Aplicación del Aprendizaje Automático para la Clasificación de la Vulnerabilidad del Cáncer de Pulmón

Miguel Chacón Lopéz
Martín Ospina Uribe
Felipe Henao Gómez
José A. Durán C.

Profesor:
Santiago Hernández Torres

Proyecto
Estadística Multivariada Avanzada
Escuela de Ciencias Aplicadas e Ingenieria
Universidad EAFIT

26 de noviembre de 2023

Índice

1. Introducción

En el ámbito de la medicina el tiempo juega un papel crucial en el proceso de diagnostico y tratamiento de cada paciente. Según datos publicados recientemente por la Organización mundial de la salud el cáncer de pulmón es la causa principal de muertes relacionadas con el cáncer en todo el mundo, y su tasa de mortalidad es la más elevada tanto entre hombres como entre mujeres. El cáncer de pulmón suele diagnosticarse en una etapa avanzada de la enfermedad, cuando las opciones de tratamiento son limitadas OMS (2023). Un tiempo oportuno de detección aumenta en gran medida las probabilidades se supervivencia de un paciente.

La inteligencia artificial es una herramienta útil en diversos contextos médicos como la predicción del padecimiento de enfermedades basándose en registros históricos. Su capacidad de procesar una enorme cantidad de datos a la vez le permite alcanzar un nivel muy alto de precisión en sus resultados y aun mas importante en un corto periodo de tiempo. Es allí donde radica su gran potencial para el diagnóstico sistemático de enfermedades.

En este trabajo utilizaremos distintos métodos de inteligencia artificial, para ser más precisos métodos de aprendizaje supervisado para clasificar el nivel de propensión de un paciente a

sufrir cáncer de pulmón en el mediano plazo. Esto con el fin de comparar el rendimiento y desempeño de distintos métodos al realizar esta tarea. Consideramos que estos resultados podrían ser de utilidad para la comunidad científica y médica en su búsqueda de crear y desarrollar herramientas que permitan mejorar la calidad de vida de los seres humanos.

2. Pregunta de Investigación y Objetivos

El objetivo principal de este proyecto es desarrollar y evaluar varios modelos de aprendizaje automático que sean capaces de predecir el nivel de enfermedad pulmonar crónica en pacientes utilizando datos clínicos y características relacionadas con los pulmones. Específicamente, buscamos lograr los siguientes objetivos:

- Analizar y preparar la base de datos de pacientes con información sobre enfermedades pulmonares y características relevantes.
- Evaluar y comparar varios algoritmos de aprendizaje automático para determinar cuál es el más efectivo en la predicción del nivel de enfermedad pulmonar crónica.
- Evaluar el rendimiento del modelo a través de métricas de evaluación, como precisión, sensibilidad, especificidad y f1-score.
- Implementar el modelo seleccionado y optimizarlo para lograr el mejor rendimiento posible.
- Comunicar los hallazgos a través de un informe técnico y visualizaciones claras que respalden los resultados del proyecto.

La pregunta de investigación fundamental que se abordará en este proyecto es la siguiente:

Basados en un muestreo estadístico de pacientes clínicos ¿Cuál es el modelo de aprendizaje automático más efectivo para predecir con precisión el nivel de enfermedad pulmonar crónica?

3. Metodología de Investigación

La presente propuesta de investigación tiene como objetivo evaluar y comparar la eficacia de distintos métodos de aprendizaje supervisado aplicados a un problema de clasificación de enfermedad pulmonar crónica. Realizaremos el estudio utilizando un conjunto de datos tabulares con diversas características como la edad del paciente, el género, el consumo de alcohol, el consumo de cigarrillos, entre otros; junto con etiquetas que indican el nivel de la enfermedad pulmonar crónica.

1. **Definición del Problema:** Definiremos el problema de clasificación y estableceremos la importancia de una predicción precisa en la enfermedad pulmonar crónica.

2. **Revisión Bibliográfica:** Realizaremos una revisión bibliográfica sobre los métodos de aprendizaje supervisado más utilizados en problemas de clasificación similares basados en conjuntos de datos tabulares. También examinaremos estudios previos relacionados con la detección y clasificación de enfermedades pulmonares crónicas, utilizando técnicas de aprendizaje automático.

3. **Adquisición y Preparación de Datos:** Utilizaremos datos sobre los pacientes como la edad, el género, consumo de alcohol y consumo de cigarrillos, junto con las etiquetas de nivel de enfermedad pulmonar crónica. Llevaremos a cabo un proceso de limpieza de datos para eliminar valores atípicos y datos faltantes, asegurando que los datos estén listos para su posterior procesamiento.

4. **Selección de Algoritmos de Aprendizaje Supervisado:** Identificaremos y seleccionaremos diversos algoritmos de aprendizaje supervisado adecuados como árboles de decisión, máquinas de soporte vectorial, regresión logística, bosques aleatorios, redes neuronales, entre otros.

5. **Implementación y Evaluación de Modelos:** Implementaremos los distintos algoritmos de aprendizaje supervisado utilizando el conjunto de datos preparado previamente. Luego, evaluaremos el rendimiento de cada modelo utilizando métricas de evaluación pertinentes como precisión, recall y f1-score. Compararemos los resultados obtenidos para cardinalizar la efectividad de cada método en la clasificación de la enfermedad pulmonar crónica. Además, se hará una transformación de los datos y se repetirá el proceso con el fin de evaluar el efecto de esta transformación sobre los resultados.

6. **Comparación de Resultados y Conclusiones:** Realizaremos una comparación del rendimiento de cada método de aprendizaje supervisado en función de las métricas de evaluación establecidas. Identificaremos el método que muestre un rendimiento más prometedor y discutiremos las posibles razones detrás de su desempeño superior. Finalmente, daremos las conclusiones y ofreceremos recomendaciones para la aplicación de técnicas de aprendizaje supervisado en problemas similares de clasificación de enfermedades pulmonares crónicas.

7. **Redacción del Informe Final y Difusión de Resultados:** Documentaremos el proceso de investigación, desde la preparación de los datos hasta el análisis de resultados y conclusiones.

4. Datos y Análisis Previo

Los datos obtenidos para el desarrollo de este proyecto se obtuvieron del dataset subido por THE DEVASTATOR (2022) en la pagina Kaggle. Este conjunto de datos se presenta en un archivo separado por comas (.csv) y contiene múltiples variables con respecto a la

condición de los pacientes que pueden relacionarse con su riesgo de tener cáncer de pulmón, las características presentadas son las siguientes:

- Edad: edad del paciente. (numérica)
- Genero: genero del paciente. (categórico)
- Contaminación del aire: nivel de exposición a la contaminación del aire de la paciente. (categórico)
- Consumo de alcohol: nivel de consumo de alcohol del paciente. (categórico)
- Alergia al polvo: nivel de alergia al polvo del paciente. (categórico)
- Riesgos laborales: nivel de riesgos laborales del paciente. (categórico)
- Riesgo genético: nivel del riesgo genético del paciente. (categórico)
- Enfermedad pulmonar crónica: nivel de enfermedad pulmonar crónica del paciente. (categórico)
- Dieta balanceada: nivel de dieta equilibrada de la paciente. (categórico)
- Obesidad: nivel de obesidad del paciente. (categórico)
- Fumar: nivel de tabaquismo del paciente. (categórico)
- Fumador pasivo: nivel de fumador pasivo del paciente. (categórico)
- Dolor torácico: nivel de dolor torácico del paciente. (categórico)
- Tos con sangre: nivel de tos con sangre del paciente. (categórico)
- Fatiga: nivel de fatiga del paciente. (categórico)
- Perdida de peso: nivel de perdida de peso del paciente. (categórico)
- Dificultad respiratoria: nivel de dificultad respiratoria del paciente. (categórico)
- Sibilancias: nivel de sibilancias del paciente. (categórico)
- Dificultad de deglución: nivel de dificultad de deglución del paciente. (categórico)
- Acropaquia de las uñas: nivel de acropaquia de las uñas del paciente. (categórico)

Todas las variables anteriores son en mayor parte una variable categórica que muestra el nivel de cierta característica sobre el paciente, y finalmente la variable dependiente corresponde a el nivel de riesgo de cáncer de pulmón, la cual se da en tres niveles distintos: bajo, medio y alto.

Para evaluar los modelos utilizados, usaremos distintas métricas correspondientes a la clasificación de las distintas clases, tales como el accuracy para cada clase, el f1 score para cada clase, y también se tendrá en cuenta la matriz de confusión de la clasificación.

Para poder trabajar con estos datos, primero haremos un análisis exploratorio de estos. Lo primero a considerar es que los datos estén correctos, esto es, que no hayan celdas vacías. Después, debemos evaluar cómo trabajar con las variables categóricas, para esto consideramos valores enteros para cada categoría, que es como se presenta en la base de datos, pues las variables categóricas se les da un valor entero entre 1 y 9. Para la variable dependiente se debe hacer una transformación de esta, para esto hacemos una codificación por números enteros, esto es, representamos el valor de bajo como 0, medio como 1 y alto como 2, para evitar problemas con modelos que no puedan trabajar con este tipo de variables. Algo que también se debe tener en cuenta es si existen correlaciones altas entre las características, para evitar darle información redundante al modelo, y poder de esta manera reducir la dimensión de los datos. Finalmente, se puede considerar una reducción de dimensionalidad del conjunto de datos, ya sea por la eliminación de algunas características que se pueden considerar irrelevantes para el problema o por medio de un análisis de componentes principales, que este segundo puede aportar también a eliminar las correlaciones que existen entre dichas variables.

