

**ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO**

**INGENIERIA BIOMÉDICA**

**PROYECTO MINERIA DE DATOS EN BIOMEDICINA**

**PRACTICA 2: MACHINE LEARNING**

**Desirée Rivera Rodríguez**

**Curso 2022-2023**

# INTRODUCCIÓN

En el siguiente proyecto de Machine Learning se desarrollará un modelo para predecir si un paciente tiene cáncer de mama o no a través de los datos de una biopsia. Para ello se implementará algoritmos de aprendizaje tanto supervisados (KNN, SVM, Naive Bayes y Arboles de Decisión) como no supervisado (K-Medias) para poder determinarla contaremos con un conjunto de datos para analizar, limpiar y entrenar el modelo y testear cuales obtienen mejores resultados.

Para conseguir que uno o más algoritmos tengan mejores resultados en sus predicciones, se someterán los datos a un análisis estadístico y tratamientos a los datos con la finalidad de adecuarlos para que los algoritmos hagan su trabajo de manera más efectiva.

La base de datos que ha sido elegida trata de la detección del cáncer de mama en los hospitales de Winsconsin, Estados Unidos. Estos datos se han obtenido a través de un diagnóstico hecho por biopsia. En este método se examinan pequeñas cantidades de tejido del tumor para obtener las características de las células individuales y propiedades contextuales importantes gracias a la cuales se ha logrado un diagnóstico exitoso.

A partir de este proceso, 10 características son modeladas numéricamente.

* **Radio**: el radio de un núcleo individual es medido promediando la longitud de la línea del segmento radial definido por centroide del límite y los puntos individuales del límite.
* **Perímetro**: es la distancia total entre los puntos del límite constituyen el perímetro nuclear.
* **Área**: se mide contando el número de pixeles en el interior de los límites y se añade la mitad de los píxeles del perímetro.
* **Compacidad**: el perímetro y el área son combinados para dar una medida de la compacidad del núcleo de la célula.
* **Suavidad**: es la diferencia entre la longitud de una línea radial y la longitud media de las líneas alrededor de esta.
* **Concavidad**: irregularidades de forma en el núcleo de la célula
* **Puntos cóncavos**: número de la concavidad del contorno.
* **Simetría**: se obtiene encontrando la línea más larga que pase por el centro de la célula.
* **Dimensión fractal**: es una característica de forma, es decir, a mayor valor corresponde a un menor contorno y por tanto a una mayor probabilidad malignidad.
* **Textura:** El valor medio, extremo (el más largo) y el error standard de cada característica son computados para cada imagen. Los valores extremos son más intuitivamente útiles para el problema en cuestión, ya que solo algunas células malignas pueden ocurrir en un ejemplo dado.

A partir de estas variables, se quiere predecir la variable de diagnosis, es decir si el tumor es benigno o maligno. Para ello se entrenarán algoritmos tanto supervisado como no supervisado que puedan identificar las dos clases

# INSERTAMOS EL DATASET

Una vez obtenida esta base de datos lo primero que hacemos es el procesamiento de datos para ello cargamos los datos a utilizar:



# ESTADÍSTICA DESCRIPTIVA

Procedemos a estudiar nuestros datos para ello visualizamos los datos de varias maneras diferentes.

La primera será ver la estructura de los datos, es decir que tipos de datos tenemos, el tamaño de la base de datos y la cantidad de variables.

Texto

Descripción generada automáticamente

De estas informaciones podemos decir que nuestra base tiene 569 muestras de tejido de pacientes distintos con 33 variables cada uno. Siendo 30 variables cuantitativas y dos cualitativas. De estas 33 variables podemos ver que la primera variable es el ID la cual no nos aporta mucha información así que la eliminaremos. Asimismo, vemos que la variable 33 sus datos son NA, los cuales tampoco en aportan mucha información por lo tanto también lo eliminamos.



Como se puede comprobar los datos se pueden dividir en tres partes: la media, el error estándar y el “peor valor” de las diez variables que se han mencionado anteriormente. Para seguir con el análisis descriptivo y observar los resultados con más claridad se dividirá en las tres categorías.

