



MACHINE
LEARNING

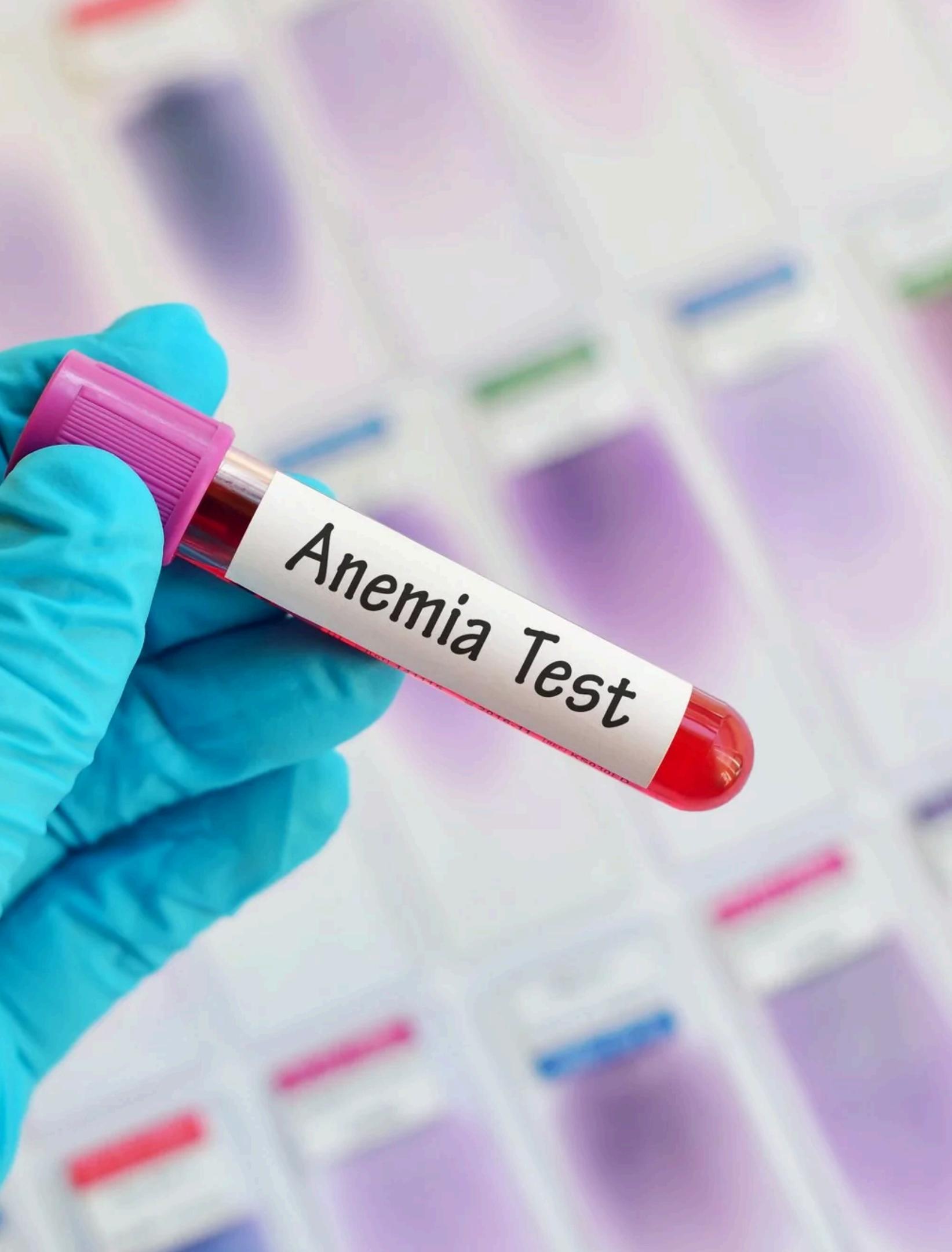
Clasificación de tipo de anemias

Análisis y datos sobre los
diagnósticos de anemia.

Integrantes:

Oscar Alejandro Castillo Naveda
Angie Katherine Gonzalez Gonzales
Mateo Andrade Ballen





Beneficios de una temprana detección de la Anemia

- Inicio rápido del tratamiento.
- Prevención de complicaciones graves.
- Mejora en la energía y la función física.
- Mejora en la función cognitiva y la calidad de vida.
- Menor riesgo de hospitalización y complicaciones de salud.
- Reducción de costos de atención médica.

¿Qué es un diagnóstico de la anemia?

La anemia es una condición en la que el cuerpo no tiene suficientes glóbulos rojos sanos o hemoglobina, lo que resulta en una capacidad reducida para transportar oxígeno a los tejidos.

El diagnóstico de la anemia implica:

1. Historia clínica y síntomas: Identificar síntomas como fatiga, debilidad, palidez, y dificultad para respirar.
2. Examen físico: Evaluar signos físicos de anemia.
3. Pruebas de laboratorio: Realizar un hemograma completo (CBC) para medir los niveles de hemoglobina y el recuento de glóbulos rojos, y pruebas adicionales para determinar la causa subyacente, como la medición de hierro sérico, ferritina, vitamina B12 y folato.



Objetivos del Proyecto

Desarrollar un modelo de Machine Learning capaz de predecir con alta precisión la presencia de enfermedades asociadas a la anemia utilizando datos clínicos y demográficos de los pacientes.



Recolectar y preparar datos clínicos relevantes

Obtener y limpiar un conjunto de datos que incluya información relevante sobre pacientes, como niveles de hemoglobina, hematocrito, hierro sérico, vitamina B12, ácido fólico, y otros marcadores hematológicos y demográficos.

Desarrollar y entrenar modelos de Machine Learning:

Probar diversos algoritmos de Machine Learning (por ejemplo, regresión logística, árboles de decisión, redes neuronales, entre otros) para identificar cuál ofrece la mejor precisión en la predicción de enfermedades asociadas a la anemia.

Validar y evaluar el rendimiento del modelo:

Implementar técnicas de validación cruzada y otras métricas de rendimiento (precisión, sensibilidad, especificidad, AUC-ROC) para evaluar la efectividad del modelo y asegurar su capacidad de generalización a nuevos datos.

Implementar el modelo en un sistema accesible para profesionales de la salud:

Desarrollar una interfaz de usuario intuitiva y fácil de usar que permita a los profesionales de la salud ingresar datos de los pacientes y obtener predicciones sobre el diagnóstico de enfermedades asociadas a la anemia en tiempo real.

