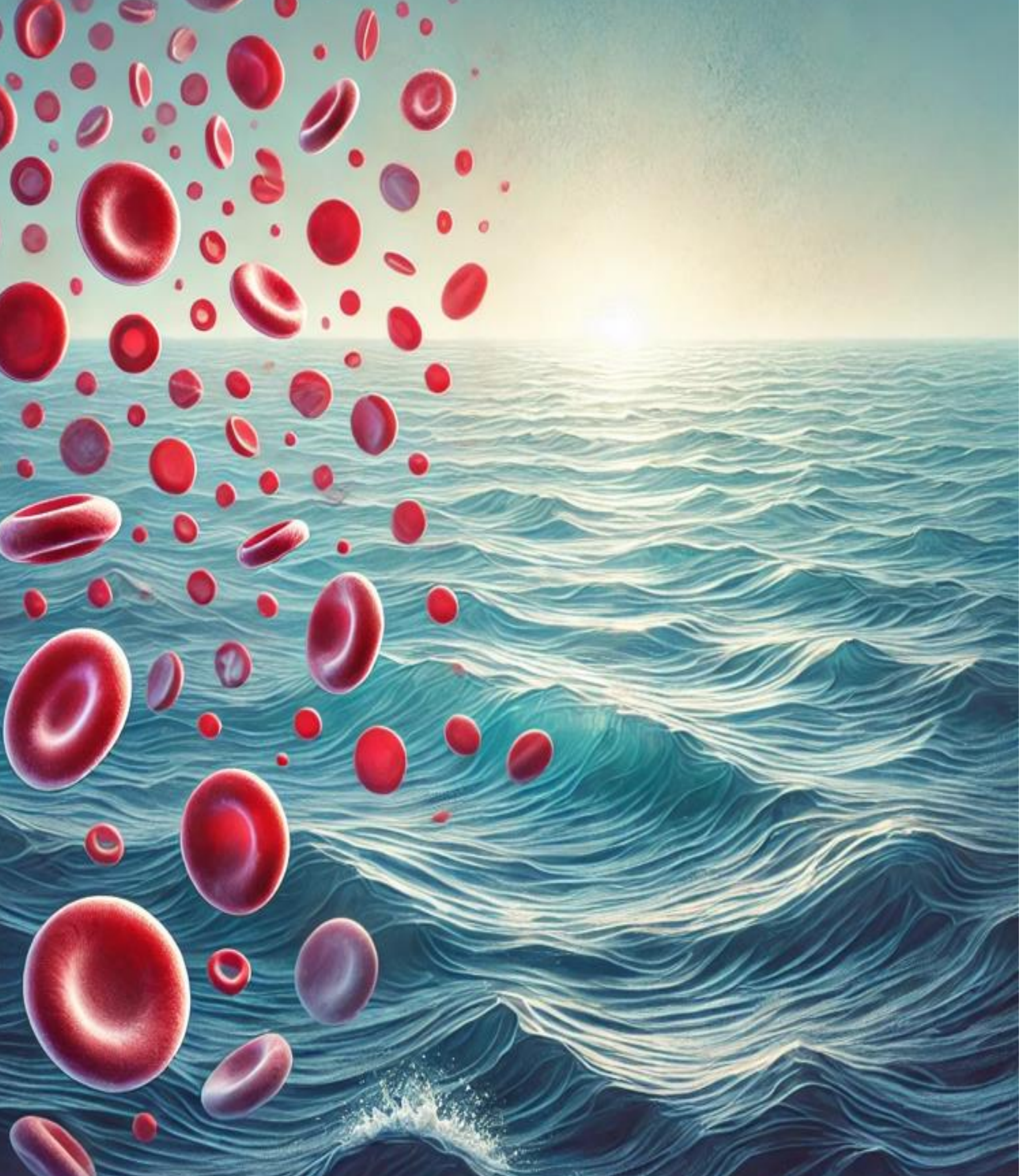


A Diagnostic Model for Alpha Thalassemia Diagnosis using Complete Blood Count Data with ML Techniques