Una captura de pantalla de un celular con texto e imagen

Descripción generada automáticamente con confianza baja

Ahora miramos las estadísticas generales de las características del conjunto de datos. Estos serán la media, mediana, desviación típica, curtosis, rango, los cuartiles entre otros. Primero lo hacemos con la primera categoría:

Un conjunto de letras blancas en un fondo blanco

Descripción generada automáticamente con confianza media

Tabla

Descripción generada automáticamente

Ahora de la segunda categoría:

Texto

Descripción generada automáticamente

Imagen que contiene marcador, teclado, grande, sostener

Descripción generada automáticamente

Y por último la tercera categoría:

Imagen que contiene marcador, firmar, sostener, azul

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Teniendo estos datos, si nos fijamos en el valor máximo y mínimo de cada una de las variables nos damos cuenta de que cada una de las variables tienen escalas diferentes lo cual puede producir imprecisiones. Por consiguiente, hay que tenerlo en consideración para normalizar los datos al aplicar algunos algoritmos de aprendizaje.

Pasamos a ver la correlación entre las variables para ello primero quitamos nuestra variable predictora:



Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Como podemos observar la misma variable en las tres categorías tiene unos valores de correlación alto como era de esperar. Asimismo, en las tres categorías podemos ver que el radio, el perímetro y el área tiene una alta correlación en las tres categorías. De la misma forma, podemos visualizar que la concavidad, compacidad y los puntos cóncavos tienen también una alta correlación.

Otra visualización interesante es ver los datos estadísticos de la variable predictora, es decir de diagnósticos. Además, convertimos la variable de predictora en factor para que tenga etiquetas.

Una captura de pantalla de un celular

Descripción generada automáticamente con confianza media

Texto

Descripción generada automáticamente

Gráfico, Gráfico de barras, Cuadrado

Descripción generada automáticamente

Por último, se muestra de la dispersión de los datos y datos atípicos por medio de un gráfico de caja.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente Diagrama

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Aquí pongo un ejemplo de la variable radio en sus tres categorías, pero al ejecutar todo el código podemos ver que las 30 variables que vamos a utilizar para predecir la enfermedad tienen outliers que pueden influir.

# ALGORITMOS DE APRENDIZAJE SUPERVISADO.

Primero aplicaremos algoritmos de aprendizaje supervisado los cuales trabajan con datos etiquetados intentado encontrar una función que con unas variables de entrada se les asigne una etiqueta de salida adecuada. En nuestro caso usaremos estos algoritmos con la finalidad de diagnosticar una enfermedad, por lo tanto, nos encontramos ante un problema de clasificación. En este apartado solo nos centraremos en las medidas de rendimiento de cada uno de los algoritmos elegidos.

* **KNN:**

Es un método el cual busca en las observaciones más cercanas a la que se está tratando de predecir y clasifica el punto de interés basado en la mayoría de los datos que le rodean.

Primero calcula la distancia entre la observación que se quiere clasificar y el resto de observaciones que pertenecen a la muestra de entrenamiento. Para ello la importancia de normalizar. En segundo lugar, se selecciona los k elementos más cercanos, es decir con menor distancia. Por último, se mira la clase a la que pertenece según la cantidad de k más cercana

En primer lugar, se probó con un k =21. Al hacer nuestra validación de nuestro algoritmo vemos que tiene una alta precisión con un nivel de Kappa cercano a 1 lo que significa los resultados obtenidos no se han dado por “casualidad”.

De la misma manera tendremos la medida F que es la combinación de las medidas de precisión y sensibilidad. Con esta medida nos será más fácil comparar los distintos algoritmos. En este caso es muy alta.

Texto

Descripción generada automáticamente

Texto, Escala de tiempo

Descripción generada automáticamente

Para ver si se podía mejorar el algoritmo hicimos un bucle y nos percatamos que k=7 es el que mejores valores de las medidas de rendimiento tiene dado que acierta más verdaderos negativos o errores subiendo tanto sus valores de especificidad como sensibilidad. Asimismo, se obtiene mejores resultados en la medida F.

Texto

Descripción generada automáticamente con confianza media

Texto

Descripción generada automáticamente con confianza media

Por último, probamos si usando la tipificación en lugar de la normalización se mejora los resultados con el mismo k vecinos más cercanos.

Texto

Descripción generada automáticamente

Viendo las medidas de rendimientos vemos que es un muy buen modelo, pero nos quedamos con el modelo normalizado con k=7 al darnos mejores medidas de rendimiento.