Diccionario de datos:

diagnosticado_cbc_data_v4.csv (125,59 kB)											
Detalle		Compacto		Columnas							
#	LYMP	#	NEUTp	#	LYMn	#	NEUTn	#	eritrocitos	#	+
45,7	8,2	91,4	0,7	5.32k	0,2	41,8	0,5	79	1,98	90,8	1
25,845	77,511	1.88876	5.14094	5,7	25,845	77,511	1.88876	5.14094	4,5	25,845	77,511
25,845	77,511	1.88876	5.14094	5,2	25,845	77,511	1.88876	5.14094	4,9	25,845	77,511
25,845	77,511	1.88876	5.14094	5,2	25,845	77,511	1.88876	5.14094	4,7	25,845	77,511
25,845	77,511	1.88876	5.14094	5,2	25,845	77,511	1.88876	5.14094	4,2	25,845	77,511
25,845	77,511	1.88876	5.14094	4,4	25,845	77,511	1.88876	5.14094	3,8	25,845	77,511
25,845	77,511	1.88876	5.14094	5,3	25,845	77,511	1.88876	5.14094	5,1	25,845	77,511
25,845	77,511	1.88876	5.14094	4,5	25,845	77,511	1.88876	5.14094	5,8	25,845	77,511
25,845	77,511	1.88876	5.14094	4,2	25,845	77,511	1.88876	5.14094	4,3	25,845	77,511
25,845	77,511	1.88876	5.14094	5,2	25,845	77,511	1.88876	5.14094	4,3	25,845	77,511
25,845	77,511	1.88876	5.14094	5,3	25,845	77,511	1.88876	5.14094	4,6	25,845	77,511
25,845	77,511	1.88876	5.14094	5,4	25,845	77,511	1.88876	5.14094	4,1	25,845	77,511
25,845	77,511	1.88876	5.14094	5,7	25,845	77,511	1.88876	5.14094	3,9	25,845	77,511
25,845	77,511	1.88876	5.14094	4,1	25,845	77,511	1.88876	5.14094	4,1	25,845	77,511

HGB La cantidad de hemoglobina en la sangre, crucial para el transporte de oxígeno.

PLT Número de plaquetas en la sangre, implicadas en la coagulación sanguínea.

WBC El recuento de glóbulos blancos, vital para la respuesta inmune.

RBC El recuento de glóbulos rojos, responsables del transporte de oxígeno.

MCV (Volumen Corpúscular Medio): Volumen promedio de un solo glóbulo rojo.

MCH (Hemoglobina Corpúscular Media): Cantidad promedio de hemoglobina por glóbulo rojo.

MCHC (Concentración media de hemoglobina corpúscular): Concentración media de hemoglobina en los glóbulos rojos.

PDW Una medida de la variabilidad en la distribución del tamaño de las plaquetas en la sangre

PCT una prueba de procalcitonina puede ayudar a su proveedor de atención médica a diagnosticar si tiene sepsis

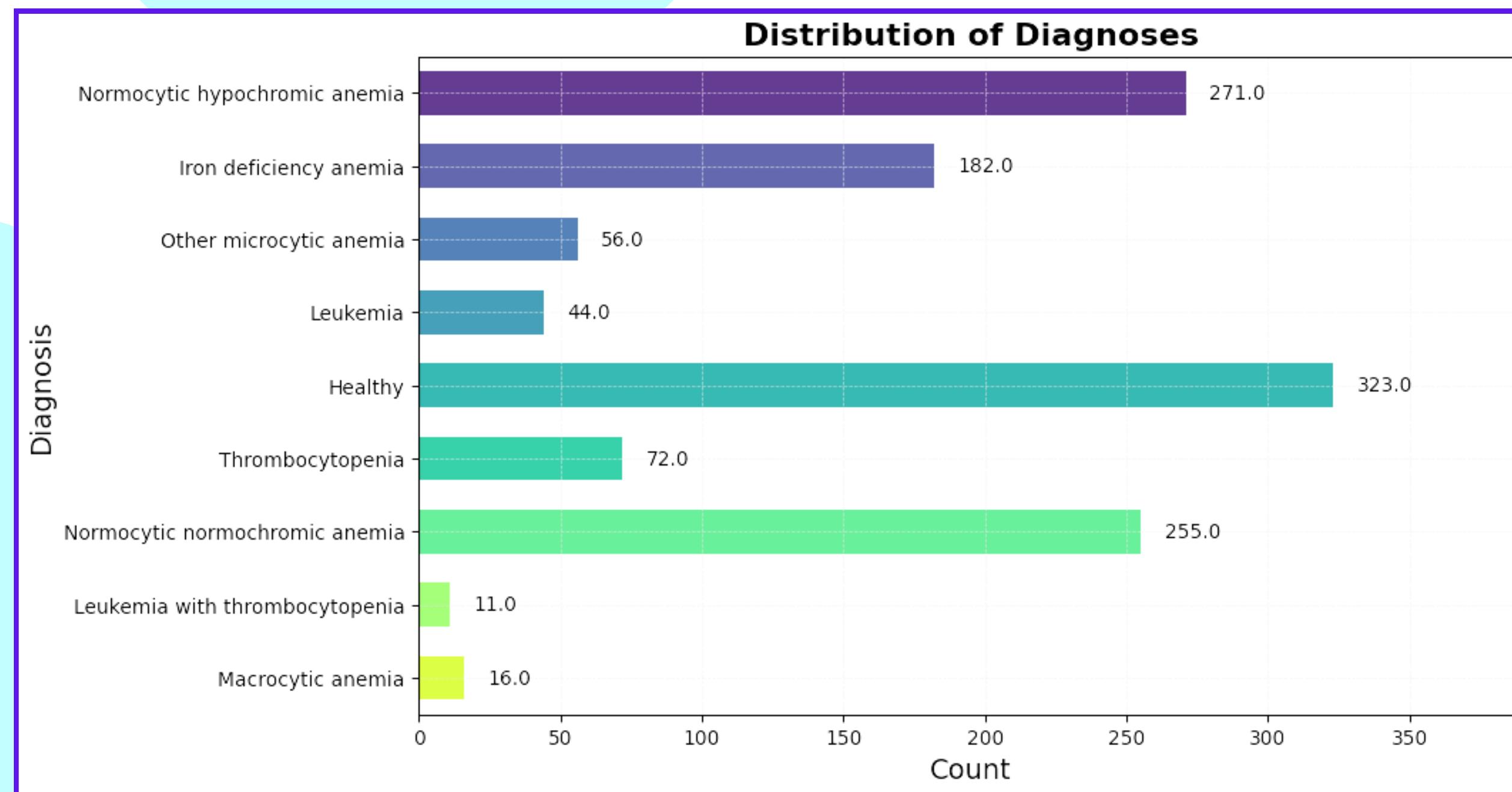
Diagnóstico Tipo de anemia según los parámetros del hemograma.