Health Data Analytics – 2023 / 2024

De Nardin Carlo
carlo.denardin@studenti.units.it

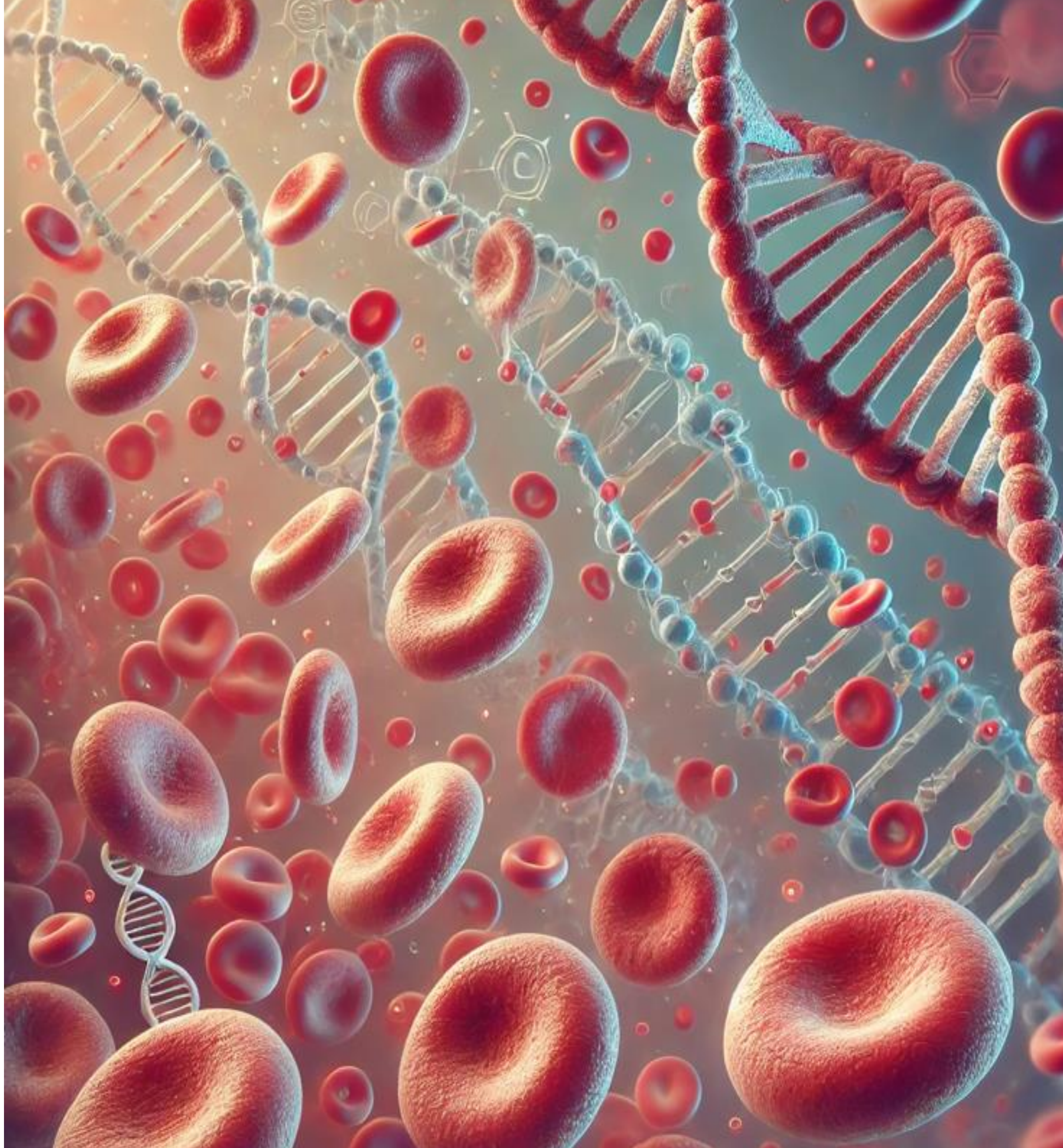
1 Introduction

2 Dataset & IDA

3 Diagnostic Models

4 Results

5 Conclusions



Introduction: what is Thalassemia?

Thalassemia, from the Greek *Thalassa* (sea) and *Haima* (blood), is a **hereditary genetic disease** that **affects the production of hemoglobin**, the protein responsible for oxygen transport in red blood cells, leading to a condition known as **anemia**.