5. Resultados

Para el desarrollo del trabajo, se consideró tres distintos modelos, estos son: máquinas de vector soporte, un random forest y el KNN (K vecinos mas cercanos), todas se hacen usando la librería scikit-learn en python. Para la selección de los hiperparámetros del modelo, se uso la librería de *Optuna* en python.

5.1. Preprocesamiento y Partición de los Datos

En primer lugar, se debe hacer un preprocesamiento de los datos para poder trabajar con estos. Lo primero que se considero fue quitar columnas de datos que no son relevantes para la predicción, pues estos son valores únicos para identificar los pacientes y el numero de registro, estas columnas son la de *index* y *Patient Id*.

Luego, se reviso que no hubieran datos nulos, esto es, que no hayan registros faltantes. De esto, se obtuvo que no habían datos nulos por lo que no hay necesidad de hacer algún tipo de cambios para esto. Algo que también se debe revisar es con respecto a que los datos estén balanceados, esto es, que no haya alguna clase con significativamente mayor numero de registros en comparación a las demás, de lo cual se obtuvo que están bien balanceados, pues se presentan 365 registros con un riesgo alto, 332 con riesgo medio y 303 con riesgo bajo.

Al mirar el tipo de dato de las variables, se pudo ver que la columna *Level*, la cual corresponde a la etiqueta, tiene un tipo de dato objeto, pues esta tiene los valores High, Medium y Low, por tanto, decidimos hacer una codificación de estos, y representamos el valor de Low como 0, Medium como 1 y High como 2.

Lo otro a tener en cuenta es la partición de los datos. Como se va realizar una búsqueda de los hiperparámetros, se parte el conjunto de datos en tres conjuntos, uno para entrenar, uno para la prueba y otro para validar. El conjunto de prueba se usa para evaluar las combinaciones de los parámetros y seleccionar la mejor, ya con la mejor se usa el conjunto de validar para evaluar los resultados. La partición se hace de 60 % para entrenar, 20 % para prueba y 20 % para validar.

Para la selección de los mejores parámetros, se considero la métrica f1 score, la cual hace una media harmónica entre el recall y precision. El recall corresponde a la fracción de predicciones de una clase que si eran en realidad de esa clase y la presicion corresponde a la cantidad de datos que eran de una clase que predijo como dicha clase. Así, el recall tiene en cuenta los falsos positivos y la presicion los falsos negativos. Como estos valores se calculan por clase, se considero un promedio macro para el calculo de estas dos métricas, esto se hace sumando todos los verdaderos positivos, falsos positivos y falsos negativos para cada clase y calculando las métricas con dichos valores, lo cual resulta como en un promedio ponderado de los valores de las métricas para cada clase. Ya con estas métricas se hace un promedio harmónico y se obtiene el f1 score, por tanto, queremos minimizar la cantidad de falsos negativos y falsos positivos ya que ambos son igual de importantes de evitar.

5.2. Máquina de Vector Soporte (SVM)

En primer lugar, se entrenó una maquina de vector soporte para la clasificación. Para la selección de hiperparámetros, se considero el valor del parámetro C, que corresponde como a una relajación en el entrenamiento, entre 10^{-10} y 10^{5} en una escala logarítmica, también, se considero el kernel, el cual puede ser lineal, polinomial, sigmoide o una función de base radial (rbf). De los ensayos hechos por *Optuna*, se obtuvo que la mejor configuración de parámetros fue con C=225.89 y un kernel rbf, con un f1 score de 1, sin embargo, esta no fue la única con f1 score de 1 pero fue la primera que encontró. A continuación, se muestra una gráfica del historial de optimización hecha por *Optuna*, la cual muestra los ensayos y el valor en el objetivo que corresponde al f1 score.

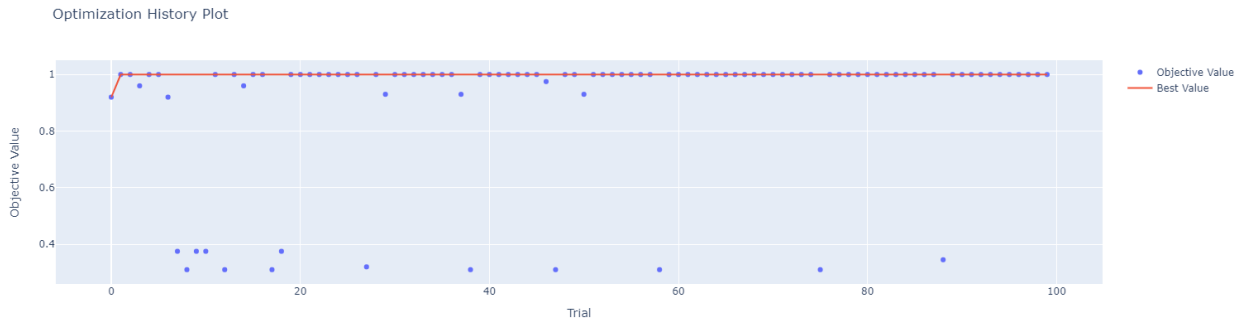

Figura 1: Historial de optimización para la SVM

De la figura, podemos ver que hubo muchos de los ensayos en los que se obtuvo un valor de 1 para el f1 score. Por tanto, se hizo la siguiente gráfica para ver la relación entre los valores de los parámetros y la función objetivo.

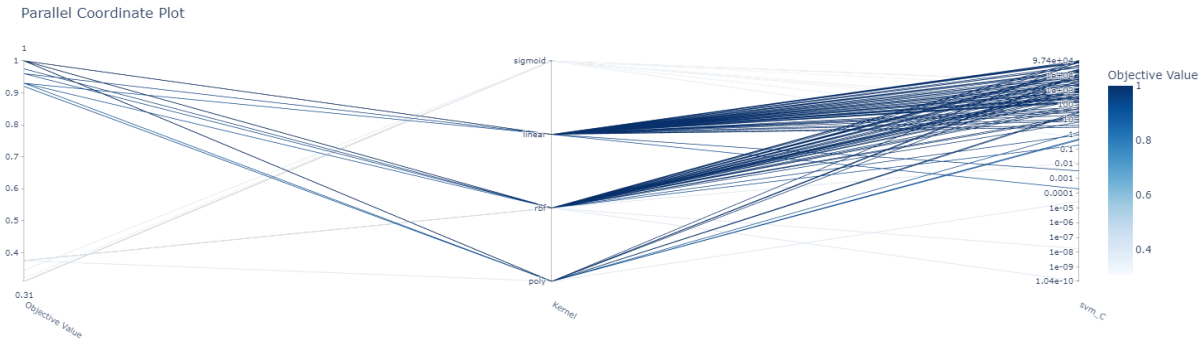

Figura 2: Parallel Coordinate Plot para la SVM

La figura anterior muestra como se relaciona cada valor de los hiperparámetros con el valor en la función objetivo, de esta podemos ver que en general un valor muy bajo de C y un kernel sigmoide resultan en un bajo valor en el f1 score, pero los valores mas altos de C y los otros kernels si resultan en buenos resultados. Adicionalmente, se evaluó la importancia de los hiperparámetros sobre el f1 score y se obtuvo la siguiente figura.

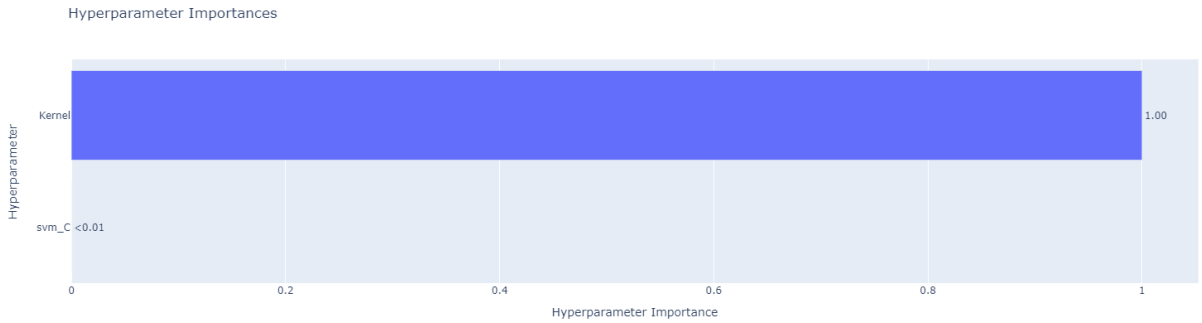

Figura 3: Importancia de los hiperparámetros para la SVM

Podemos ver que el kernel es lo más importante de la SVM y el valor de C no tiene tanta importancia como el kernel.