CARGA DE DATOS

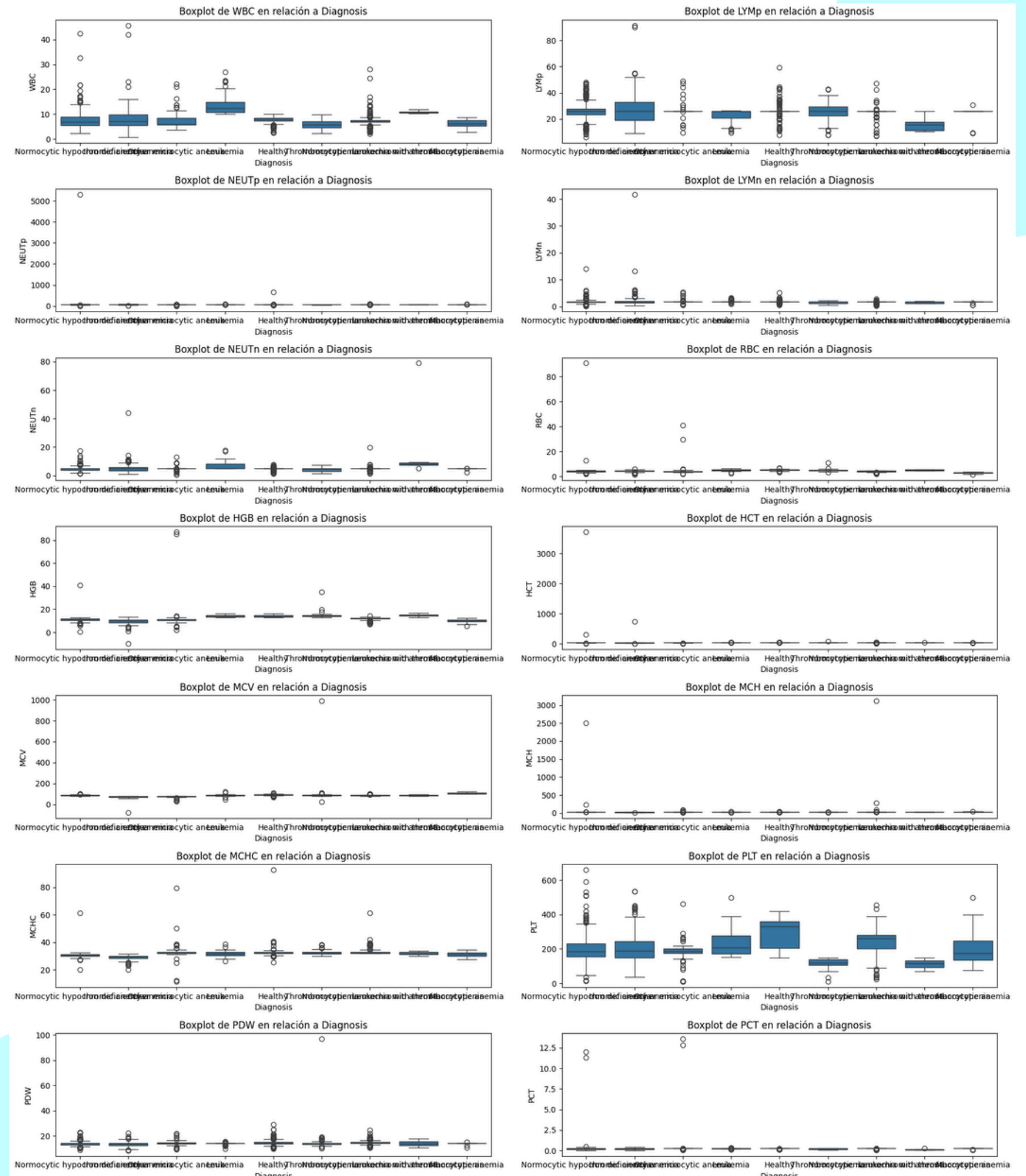
	WBC	LYMp	NEUTp	LYMn	NEUTn	RBC	HGB	HCT	MCV	MCH	MCHC	PLT	PDW	PCT	Diagnosis
0	10.00	43.200	50.100	4.30000	5.00000	2.77	7.3	24.2000	87.7	26.3	30.1	189.0	12.500000	0.17000	Normocytic hypochromic anemia
1	10.00	42.400	52.300	4.20000	5.30000	2.84	7.3	25.0000	88.2	25.7	20.2	180.0	12.500000	0.16000	Normocytic hypochromic anemia
2	7.20	30.700	60.700	2.20000	4.40000	3.97	9.0	30.5000	77.0	22.6	29.5	148.0	14.300000	0.14000	Iron deficiency anemia
3	6.00	30.200	63.500	1.80000	3.80000	4.22	3.8	32.8000	77.9	23.2	29.8	143.0	11.300000	0.12000	Iron deficiency anemia
4	4.20	39.100	53.700	1.60000	2.30000	3.93	0.4	316.0000	80.6	23.9	29.7	236.0	12.800000	0.22000	Normocytic hypochromic anemia
...	
1276	4.40	25.845	77.511	1.88076	5.14094	4.86	13.5	46.1526	80.7	27.7	34.4	180.0	14.312512	0.26028	Healthy
1277	5.60	25.845	77.511	1.88076	5.14094	4.85	15.0	46.1526	91.7	31.0	33.8	215.0	14.312512	0.26028	Healthy
1278	9.20	25.845	77.511	1.88076	5.14094	4.47	13.1	46.1526	88.7	29.3	33.0	329.0	14.312512	0.26028	Healthy
1279	6.48	25.845	77.511	1.88076	5.14094	4.75	13.2	46.1526	86.7	27.9	32.1	174.0	14.312512	0.26028	Healthy
1280	8.80	25.845	77.511	1.88076	5.14094	4.95	15.2	46.1526	89.7	30.6	34.2	279.0	14.312512	0.26028	Healthy

1230 rows × 15 columns

Cuadro para establecer la comparación entre los diferentes diagnósticos e identificar cuáles son los más comunes y cuáles menos frecuentes.



Se analiza cada variable hematológica a partir de la clasificación según el diagnóstico