* **SVM:**

Las Máquinas de Vector Soporte se basa en el concepto de hiperplano. En este modelo se busca de un hiperplano que separe de forma óptima a los puntos de una clase de la de otra, que eventualmente han podido ser previamente proyectados a un espacio de dimensionalidad superior. La característica fundamental es buscar el hiperplano que tenga la máxima distancia (margen) con los puntos que estén más cerca de él mismo. De esta forma, los puntos del vector que son etiquetados con una categoría estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado.

Primero probamos con el parámetro de Kernel ‘vanilladot’ y analizamos las siguientes medidas de rendimiento

Texto

Descripción generada automáticamente

AL hacer nuestra validación podemos observar que tiene una precisión muy cercana a 1 por lo que se puede decir que predice si una persona tiene cáncer o no de manera casi segura. Con el valor del coeficiente de Kappa se puede demostrar que estas predicciones no se dan por “casualidad”. Junto con la medida F y la precisión podemos decir que es un modelo muy bueno.

De la misma forma vamos a probar si mejora aun más el modelo cambiando el parámetro de Kernel a rbfdot

Texto

Descripción generada automáticamente

Aquí vemos que tanto la accuracy como el coeficiente Kappa es igual. No obstante, podemos ver que en este modelo no hay ningún falso positivo conllevando que la especificidad y la precisión sea del valor máximo. Asimismo, la medida F es un pelín más alta que en el anterior por lo que nos quedamos de momento con este modelo.

Por último, vamos a comprobar cambiando otra vez el parámetro de Kernel a "tanhdot".

Texto

Descripción generada automáticamente

Viendo los resultados de la medida de rendimiento son buenos, pero no tanto como el del modelo que hemos usado con el parámetro de Kernel rdfdot.

* **Arboles de Decisión:**

Es un gran modelo para darle un toque explicativo a las decisiones que se toman para predecir si es en nuestro caso Benigno o Maligno. Aplicando el modelo vemos su rendimiento:

Texto

Descripción generada automáticamente

Podemos apreciar que tiene valores muy buenos en las medidas de rendimiento, pero vamos a ver si se puede mejorar el modelo usando la técnica de poda. Para ello utilizamos el concepto de Boosting para mejorar el modelo. En este agregaremos unos parámetros que será el número de árboles separados. Estos árboles juntaran su fortalezas y debilidades para mejorar el modelo. Probamos con 10, 20, 30 árboles y con las medidas de rendimiento vemos que el mejor modelo es el que utiliza 20 trials:

Texto

Descripción generada automáticamente

Dibujamos su árbol de decisión para ver las decisiones que toma para elegir una clase u otra.

Diagrama

Descripción generada automáticamente

* **Naives Bayes:**

Naïve Bayes es un algoritmo de aprendizaje automático basado en el teorema de Bayes. Usa las probabilidades condicionales de las variables independientes para poder clasificar cada observación de la base de datos. Procedemos a ver las medidas de rendimiento.

Texto

Descripción generada automáticamente

Podemos comentar que es un modelo sencillo y muy efectivo. No obstante, para este modelo comparándolo con otros no es el mejor aunque tendría una gran utilidad.

* **Comparación**

Comparando todos los modelos podemos destacar que los 4 son muy buenos algoritmo de aprendizaje para predecir el cancer de mama con los datos de la base proporcionada. Pero si nos tuviéramos que quedar con uno y analizando las medidas de rendimiento sería el SVM con un parámetro Kernel de rbfdot.

# ALGORITMOS DE APRENDIZAJE SUPERVISADO.

Utilizamos el K-Means que es un tipo de aprendizaje no supervisado, que se utiliza cuando se tienen datos no etiquetados, es decir, datos sin categorías o grupos definidos. El objetivo de este algoritmo es encontrar grupos en los datos. Los puntos de datos se agrupan según la similitud de características.

En este caso tenemos que suponer que la variable categórica diagnóstico no nos la dan entonces a partir de las otras variables vamos a ver si el algoritmo reconoce los dos grupos que tenemos en la variable diagnosis.

Se aplican unos centroides aleatorios y con un bucle vamos a ver la inercia intra total en la dispersión para escoger nuestros números de clusters. Al aplicar el método del codo no se ve muy bien:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Por esa razón vamos a usar el método de ‘silhouette’ que es más preciso. En este se puede ver que hay dos grupos que se pueden corresponder a nuestras etiquetas que hemos usado en los algoritmos supervisados:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Por último, vemos nuestros clusters:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente