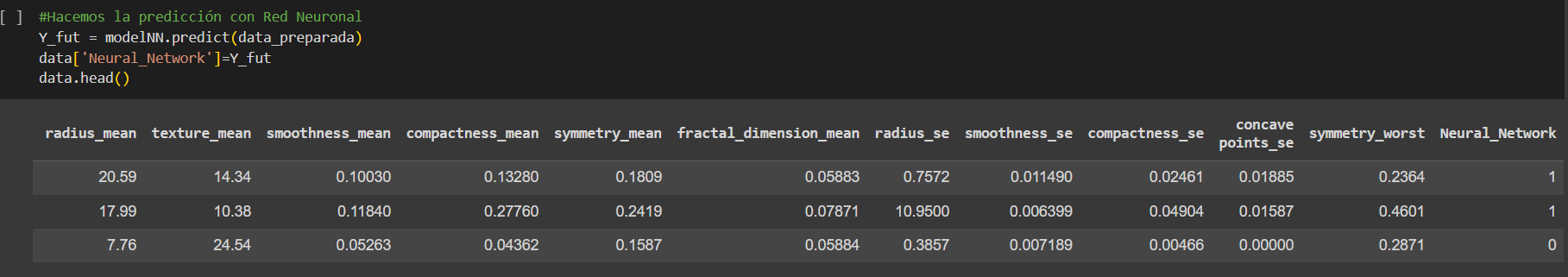
|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo** | **VP** | **FN** | **RECALL** |
| KNN | 59 | 5 | 0.95 |
| Red Neuronal | 62 | 2 | 0.96 |
| Boosting | 64 | 0 | 1.00 |

Aunque el Boosting haya obtenido un recall de 1.00 éste erró todos los intentos durante la prueba del modelo mientras que, por ejemplo, la red neuronal acertó todas las predicciones, eso junto con el hecho de haber obtenido una mejor curva ROC.



1. **DESPLIEGUE**
   1. PREDICCIÓN DE DATOS FUTUROS

Para el despliegue, se cargaron los datos futuros y se importó el pickle para cargar el modelo, la predicción para estos datos se realizó con la red neuronal y estos fueron los resultados obtenidos:



Se aprecia que la red acertó adecuadamente los tipos de tumor para cada caso que se le planteó en los datos futuros, por lo que se tiene un buen rendimiento del modelo, por esto mismo se optó por la red neuronal ya que el Boosting aunque haya presentado muy buenas métricas de evaluación, al momento del despliegue evidenciaba un caso de overfitting debido a que en la predicción de los datos futuros erraba cada caso que se le propuso.

* 1. CRONOGRAMA DE RE-ENTRENAMIENTO DEL MODELO

Debido a que el modelo está enfocado en el área de medicina, es muy importante que siempre tenga una efectividad alta y certera debido a que es una herramienta primordial al momento de ayudar a personas que puedan depender de un buen diagnóstico, por lo tanto se recomienda una actualización frecuente de los datos de los paciente para poder re - entrenar el modelo en lapsos de máximo cada 6 meses, sin embargo es importante realizar un seguimiento continuo del rendimiento del modelo en la práctica clínica. Si se observa una disminución en la precisión, una tasa de falsos positivos o falsos negativos inaceptablemente alta, o si el modelo no se ajusta bien a los nuevos datos, puede ser necesario reentrenar el modelo para mejorar su desempeño.