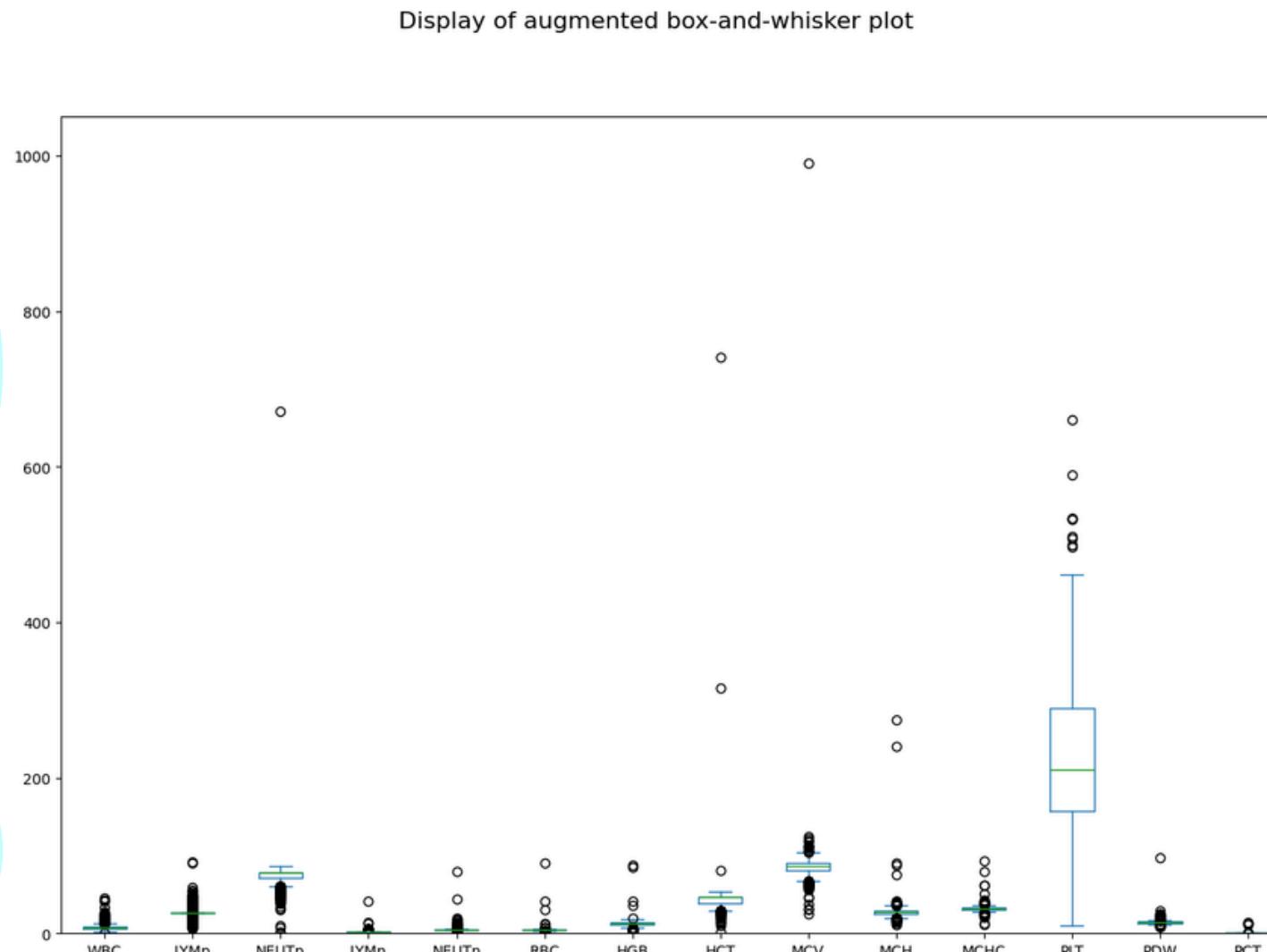


Para tratar con las variables numéricas, se transforma la variable cualitativa con la función LabelEncoder()

	WBC	LYMp	NEUTp	LYMn	NEUTn	RBC	HGB	HCT	MCV	MCH	MCHC	PLT	PDW	PCT	Diagnosis_INT
0	10.00	43.200	50.100	4.30000	5.00000	2.77	7.3	24.2000	87.7	26.3	30.1	189.0	12.500000	0.17000	5
2	7.20	30.700	60.700	2.20000	4.40000	3.97	9.0	30.5000	77.0	22.6	29.5	148.0	14.300000	0.14000	1
3	6.00	30.200	63.500	1.80000	3.80000	4.22	3.8	32.8000	77.9	23.2	29.8	143.0	11.300000	0.12000	1
6	16.70	19.100	68.200	3.20000	11.40000	5.15	14.2	44.8000	87.1	27.5	31.6	151.0	12.800000	0.14000	2
7	9.30	27.400	64.000	2.60000	5.90000	4.39	12.0	37.9000	86.4	27.3	31.6	194.0	15.900000	0.19000	5
...
1276	4.40	25.845	77.511	1.88076	5.14094	4.86	13.5	46.1526	80.7	27.7	34.4	180.0	14.312512	0.26028	0
1277	5.60	25.845	77.511	1.88076	5.14094	4.85	15.0	46.1526	91.7	31.0	33.8	215.0	14.312512	0.26028	0
1278	9.20	25.845	77.511	1.88076	5.14094	4.47	13.1	46.1526	88.7	29.3	33.0	329.0	14.312512	0.26028	0
1279	6.48	25.845	77.511	1.88076	5.14094	4.75	13.2	46.1526	86.7	27.9	32.1	174.0	14.312512	0.26028	0
1280	8.80	25.845	77.511	1.88076	5.14094	4.95	15.2	46.1526	89.7	30.6	34.2	279.0	14.312512	0.26028	0

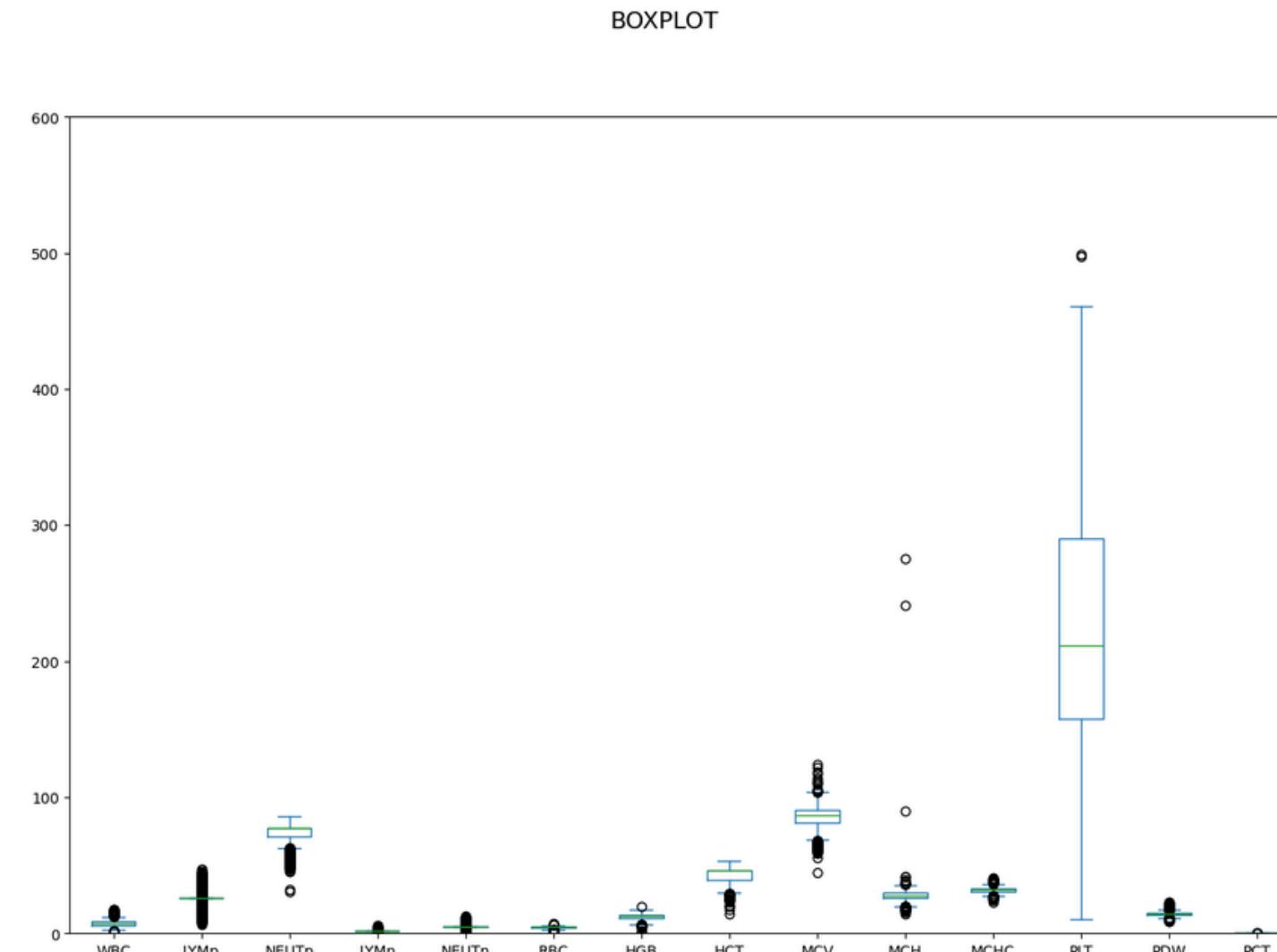
1.

Se realiza el análisis de outliers a través del gráfico de cajas y bigotes analizando cada variable hematológica.



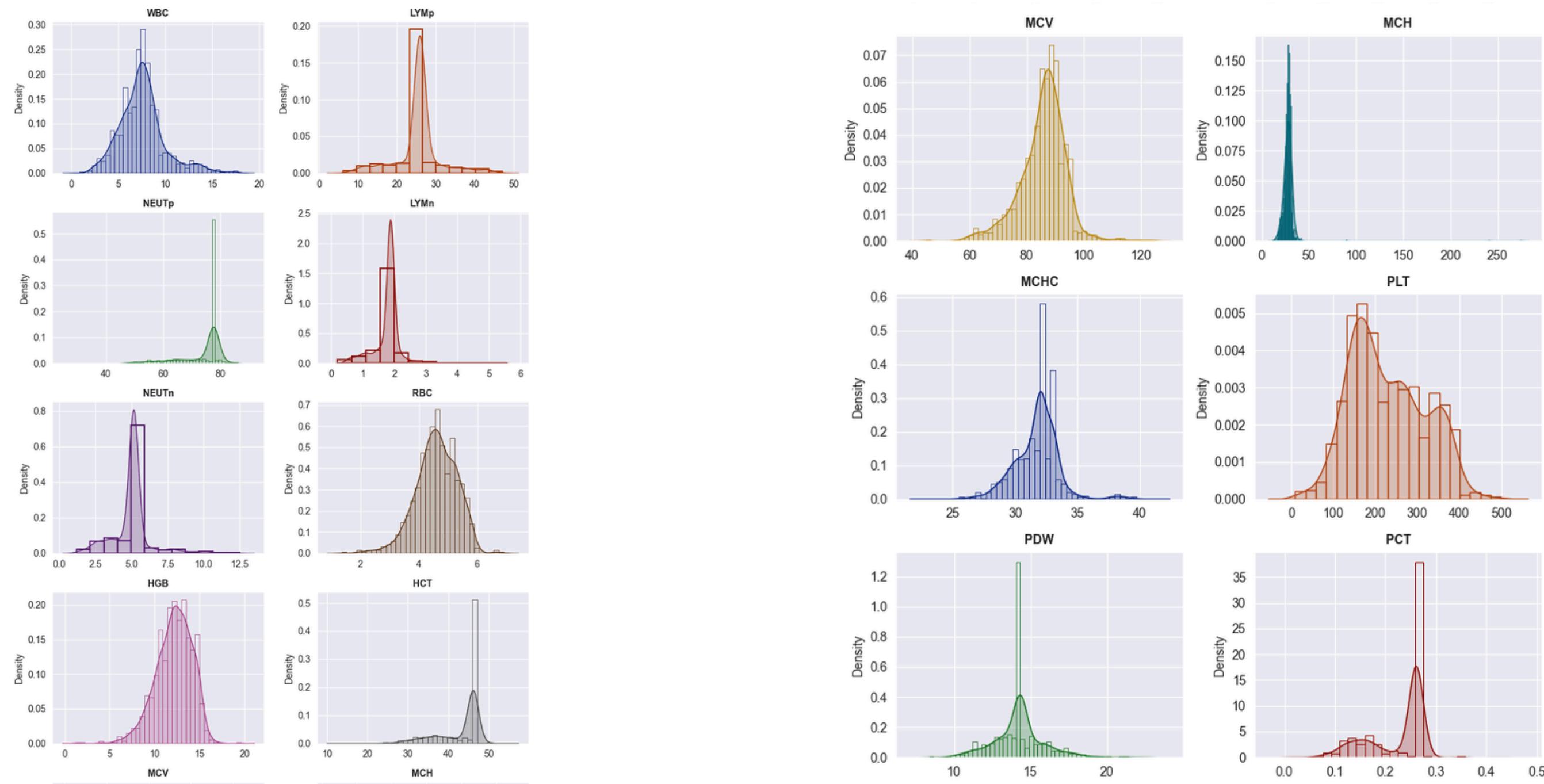
2.

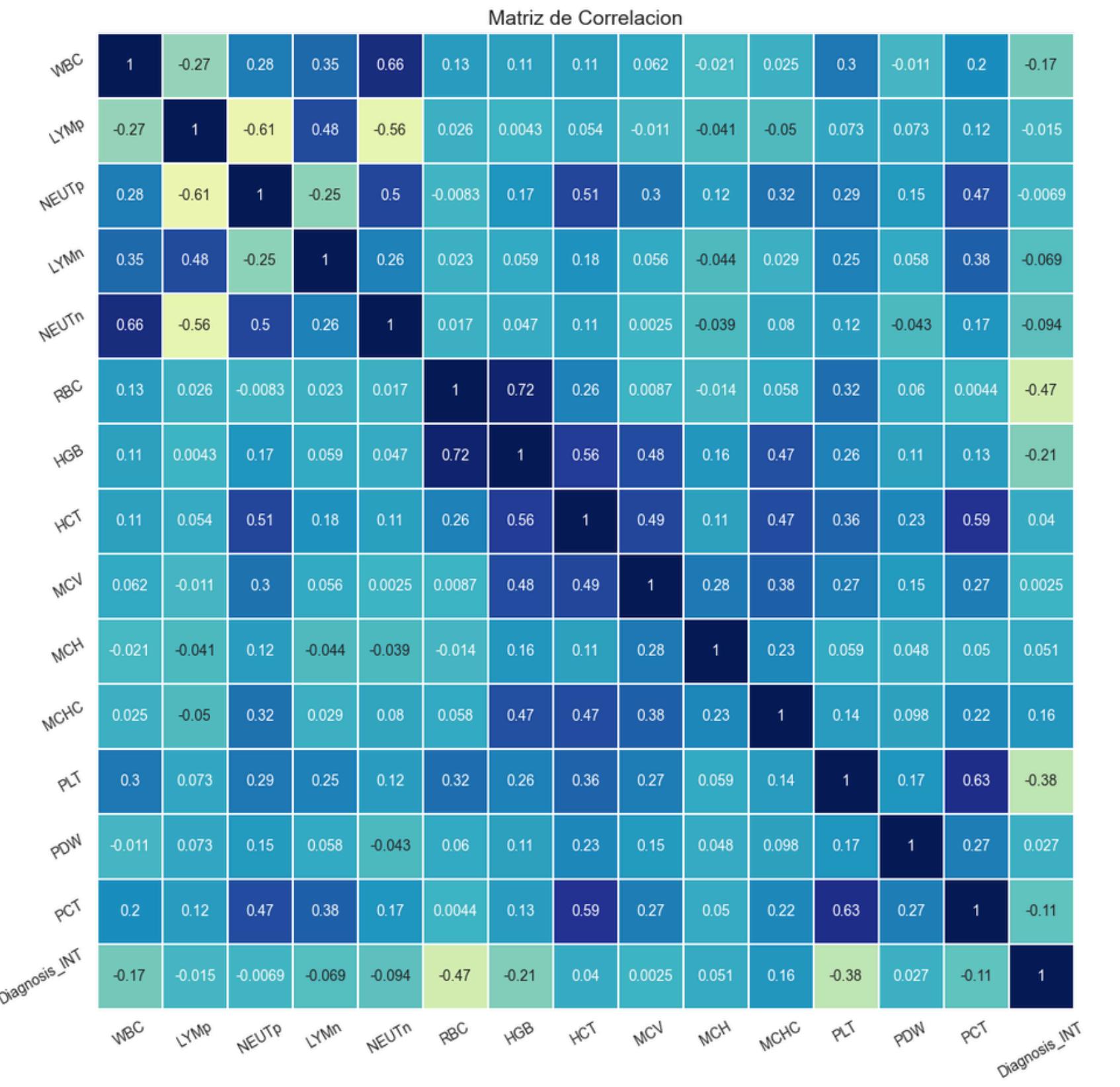
Se implementa el algoritmo z_score para el tratamiento de outliers



3.

Se analizan las distribuciones de las variables independientes



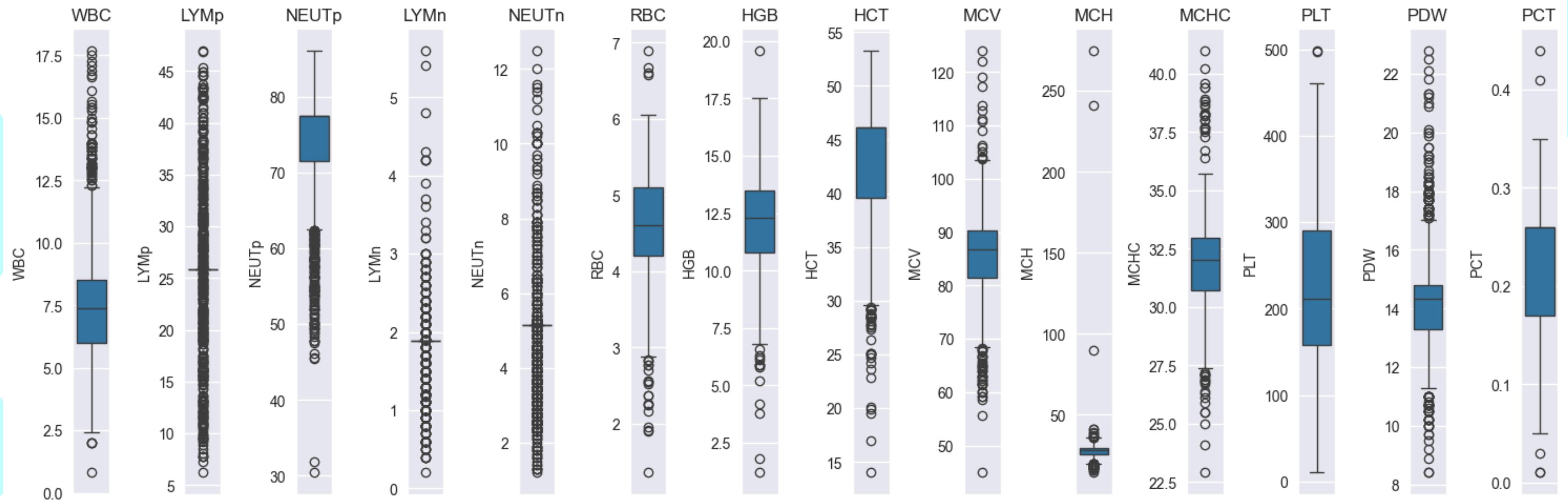


4.

Se busca alta correlación con el target, sin alcanzar el 0.6 con el resto de variables. Se observan que los mayores pesos corresponden a RBC y PLT. Sin embargo, este no es suficiente y se procederá a tratar los datos buscando adquirir la mayor cantidad de variables finales para el modelamiento.

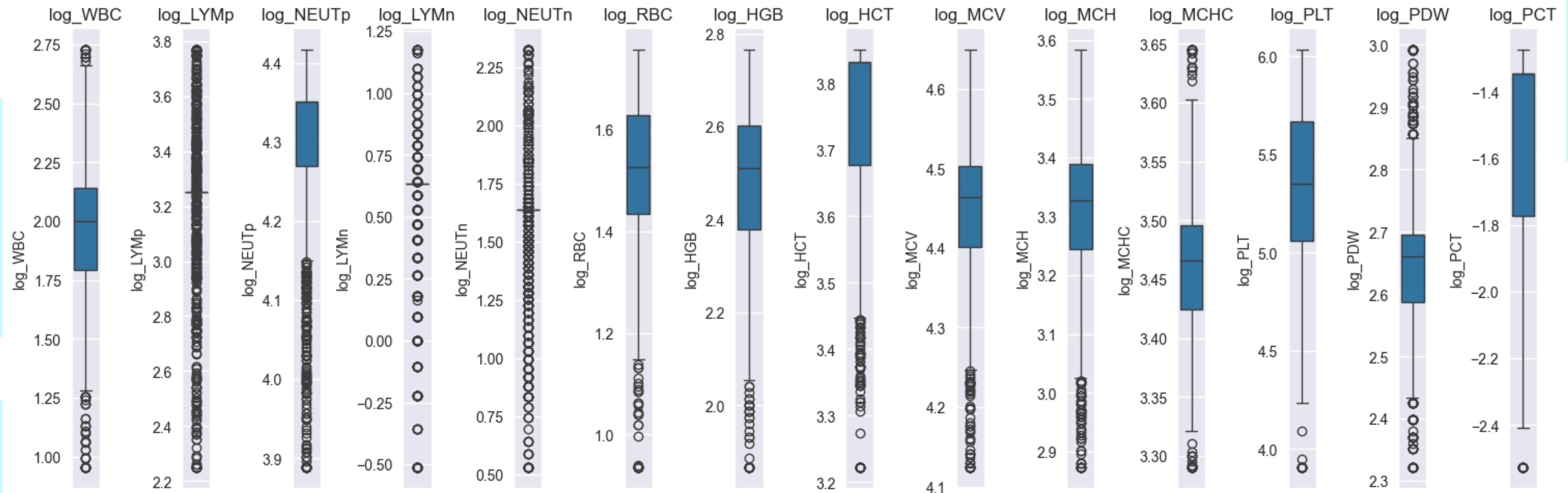
5.

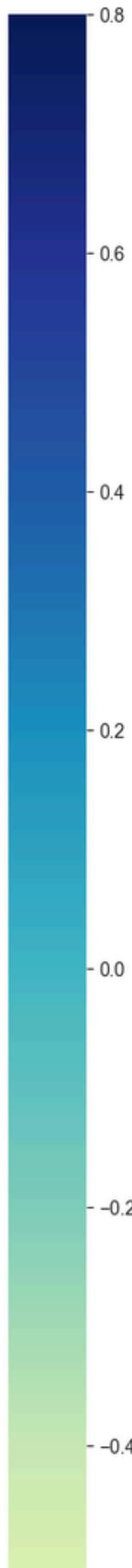
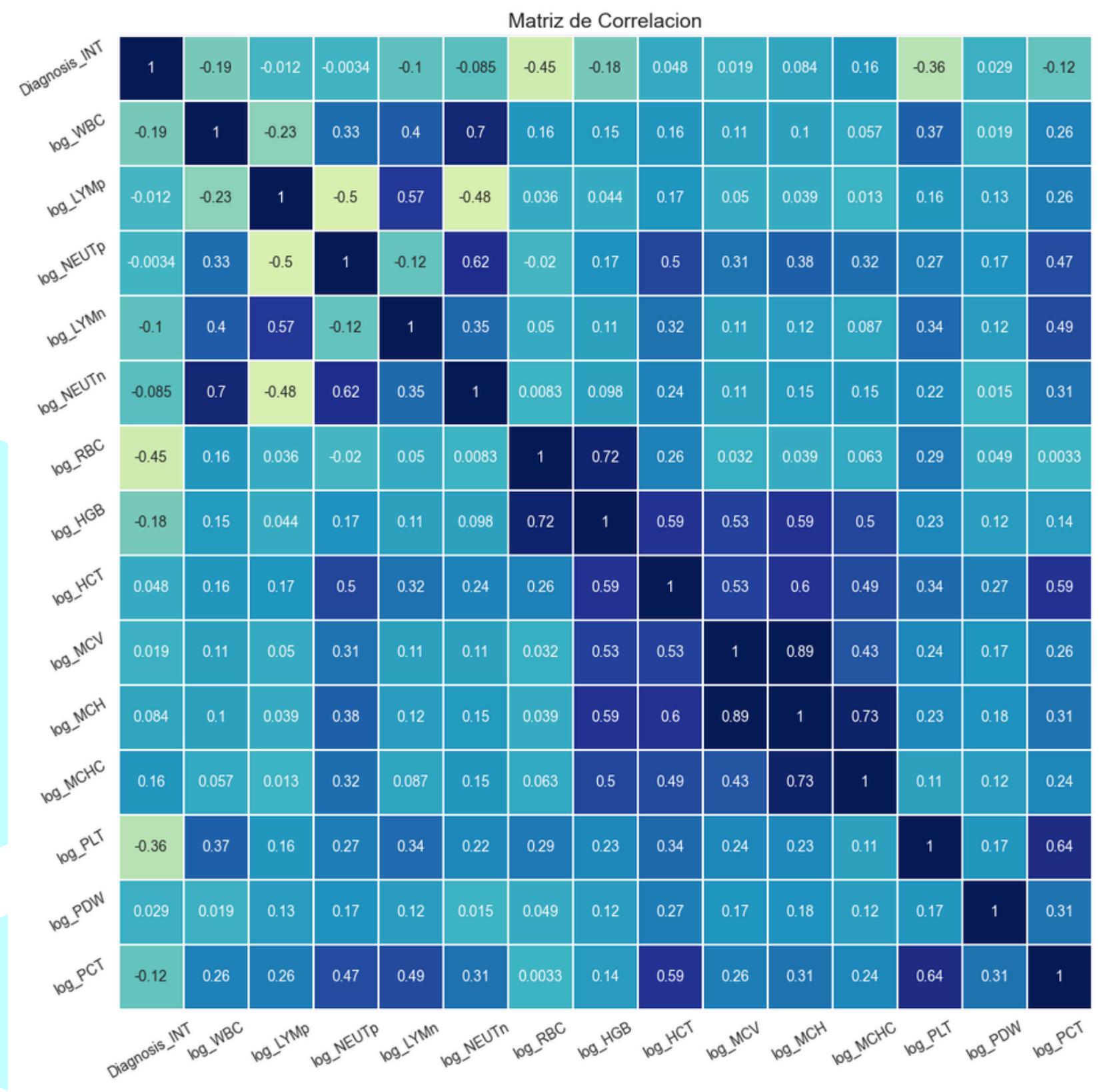
La presencia de outliers puede estar afectando la correlación de variables. Se verifican valores atípicos en el boxplot



6.

**Se aplica un filtro logaritmico que permite
un mejor tratado de los datos**



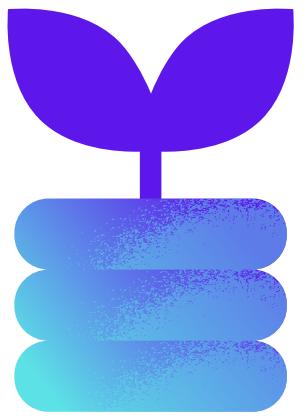


7.

Se verifica un cambio en las variables. Por lo que se procede a escoger las variables con mayor correlación, asumiendo algunas que presentan una correlación mínimamente mayor al 0.6 con respecto de las otras variables

PCT	CHC	RBC
PLT	HGB	WBC

**PARA EL DIAGNÓSTICO SE EVALUARON TRES
MODELOS DIFERENTES DE APRENDIZAJE
AUTOMÁTICO:**



Árbol de decisión
`(`DecisionTreeClassifier`)`



K-Near más cercano
`(`KNeighborsClassifier`)`

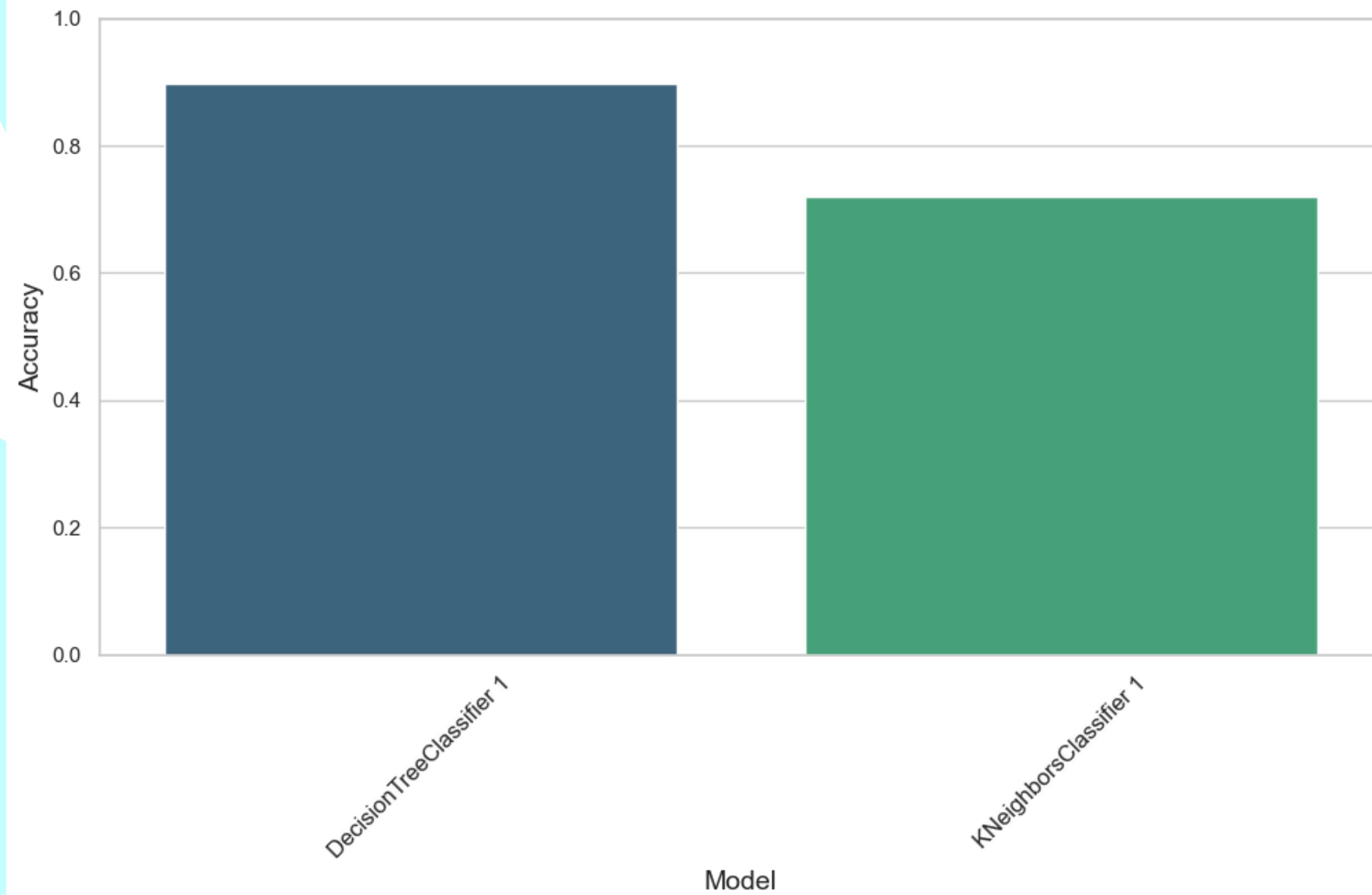
ENTRENAMIENTO DE MODELOS

Se tratan los datos desbalanceados y se observa la respuesta de las métricas al dataset entrenado

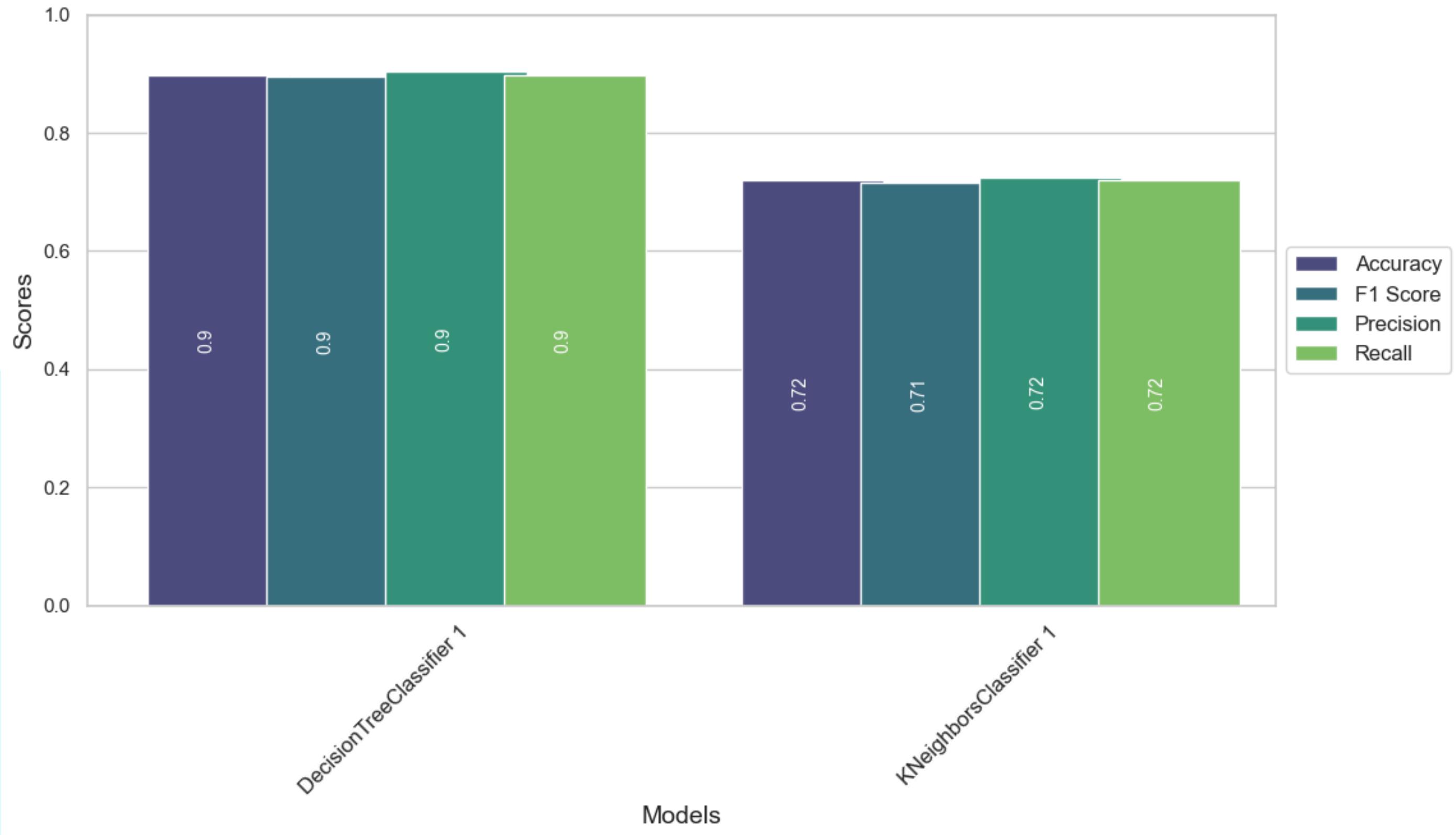
TRATAMIENTO	DecisionTreeClassifier	KNeighborsClassifier
DESBALANCEADO	0.896551724137931	0.7198275862068966

LA MÉTRICA IMPLEMENTADA FUE 'ACUARICY'

Models Accuracy



Models Performance Metrics



SE IMPLEMENTARON
OTRAS MÉTRICAS
PARA AUMENTAR LA
CREDIBILIDAD DEL
MODELO

CONCLUSIONES FINALES

El desarrollo e implementación de un modelo de Machine Learning para predecir el diagnóstico de enfermedades asociadas a la anemia ha demostrado ser altamente efectivo, alcanzando una precisión del 70% - 90%. Este modelo tiene el potencial de transformar significativamente la forma en que se detectan y manejan estas enfermedades, mejorando la calidad de vida de los pacientes mediante diagnósticos más tempranos y precisos.

Estas conclusiones resaltan la efectividad y el impacto positivo que el modelo de Machine Learning puede tener en el diagnóstico y tratamiento de enfermedades asociadas a la anemia, marcando un avance significativo en la aplicación de inteligencia artificial en el campo de la salud.



- **Alta precisión del modelo:** El modelo desarrollado ha alcanzado una precisión del 99%, lo que indica su alta capacidad para predecir correctamente la presencia de enfermedades asociadas a la anemia. Esto sugiere que el uso de técnicas avanzadas de Machine Learning puede ser extremadamente eficaz en el diagnóstico médico.
- **Impacto positivo en la atención médica:** La implementación de este modelo en un entorno clínico puede reducir significativamente el tiempo y los recursos necesarios para diagnosticar enfermedades relacionadas con la anemia, permitiendo a los profesionales de la salud tomar decisiones más informadas y rápidas, y mejorar así el cuidado del paciente.
- **Validación y robustez del modelo:** La alta precisión lograda se debe a un riguroso proceso de validación y evaluación del modelo, que incluyó técnicas de validación cruzada y el uso de diversas métricas de rendimiento. Esto asegura que el modelo no solo es preciso, sino también robusto y confiable para su aplicación en diferentes conjuntos de datos.
- **Potencial para expandir el uso del modelo:** Dado el éxito del modelo en predecir enfermedades asociadas a la anemia, existe un gran potencial para ampliar su uso a otras condiciones médicas. Adaptar este enfoque a diferentes enfermedades podría mejorar aún más los resultados de salud a nivel global, demostrando la versatilidad y el poder de las técnicas de Machine Learning en la medicina.

MUCHAS GRACIAS



MACHINE
LEARNING

Clasificación de tipo de anemias

Análisis y datos sobre los
diagnósticos de anemia.