Finalmente, con base en los mejores hiperparámetros obtenidos dichos anteriormente, se hizo una validación del modelo con el conjunto de validación y se obtuvo los siguientes resultados.

	precision	recall	f1-score	support
Low	1.00	1.00	1.00	63
Medium	1.00	1.00	1.00	72
High	1.00	1.00	1.00	65
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Figura 4: Métricas de los resultados para la SVM

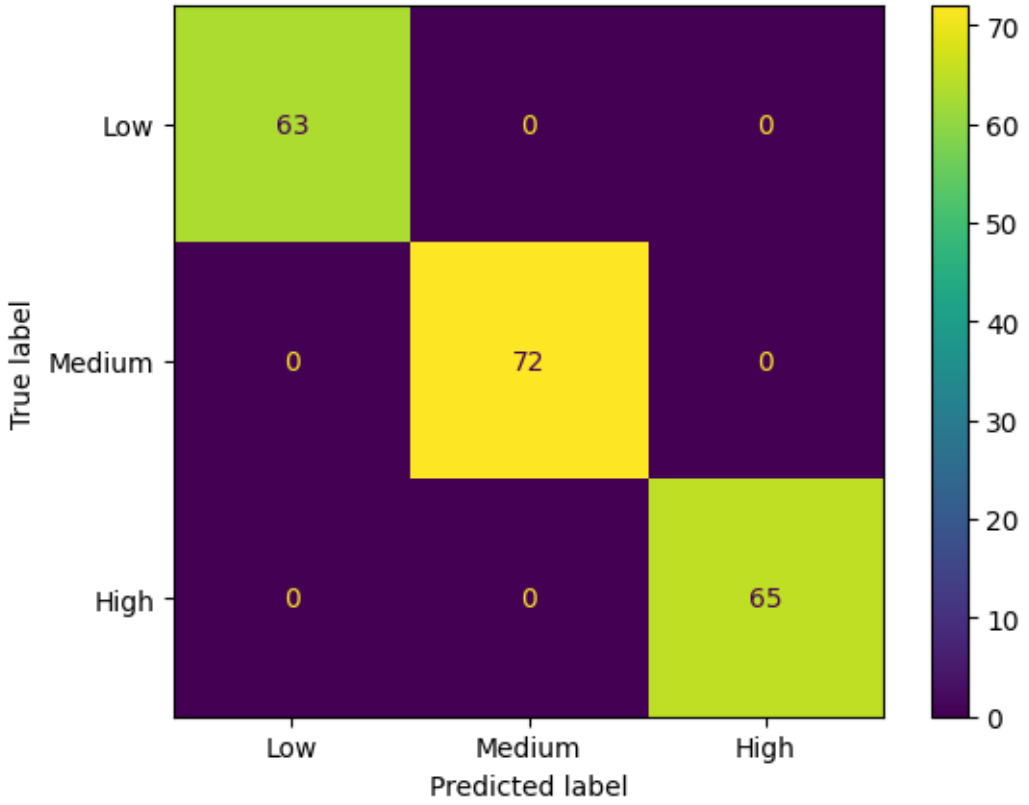

Figura 5: Matriz de confusión para la SVM

Con base en estos resultados, podemos decir que el modelo desarrollado es perfecto, pues no se equivoca en las predicciones y las métricas para cada clase y sus promedios son 1, lo que indica que no se equivoca.

Adicionalmente queremos estudiar y analizar el impacto que las distintas variables tienen en las clasificaciones realizadas por el modelo. Para esto, decidimos utilizar la librería *SHAP* una librería basada en la teoría de juegos cuya finalidad es dar una mayor intratabilidad a los distintos modelos de inteligencia artificial.

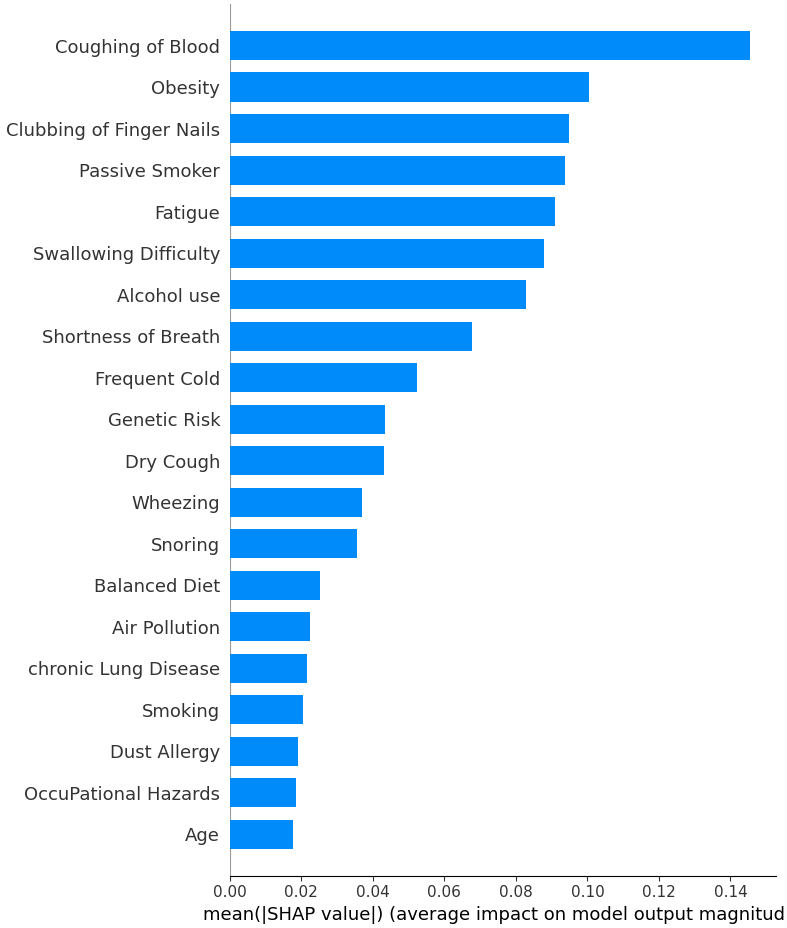

Figura 6: Valores SHAP medios del impacto de cada variable

Podemos observar como en este modelo de maquinas vector soporte las 5 variables que tienen un mayor impacto al momento de realizar las clasificaciones son *Coughing of Blood, Obesity,Clubbing od Finger Nails, Pasive Smoker* y *Fatigue*.

5.3. Random Forest

Nuestro segundo modelo es *Random forest*. Utilizamos *Optuna* para la optimización de los siguientes hiperparámetros. Primero tenemos *Criterion* que corresponde a la función para medir la calidad de una división, asignamos dentro de sus posibles valores *gini* , *entropy* y *log loss*. El segundo hiperparámetro es *Max depth* que corresponde a la profundidad máxima del árbol, en esto caso hicimos *Optuna* experimentara con valores del 1 al 15. El tercer hiperparámetro es *estimators* que corresponde a la colección de subestimadores ajustados. Obtuvimos por medio de *Optuna* que la mejor configuración de parámetros para este caso es *Criterion* tomando como función *log loss*, *Max depth* con un valor de 8 y *estimators* con un valor de 111. Esta configuración logro obtener un f1-score de 1.0 al igual que en caso anterior no fue la única configuración en lograrlo pero si la primera. La gráfica del historial

de optimización se encuentra a continuación

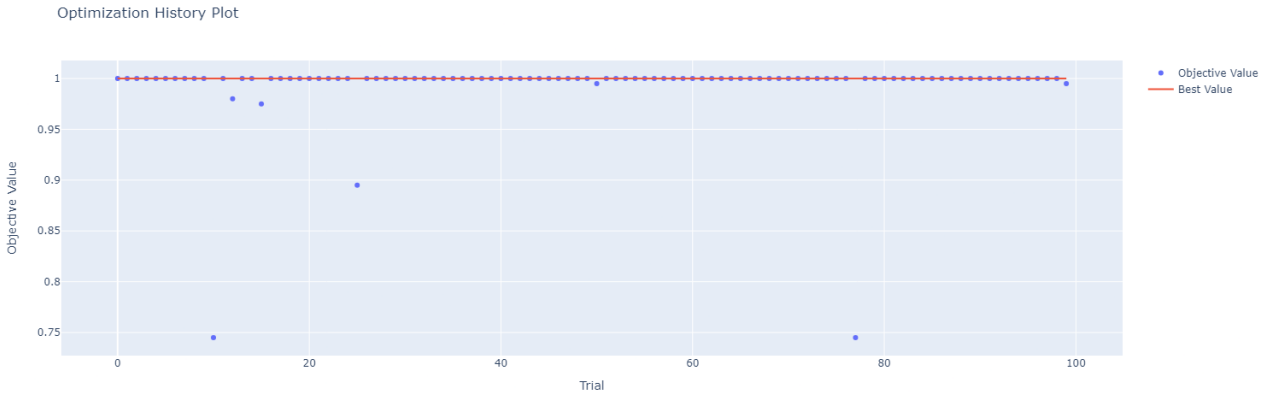

Figura 7: Historial de optimización RF

Podemos observar que hubo una gran cantidad de configuración que logro un f1 score de 1.0 es por esto que decidimos realizar la siguiente gráfica para visualizar la relación entre los valores de los parámetros y la función objetivo

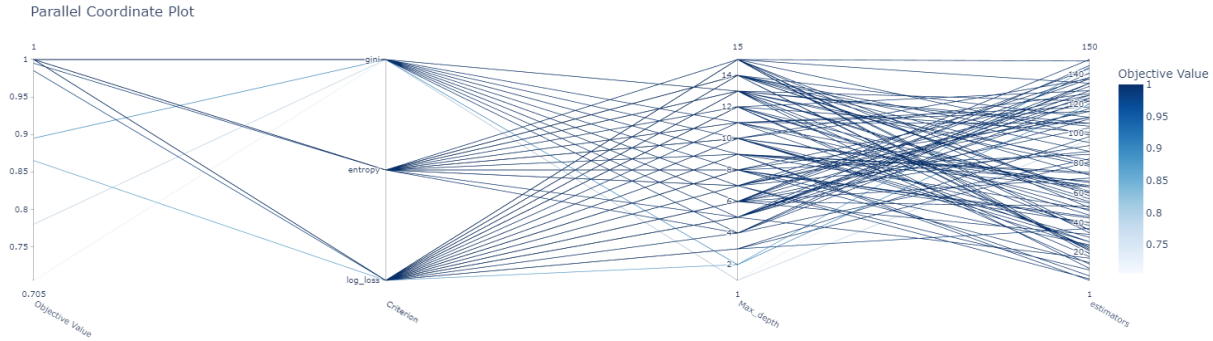

Figura 8: Parallel Coordinate Plot RF

En este caso no le logra visualizar una tendencia muy clara de los valores de los hiperparámetros sobre la función objetivo

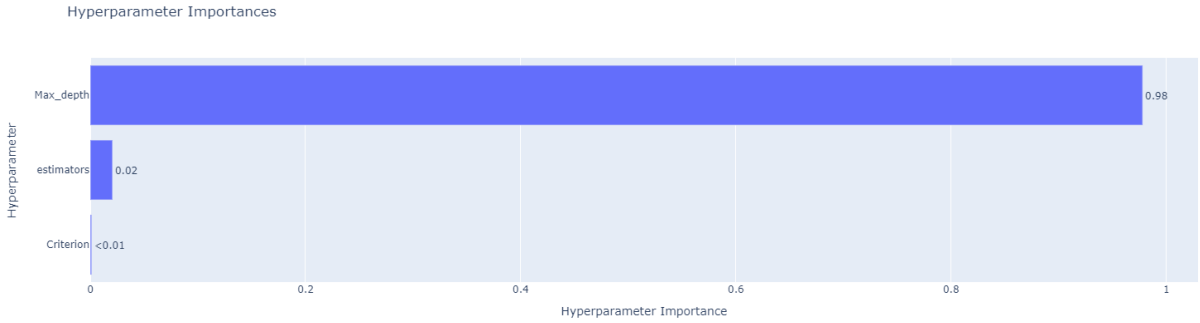

Figura 9: Importancia de los hiperparámetros del RF

La gráfica nos muestra que el hiperparámetro que mas importancia tiene es la profundidad máxima, mientras la importancia de los otros dos hiperparámetros es mucho menor.
Con los valores de los hiperparámetros obtenidos validamos el modelo, los resultados se presentan a cotinuacion

	precision	recall	f1-score	support
Low	1.00	1.00	1.00	63
Medium	1.00	1.00	1.00	72
High	1.00	1.00	1.00	65
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Figura 10: Métricas de los resultados para el RF

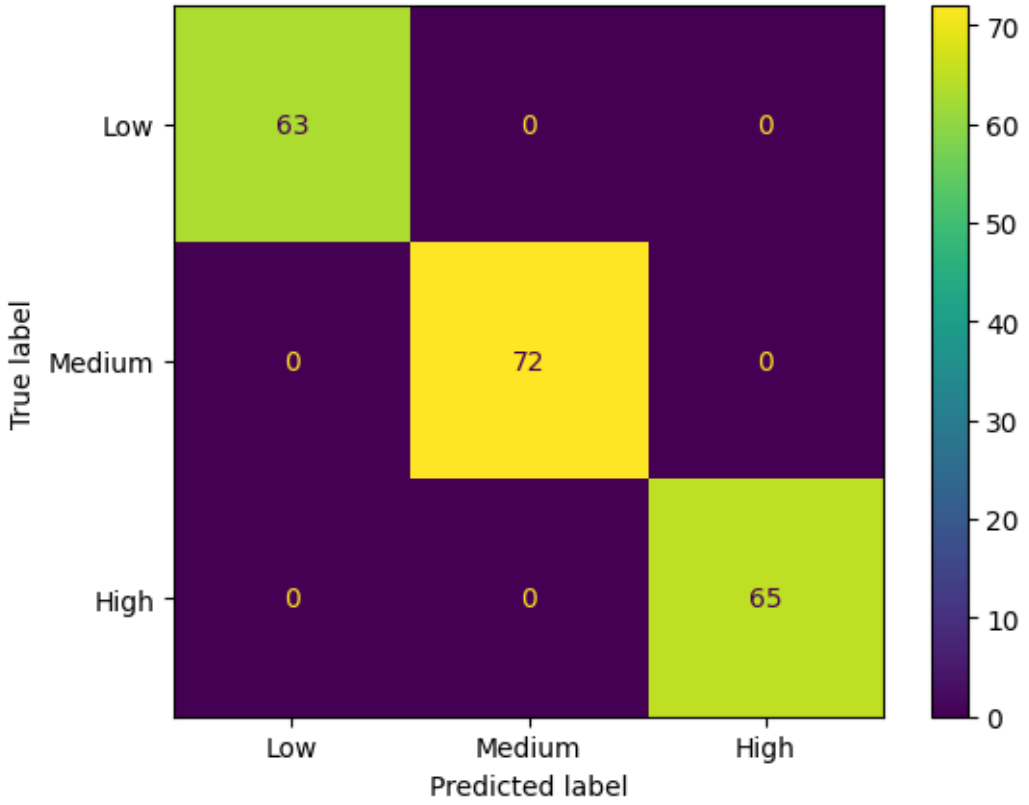

Figura 11: Matriz de confusión para el RF

Basándonos en los resultados podemos decir que el desempeño del modelo implementado fue perfecto.
Ahora procedemos a realizar el impacto de las variables en los resultados del modelo. En este caso podemos visualizar una mayor variedad de gráficos ya que *SHAP* ofrece un método optimizado para random forest que cuenta con mas gráficos disponibles para el análisis. Por ejemplo podemos obtener un gráfico del impacto de las variables para las clasificaciones de cada posible valor de la etiqueta. A continuación se presenta del gráfico de los valores SHAP medios del impacto de las variables en las clasificaciones etiquetadas como 0, es decir, un nivel bajo de propensión a la enfermedad.

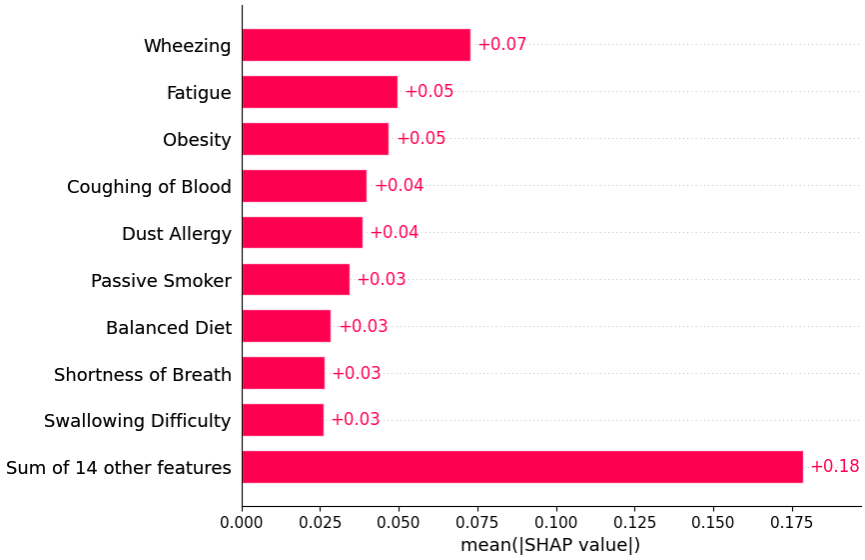

Figura 12: Promedio de los valores SHAP etiqueta 0

En el gráfico se muestra que las 5 variables que mas peso tienen al momento en el que el modelo clasifica al paciente con propensión baja son *Whezzing*, *Fatigue*,*Obesity*,*Coughing of Blood* y *Dust Allergy*

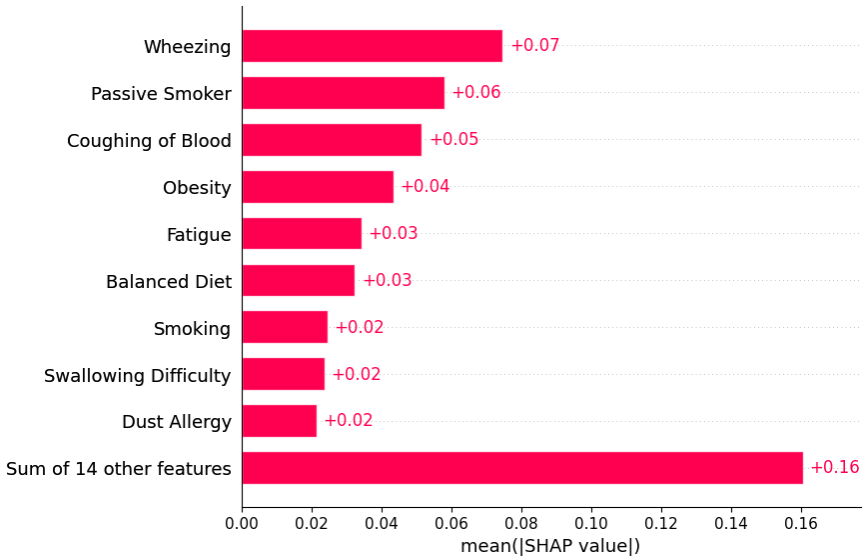

Figura 13: Promedio de los valores SHAP etiqueta 1

Para los pacientes etiquetados con un nivel medio de propensión a la enfermedad tenemos que las 5 variables que mas impacto tuvieron son *Whezzing*, *Pasive Smoker*,*Coughing of Blood*,*Obesity*, y *Fatigue*.

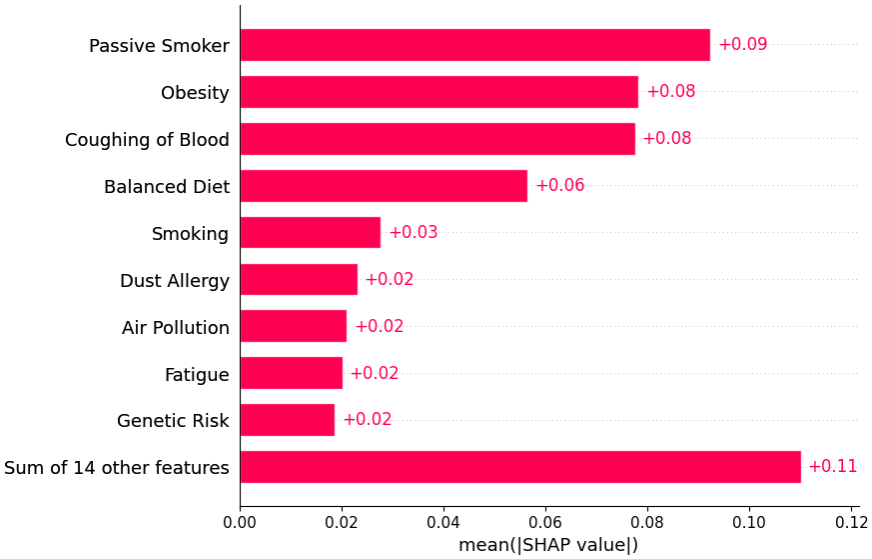

Figura 14: Promedio de los valores SHAP etiqueta 2

Para el caso de los pacientes diagnosticados con alta propensión a la enfermedad tenemos que las 5variables que mas peso tuvieron en el resultado de esta clasificación fueron *Pasive Smoker*,*Obesity*, *Coughing of Blood*, *Balance Diet* y *Smoking*.
Con esta herramienta también podemos ver como el valor que tome cierta variable afecta su probabilidad de ser clasificada en distintos niveles de propensión. Por ejemplo, utilicemos las variable *Obesity* para ver como los valores que toma afectan la probabilidad de clasificación en los niveles de propensión bajo y alto.

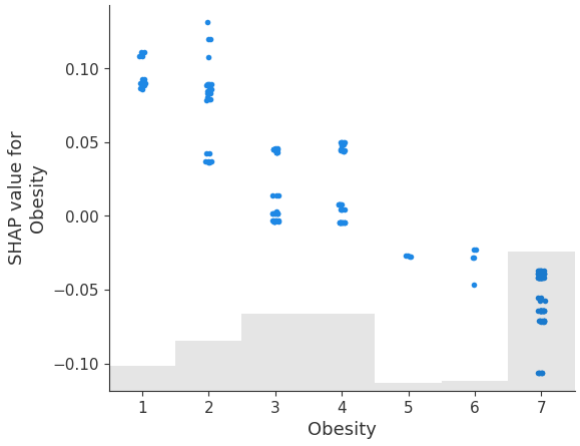

Figura 15: Efecto valor Obesity en valores SHAP etiqueta 0

Podemos ver como a medida que aumenta el valor de *Obesity* en el individuo decrece el valor SHAP de impacto de la variable para que el individuo sea clasificado con un nivel bajo de propension.

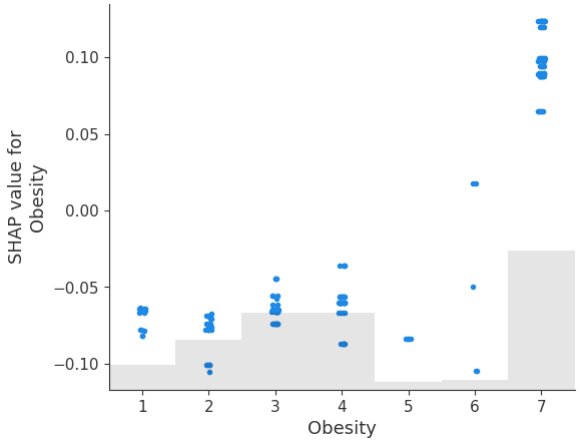

Figura 16: Efecto valor Obesity en valores SHAP etiqueta 2

Por otro lado cuando analizamos como afecta el valor de la variables *Obesity* el valor SHAP de impacto para que el individuo sea clasificado con propensión alta, a medida que aumenta el valor de la variable aumenta el valor SHAP. A continuación se presenta un grafico en el que se ve el promedio de los valores SHAP del impacto de cada variable en cada tipo de diagnostico

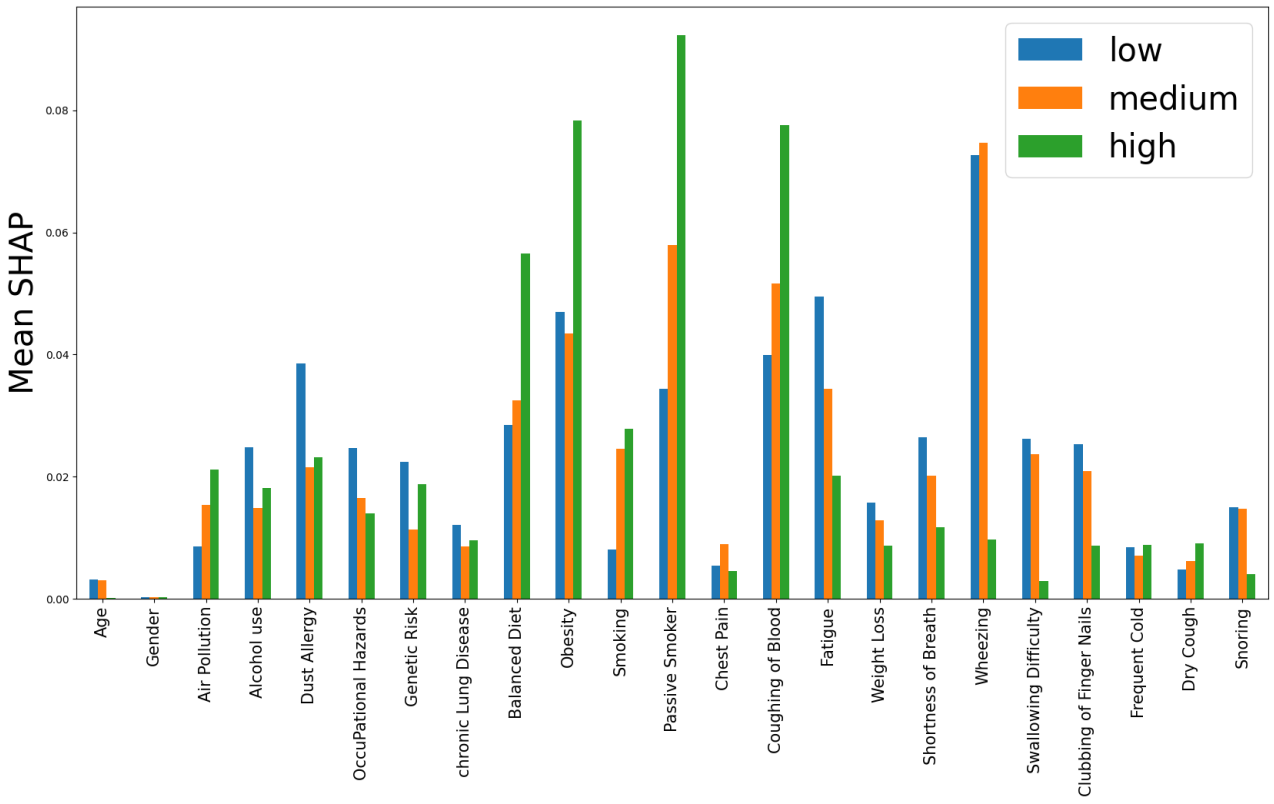

Figura 17: Promedio valores SHAP para cada tipo de diagnostico

Al igual que con el modelo SVM también podemos ver el impacto de las variables los diagnósticos realizados por el modelo desde una perspectiva general

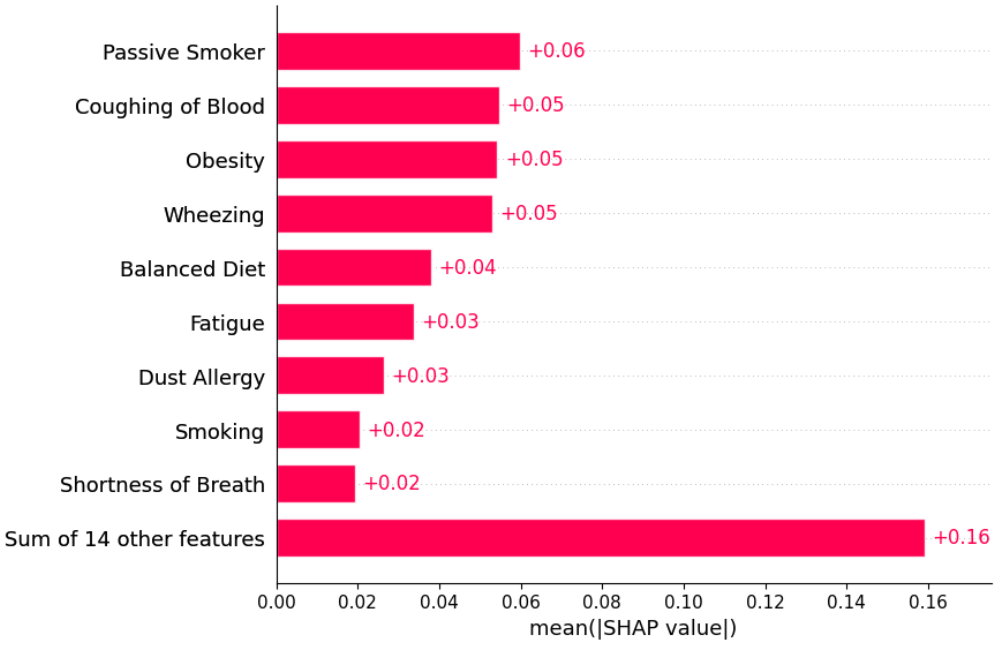

Figura 18: Promedio valores SHAP impacto en el modelo

Podemos visualizar que las variables que mas impacto tienen en los resultados ofrecidos por el modelo son *Pasive Smoker*, *Coughing of Blood*, *Obesity*, *Whezzing* y *Balance Diet*.

5.4. KNN

Como ultimo modelo, se entreno un modelo KNN para la clasificación. Similar a los modelos anteriores, se uso *Optuna* para optimizar el valor de los hiperparámetros, que para este modelo se considero el valor de k, que corresponde a la cantidad de vecinos, y se tomo valores de 1 a 20. De los ensayos hechos por *Optuna*, se obtuvo que el mejor valor de k fue de 5, con un f1 score de 1, sin embargo, como con los anteriores, esta no fue el único valor con f1 score de 1 pero fue el primero que encontró. A continuación, se muestra una gráfica del historial de optimización hecha por *Optuna*, la cual muestra los ensayos y el valor en el objetivo que corresponde al f1 score.

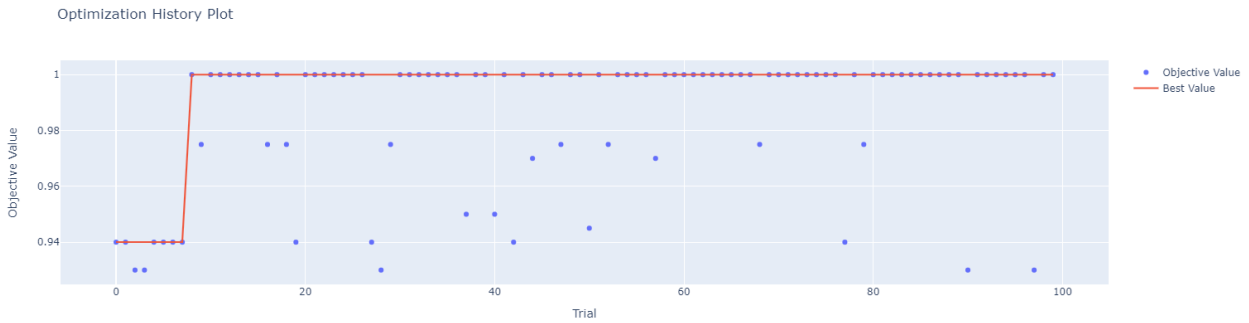

Figura 19: Historial de optimización para el KNN

Como con los modelos anteriores, de la figura, podemos ver que hubo muchos de los ensayos en los que se obtuvo un valor de 1 para el f1 score. Por tanto, se hizo también una gráfica para ver la relación entre los valores de los parámetros y la función objetivo.

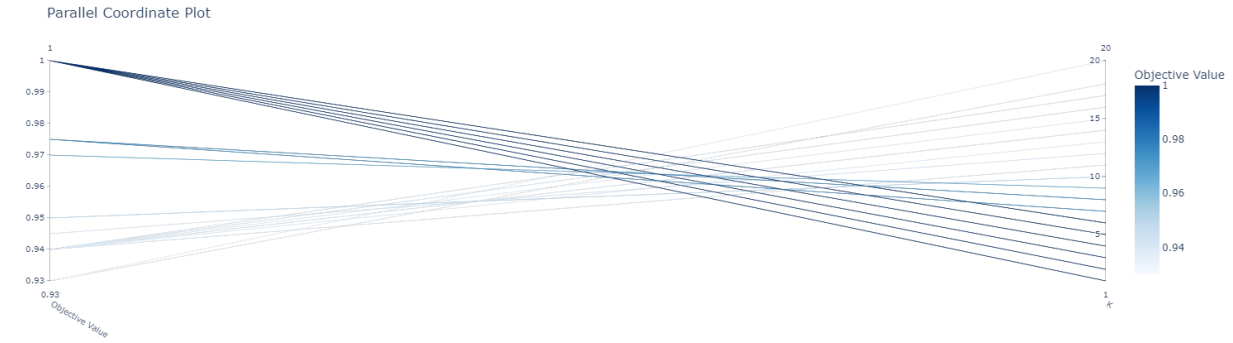

Figura 20: Parallel Coordinate Plot para la KNN

Para el KNN, podemos ver que valores altos de k resulta en un menor f1 score en el entrenamiento, mientras que los valores bajos resultan en un f1 score mayor, sin embargo, no se puede tomar un valor demasiado bajo ya que puede llevar a un sobreajuste del modelo.

Finalmente, con el valor de k=5, se hizo una validación del modelo con el conjunto de validación y se obtuvo los siguientes resultados.

	precision	recall	f1-score	support
Low	1.00	1.00	1.00	63
Medium	1.00	1.00	1.00	72
High	1.00	1.00	1.00	65
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Figura 21: Métricas de los resultados para el KNN

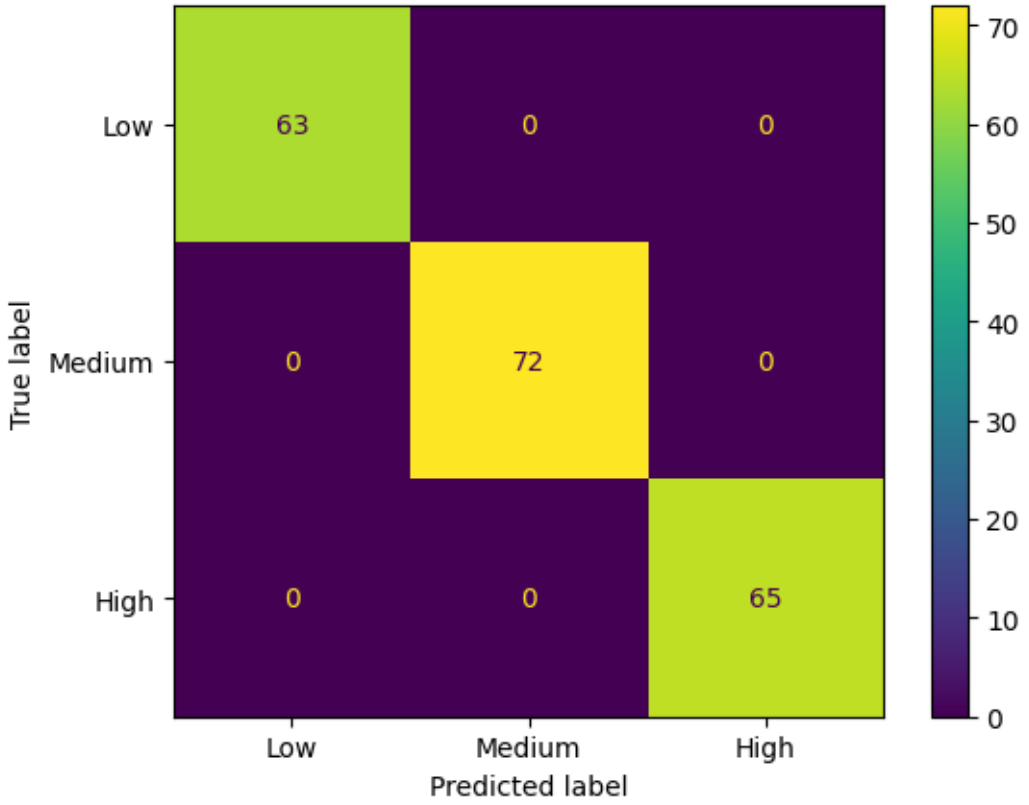

Figura 22: Matriz de confusión para el KNN

Con base en estos resultados, podemos decir que el modelo desarrollado también es perfecto, pues no se equivoca en las predicciones y las métricas para cada clase y sus promedios son 1, lo que indica que no se equivoca.
Ahora procederemos a estudiar el impacto de las variables en las predicciones del modelo utilizando de nuevo la librería *SHAP*. Inicialmente veremos el impacto promedio en cada uno de los tipos de diagnostico.

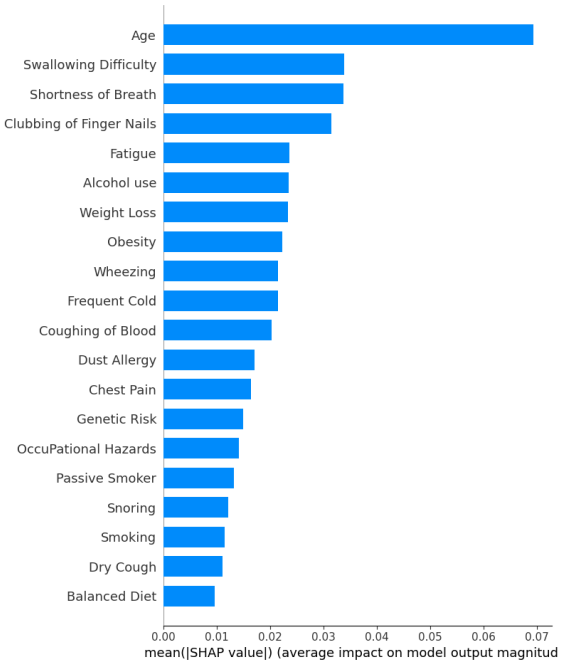

Figura 23: Impacto promedio en la salida del modelo etiqueta 0

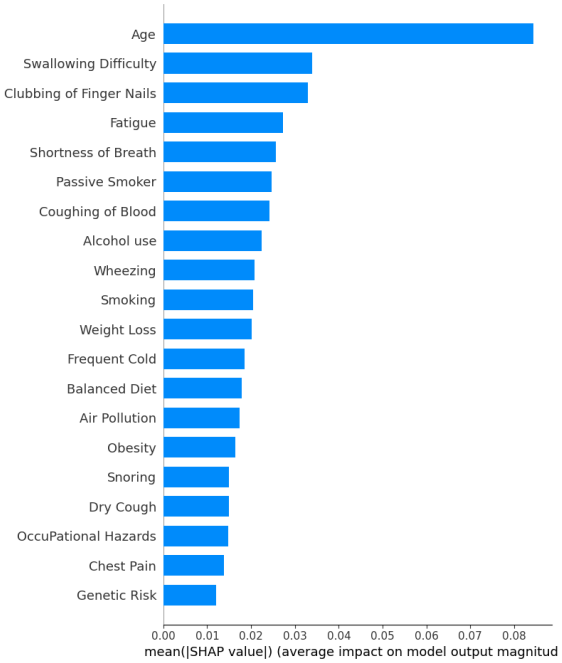

Figura 24: Impacto promedio en la salida del modelo etiqueta 1

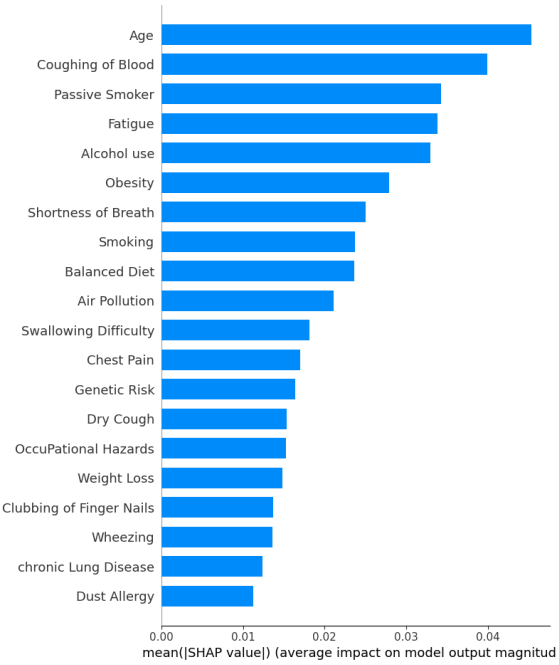

Figura 25: Impacto promedio en la salida del modelo etiqueta 2

Ahora veremos el impacto de las variables en el modelo en general.

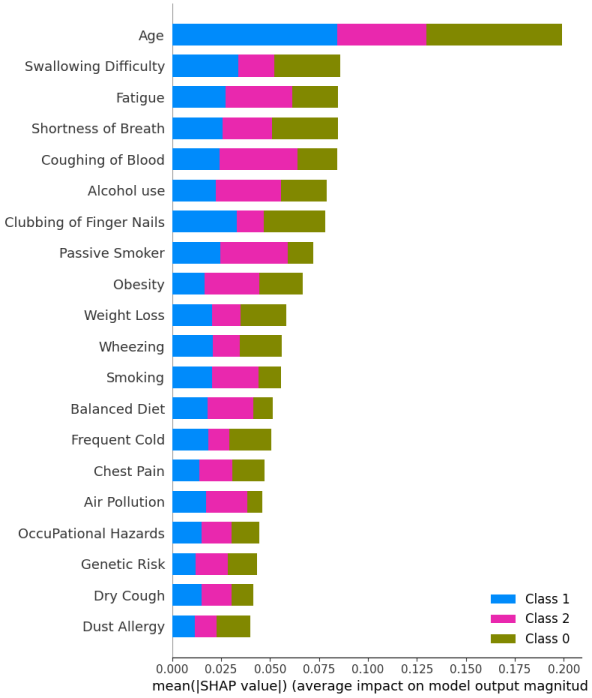

Figura 26: Impacto promedio en la salida del modelo

Podemos ver que en promedio las 5 variables que mas impacto tienen en la salida del modelo son *Age*, *Swallowing Difficulty*, *Fatigue*, *Shortness of Breath*, *Coughing of Blood*

5.5. Transformación en los Datos

Ahora, se quiere evaluar el efecto que puede tener la transformación en los datos sobre los resultados en los modelos. Para transformar los datos, primero se hizo una estandarización de esto, esto es, se transforman los datos para que tengan media cero y desviación estándar 1, lo cual se hace restándole a cada dato la media y dividiéndolo por la desviación estándar. Esta media y desviación estándar se calculan con base en los datos de entrenamiento y se usan para estandarizar los tres conjuntos de datos.

Después de estandarizar, se hizo un análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos. Este análisis se hizo sobre los datos de entrenamiento. Primero, vemos que tanta varianza explica cada una de las componentes y la explicada al acumular dichas componentes, para ver cuantas escogemos, esto se muestra en la siguiente gráfica:

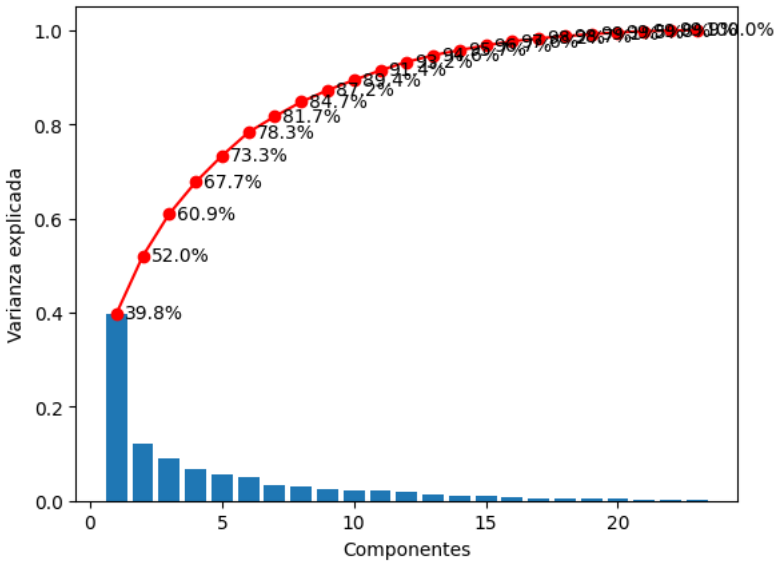

Figura 27: Varianza explicada por cada componente y acumulada

Se decidió escoger 4 componentes, lo que corresponde a explicar aproximadamente un 68 % de la varianza, para ver que ocurre al llevar los datos a un espacio de 4 dimensiones, es decir, reducir la dimensión de 20 a 4. Además, esto ayuda a eliminar la correlación entre las variables, lo que elimina variables redundantes. Con base en la transformación hecha por PCA a el conjunto de entrenamiento, se hace para el conjunto de prueba y validación.

5.5.1. SVM

Se entreno primero una maquina de vector soporte como se hizo anteriormente, utilizando *Optuna* para seleccionar los hiperparámetros. Luego de seleccionar los mejores hiperparámetros, se valido el modelo con el conjunto de validación y se obtuvo lo siguiente:

	precision	recall	f1-score	support
Low	1.00	1.00	1.00	63
Medium	1.00	1.00	1.00	72
High	1.00	1.00	1.00	65
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Figura 28: Métricas de los resultados para el KNN

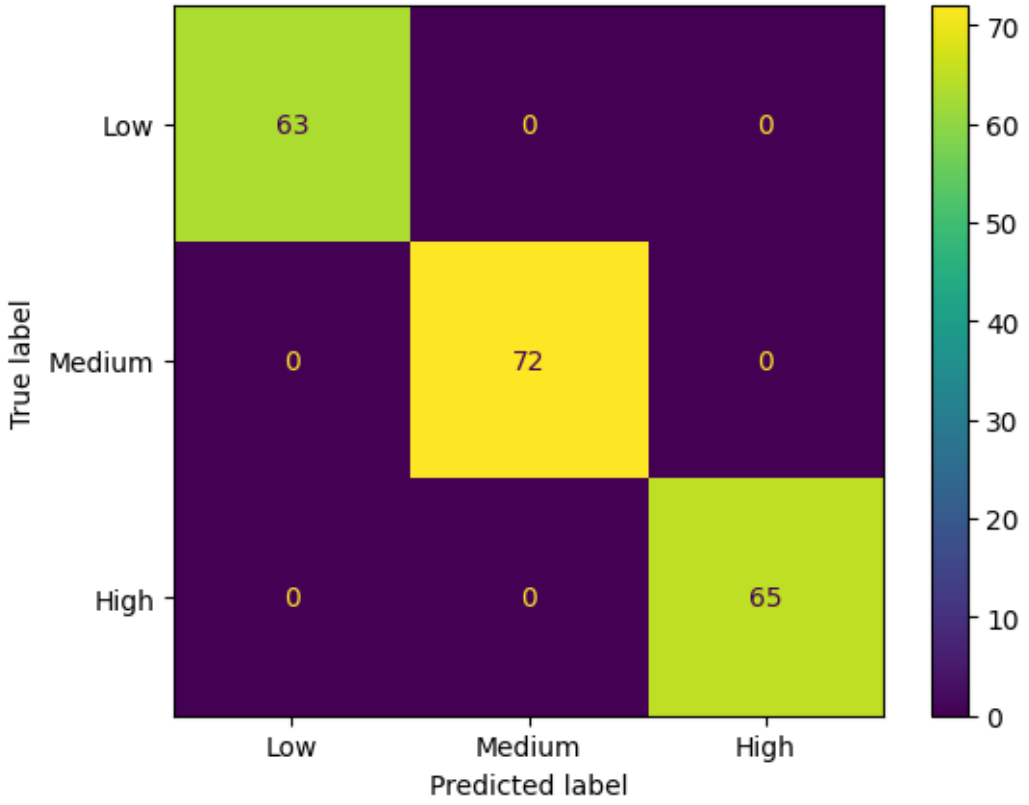

Figura 29: Matriz de confusión para el KNN

Como se puede ver en los resultados, no hay cambios con respecto a los datos sin la transformación, lo que indica que la transformación de estos datos no afecta negativamente los resultados.

5.5.2. Random Forest

Luego, se entreno un random forest como se hizo anteriormente, utilizando *Optuna* para seleccionar los hiperparámetros. Luego de seleccionar los mejores hiperparámetros, se valido el modelo con el conjunto de validación y se obtuvo lo siguiente:

	precision	recall	f1-score	support
Low	1.00	1.00	1.00	63
Medium	1.00	1.00	1.00	72
High	1.00	1.00	1.00	65
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Figura 30: Métricas de los resultados para el KNN

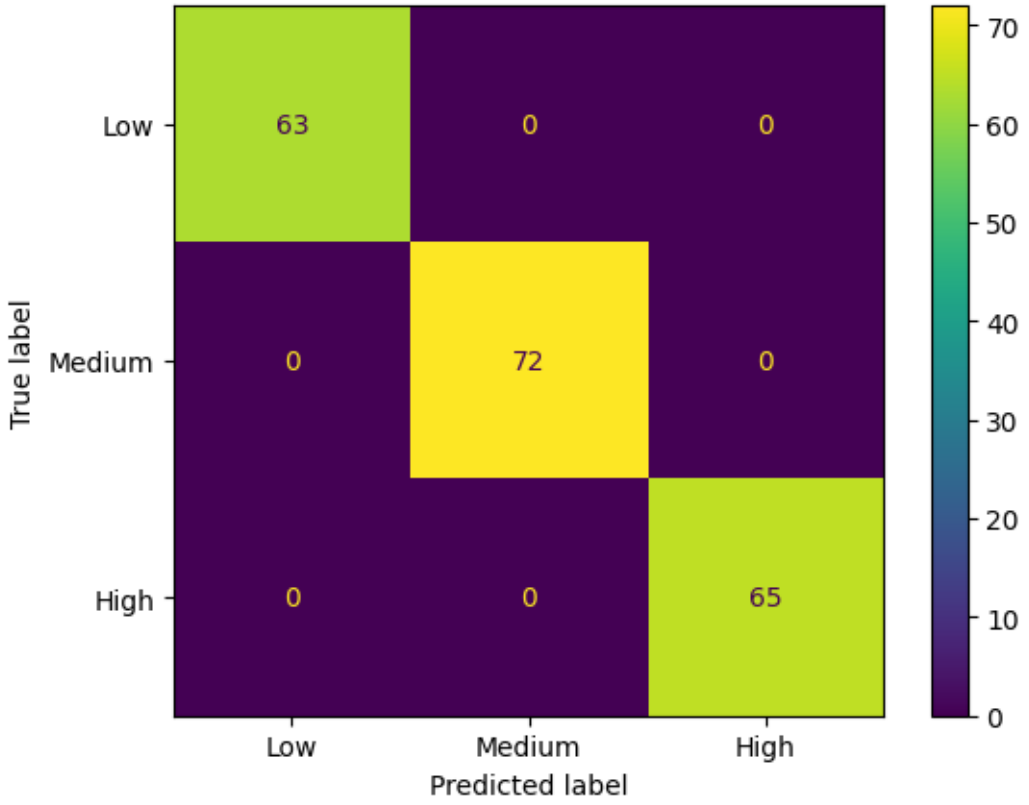

Figura 31: Matriz de confusión para el KNN

Similar a con la SVM, se puede ver en los resultados que no hay cambios con respecto a los datos sin la transformación, lo que indica que la transformación de estos datos no afecta negativamente los resultados.

5.5.3. KNN

Finalmente, se entreno un modelo KNN, utilizando *Optuna* para seleccionar los hiperparámetros.Se selecciono los mejores hiperparámetros y se valido el modelo con el conjunto de validación, obteniendo lo siguiente:

	precision	recall	f1-score	support
Low	1.00	1.00	1.00	63
Medium	1.00	1.00	1.00	72
High	1.00	1.00	1.00	65
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Figura 32: Métricas de los resultados para el KNN

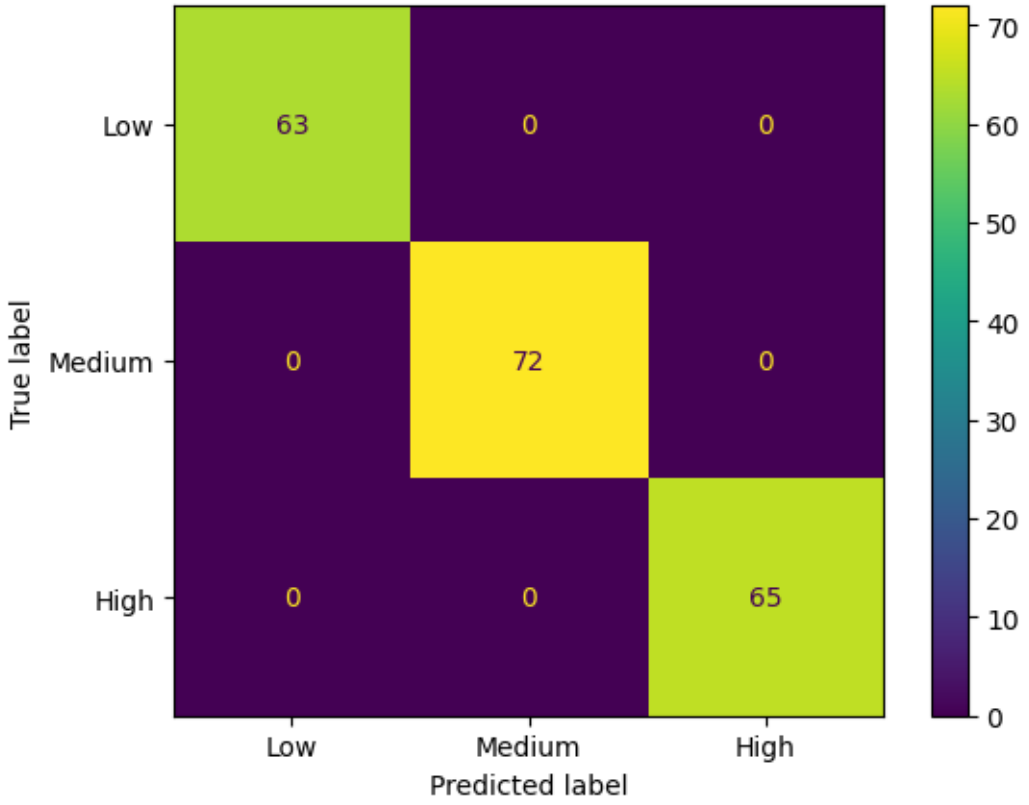

Figura 33: Matriz de confusión para el KNN

Como con los modelos anteriores, no hay cambios con respecto a los datos sin la transformación, lo que indica que la transformación de estos datos no afecta negativamente los resultados.

6. Conclusiones

Respecto al rendimiento de los modelos, los tres modelos demostraron un excelente rendimiento en la precisión de sus predicciones. Creemos que principalmente esto se debe a la calidad de los datos utilizados. Es posible que las variables seleccionadas al momento de construir el dataset hubieran estudiadas por expertos en el campo.

Con respecto a las transformaciones en los datos, podemos ver que transformar los datos para este conjunto no tiene ningún efecto negativo sobre los resultados, esto resulta ser beneficioso en la práctica si se trabaja a gran escala, ya que esto nos indica que podemos almacenar los datos con únicamente 4 variables con base en la transformación hecha, en lugar de 20, lo que resulta en mucho menos gasto en memoria, y aun así obteniendo buenos resultados.

Con respecto a la intratabilidad de los modelos podemos visualizar las variables que mas impacto tienen en las predicciones, algunas variables que presentaron en común son las siguientes: *Coughing of Blood*, *Obesity*, *Pasive Smoker*, y *Fatigue*. Estos resultados pueden ser útiles tanto para posibles estudios llevados por investigadores que busquen mejorar la forma de diagnosticar el cáncer de pulmón y también para que los pacientes tengan estas se relaciona su estilo de vida con estas variables, esto puede mejorar el enfoque de pretensión de este tipo de enfermedad.

7. Implicaciones Éticas

Identificamos las siguientes implicaciones éticas en el manejo de datos personales de salud, como en el uso de algoritmos de aprendizaje automático en la clasificación de enfermedades pulmonares crónicas:

- **Sesgo y Equidad en los Modelos de Aprendizaje Automático:** Identificaremos y mitigaremos posibles sesgos en los datos y los algoritmos, con el fin de evitar discriminación o injusticias con la clasificación de los pacientes.
- **Interpretabilidad y Explicabilidad de los Modelos:** Priorizaremos la interpretabilidad y explicabilidad de los modelos de aprendizaje automático para garantizar que las decisiones resultantes sean comprensibles y justificables tanto para los profesionales de la salud como para los pacientes. Además, proporcionaremos explicaciones sobre cómo realizaremos las predicciones y cómo se utilizaremos los diferentes atributos de los datos para llegar a conclusiones sobre la enfermedad pulmonar crónica.
- **Responsabilidad Social y Impacto en la Salud Pública:** Consideraremos los posibles impactos sociales y de salud pública de la implementación de los resultados de la investigación, asegurando que las conclusiones y recomendaciones sean éticamente responsables y beneficiosas para la sociedad en su conjunto.

8. Aspectos Legales y Comerciales

Nuestro proyecto se compromete a cumplir con todas las regulaciones legales pertinentes, incluidas pero no limitados a las normativas de protección de datos en salud. Estableceremos un marco de propiedad intelectual para proteger los resultados de nuestra investigación considerando sus posibles usos comerciales. Implementaremos medidas de seguridad de datos y garantizaremos el cumplimiento normativo y ético en todas las etapas del proyecto.

Referencias

OMS. 2023. *Cáncer de pulmón*. Accedido el: 26 de junio de 2023.

THE DEVASTATOR. 2022. *Lung Cancer Prediction*. Recuperado el 27 de octubre de 2023, de `https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/code`.