Hemoglobin is composed by four protein chains: **two alpha (4 genes)** and two beta (2 genes). Each person inherits half of these genes from each parent. Mutations in one or more of these genes compromise the production of the protein chains, causing thalassemia. **The more genes are affected, the more severe the thalassemia will be**.

Affected Genes	Name	Symptoms
0	Normal	-
1	Silent Alpha Thalassemia	Asymptomatic
2	Trait Alpha Thalassemia	Asymptomatic or may have mild anemia symptoms (fatigue)
3	Intermidia Alpha Thalassemia (HbH)	Anemia symptoms at birth and leads to severe lifelong anemia (poor appetite, pale skin, dark urine, irregular face bone structure)
4	Hb Bart’s Hydrops Fetails Syndrome	Fetal death

Table: Types of Alpha Thalassemia

Introduction: where is Thalassemia?

Thalassemia is prevalent in areas of the world where malaria is or was present, such as **Southeast Asia**, **Africa**, and **parts of the Mediterranean***.

Thalassemia carriers have compromised red blood cells that prevent the malaria parasite from completing its life cycle and spreading throughout the carrier's body**.

Approximately **5% of the populations** has a form of thalassemia but only **1.7% of the population show traits** of the disease***.

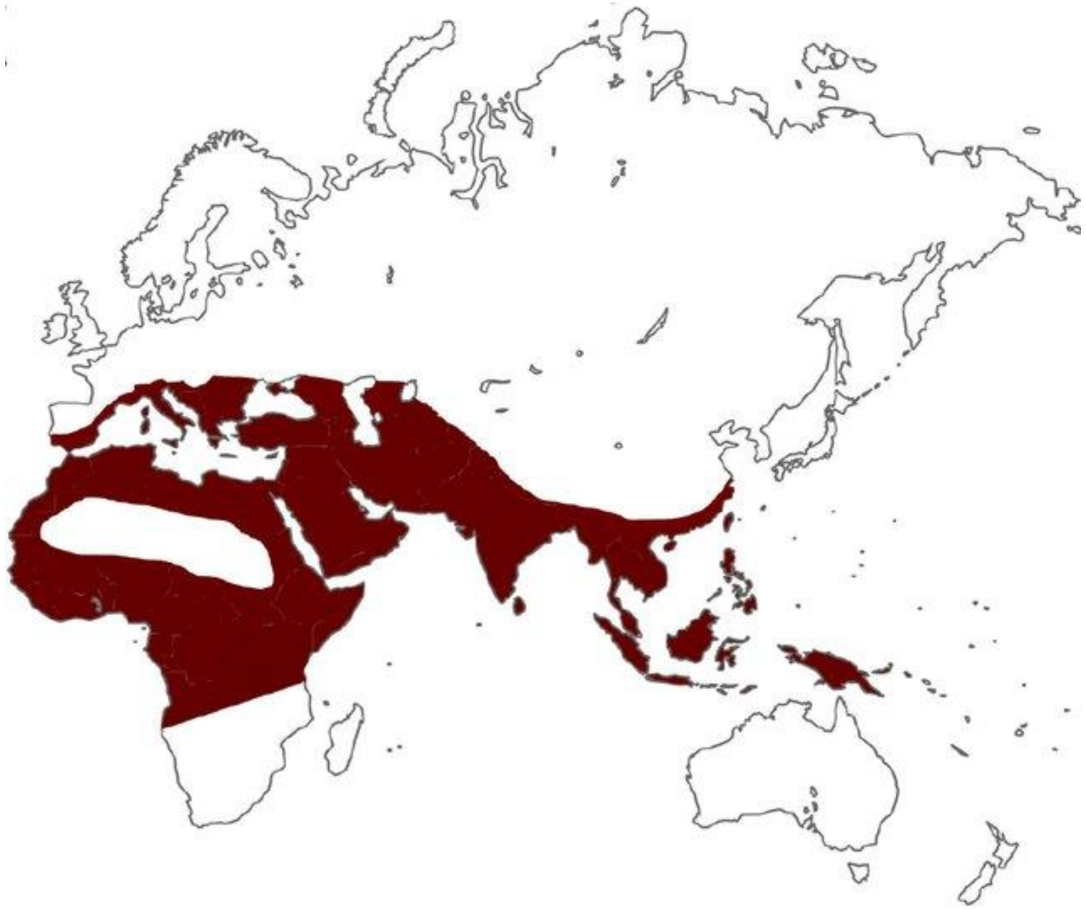


Figure: Thalassemia most affected countries

01
02
03
04
05

* Singer, Titi. (2009). Variable Clinical Phenotypes of α-Thalassemia Syndromes. TheScientificWorldJournal. 9. 615-25. 10.1100/tsw.2009.69.
** https://malattierarealcentro.it/talassemia-e-malaria-quando-una-malattia-offre-un-vantaggio-selettivo-rispetto-a-unaltra/
*** https://www.news-medical.net/health/Thalassemia-Prevalence.aspx

Introduction: how is Alpha Thalassemia diagnosed?

The diagnosis of Alpha Thalassemia is carried out using different methodologies*:

- **Complete Blood Count (CBC):** a blood test that measures hemoglobin levels and analyzes other red blood cells parameters
- **Hemoglobin electrophoresis:** separates different types of hemoglobin to identify any anomalies (not always effective in detecting Alpha thalassemia**)
- **Genetic test:** used to identify genes mutations (expensive)

Alpha Thalassemia is rarely treatable. The best solution is **prevention** and **awareness**, as two silent carriers can have children with severe forms of the disease.

* https://www.msdmanuals.com/it-it/casa/disturbi-del-sangue/anemia/talassemie

** https://www.aafp.org/pubs/afp/issues/2009/0815/p339.html#:~:text=In%20the%20newborn%20period%2C%20if,major%20are%20diagnosed%20during%20infancy.

Introduction: research purpose

The goal of this research is to develop a diagnostic model that categorized Alpha Thalassemia and its severity using only a **Complete Blood Count**.

Alpha thalassemia is prevalent in many poor countries that cannot afford expensive methodologies like genetic test*. A diagnostic model based on a simple **CBC** may offer an accessible solution that can help these countries reducing new cases of Alpha Thalassemia.

To reduce the incidence, it is essential to **raise awareness**, especially for **individuals who may be asymptomatic** so the discrimination between ***normal*** and ***silent*** conditions is crucial **(maximize recall)**.

Research questions:

- *What are the parameters of a CBC that are related to Alpha Thalassemia?*
- *Can ML models diagnose Alpha Thalassemia in asymptomatic carriers to help reduce its incidence?*

* Kolambage, Nival & Goonasekara, Hemali & Hewapathirana, Roshan. (2020). Design, Development and Implementation of a Machine Learning-based Predictive Modelling Tool to Accurately Predict Thalassemia Carrier state using Full Blood Count Indices and Haemoglobin Variants.

Initial Data Analysis: dataset description

The dataset* used consists in various parameters related to a CBC and patient information. The variables contained in the dataset are listed below.

Variable	Description
Hemoglobin (HB)	Amount of hemoglobin in the blood (g/dL)
Hematrocrit (HCT)	Percentage of blood volume occupied by red blood cells (%)
Mean Corpuscular Volume (MCV)	Average size of red blood cells (fL)
Mean Corpuscular Hemoglobin (MCH)	Average amount of hemoglobin contained in a single red blood cell (pg)
Mean Corpuscular Hemoglobin Concentration (MCHC)	Average concentration of hemoglobin within a red blood cell (g/dL)
Red Cell Distribution Width (RDW)	Variation in size among red blood cells (%)
Red Blood Cell Count(RBC)	Total number of red blood cells in the blood (milions/µL)
Age	Patient's age (years)
Gender	Patient's gender (1: female, 2: male)
Target	Thalassemia type (1: normal, 2: silent, 3: trait, 4: HbH)

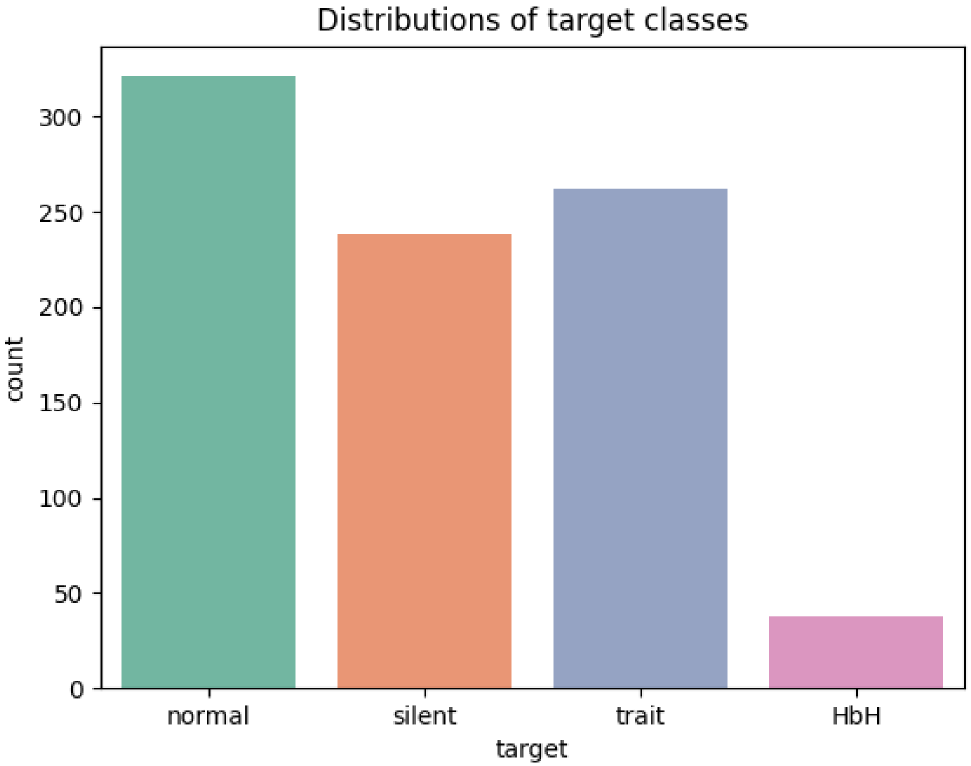
Table: Alpha Thalassemia dataset description

* https://www.kaggle.com/datasets/plenoi/thalassemia/data

** Longo F, Corrieri P, Origa R, Barella S, Sanna PMG, Bitti PP, Zuccarelli A, Commendatore FV, Vitucci A, Quarta A, Lisi R, Cappellini MD, Massei F, Forni GL, Piga A. Changing patterns of thalassaemia in Italy: a WebThal perspective. Blood Transfus. 2021 May;19(3):261-268. doi: 10.2450/2020.0143-20. Epub 2020 Nov 3. PMID: 33196416; PMCID: PMC8092041.

Initial Data Analysis: target variable & missing data

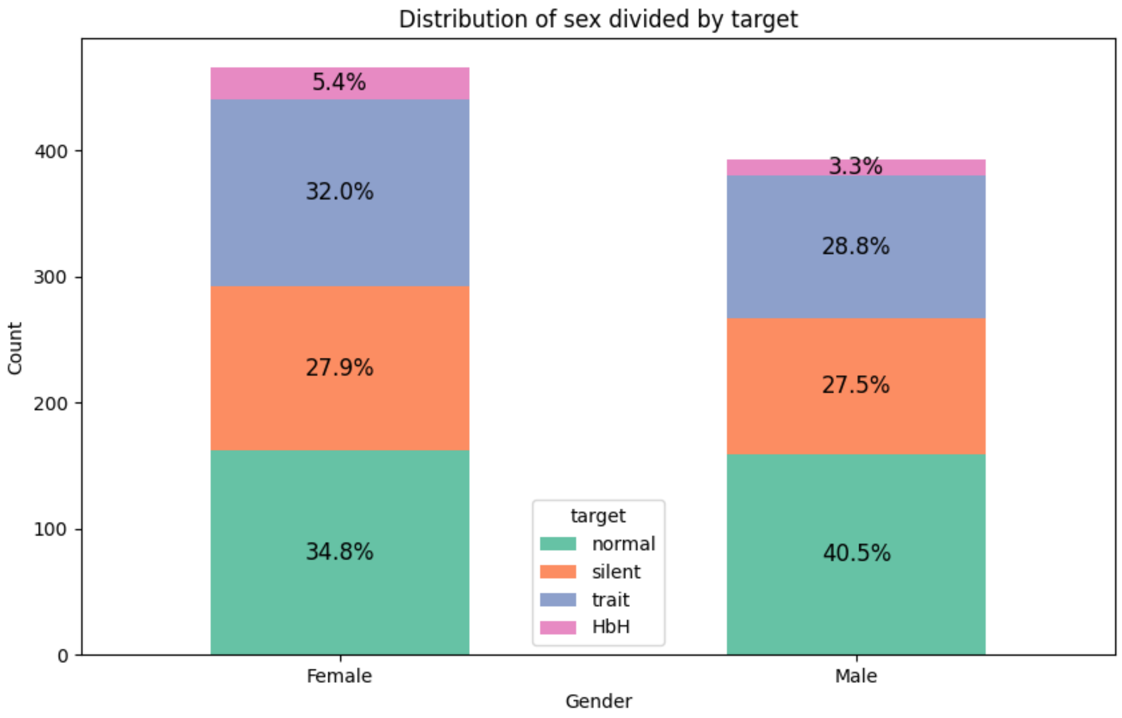
The **target** variable is categorical and represents different types of Alpha Thalassemia. It is divided into four categories (1: normal, 2: silent, 3: trait, 4: HbH). Below, the distributions of each category in the dataset are shown.

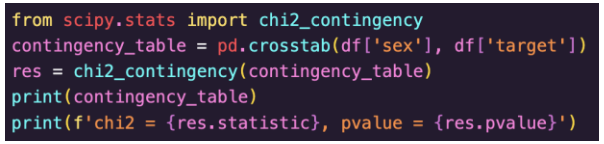


The **first three classes are slightly imbalanced**, while the **HbH class is significantly underrepresented**. This imbalance aspect must be considered when developing the diagnostic model.

There are no missing values in the dataset.

Initial Data Analysis: gender





```
from scipy.stats import chi2_contingency
contingency_table = pd.crosstab(df['sex'], df['target'])
res = chi2_contingency(contingency_table)
print(contingency_table)
print(f'chi2 = {res.statistic}, pvalue = {res.pvalue}')
```

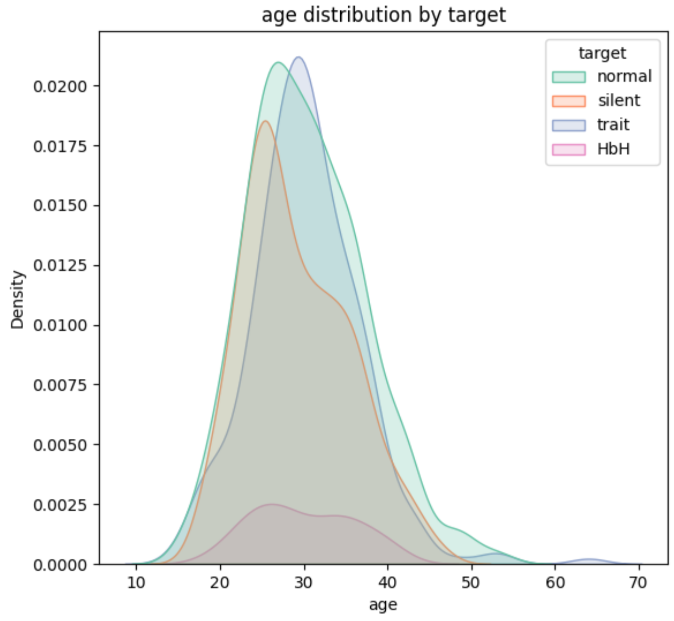
x^2	p-value
4.6273	0.2012

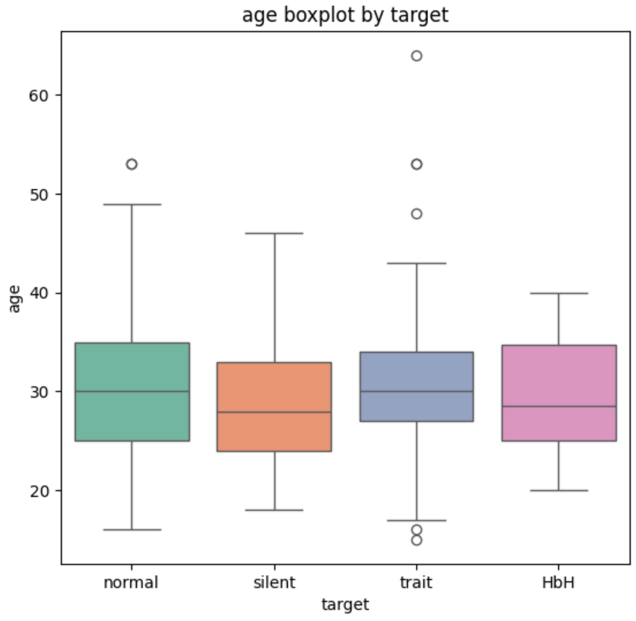
The dataset contains a higher percentage of female patients. A higher percentage of male patients is seen in the ***normal*** class, while the ***trait*** class shows a greater presence of female patients.

There does not appear to be a significant correlation between these two variables. To assess this statement, a Chi-square test can be performed to determine the independence between gender and target.

Since the p-value is greater than 0.05, there is not enough evidence to assert that a significant relationship exists.

Initial Data Analysis: age


age distribution by target


age boxplot by target

H_0: The data follow a normal distribution

Degree of asymmetry

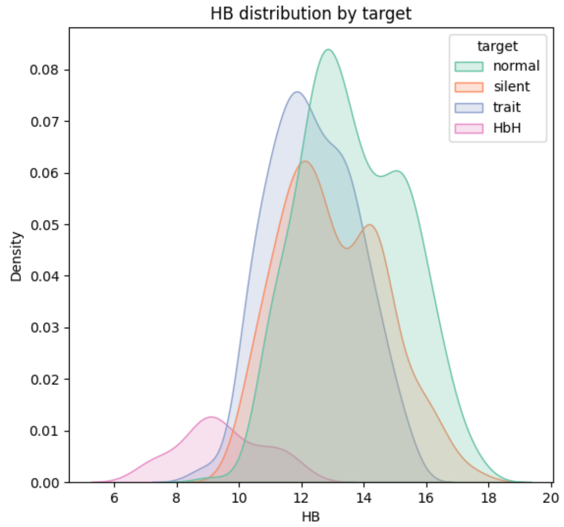
	Shapiro-Wilk (p-value)	Skewness	Kruskal-Wallis
🟩	9.38e-05	0.49	H-Value: 6.33 P-value: 0.09
🟧	1.03e-05	0.80	
🟦	5.94e-07	0.16	
🟪	0.07	0.52	

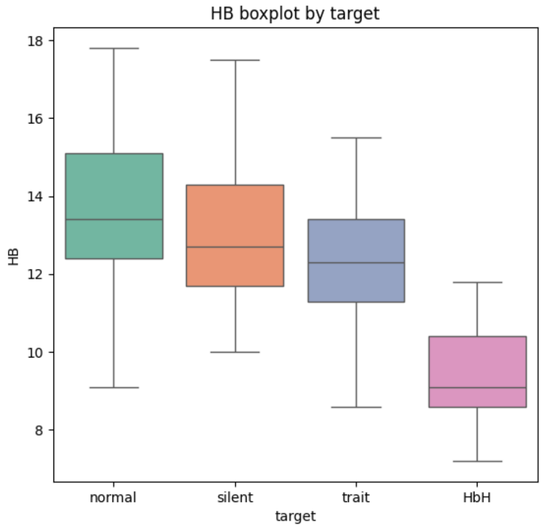
H_0: The rank distributions across the different groups are the same

The categories show a similar distribution, with the ***silent*** and ***HbH*** variables having a slightly lower median. *Alpha Thalassemia is a hereditary disease present from birth, this variable seems to be more related to the methodology of data collection used in the study rather than the target variable.*

01
02
03
04
05

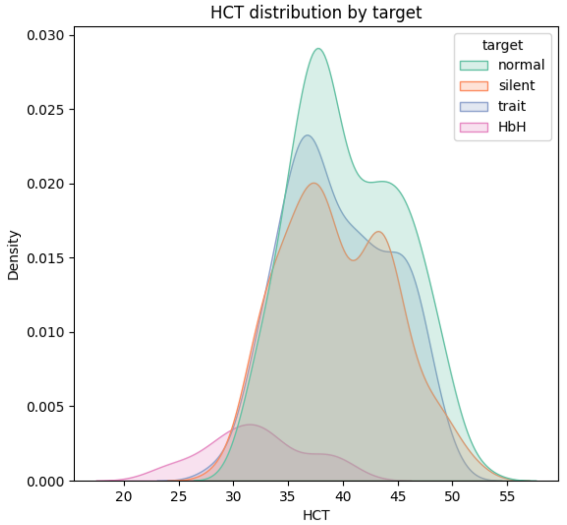
Initial Data Analysis: HB & HCT

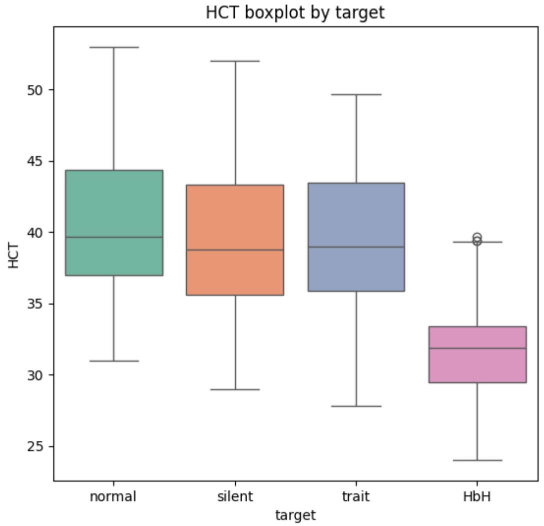

HB distribution by target


HB boxplot by target

	Shapiro-Wilk (p-value)	Skewness	Kruskal-Wallis
🟩	0.0006	0.16	H-Value: 165.96 P-value: 9.4e-36
🟧	0.0002	0.35	
🟦	0.0214	0.09	
🟪	0.0719	0.09	

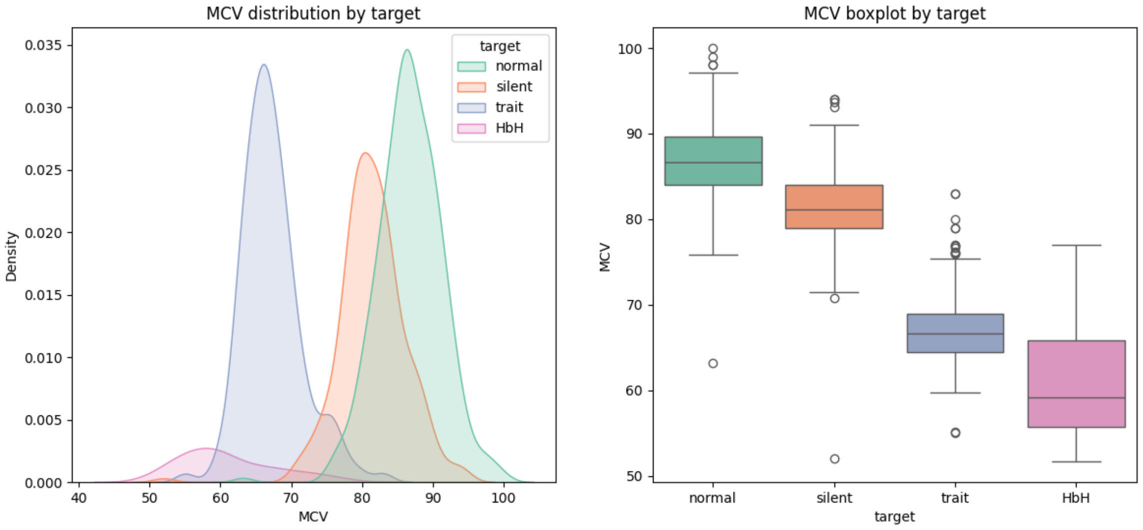
A **gradual decrease** in both **HB** and **HCT** (less evident) values can be observed from the ***normal*** class to the ***trait*** class. The ***HbH*** class shows a much more **evident reduction** for both **HB** and **HCT**, highlighting a difference with respect to other groups.


HCT distribution by target


HCT boxplot by target

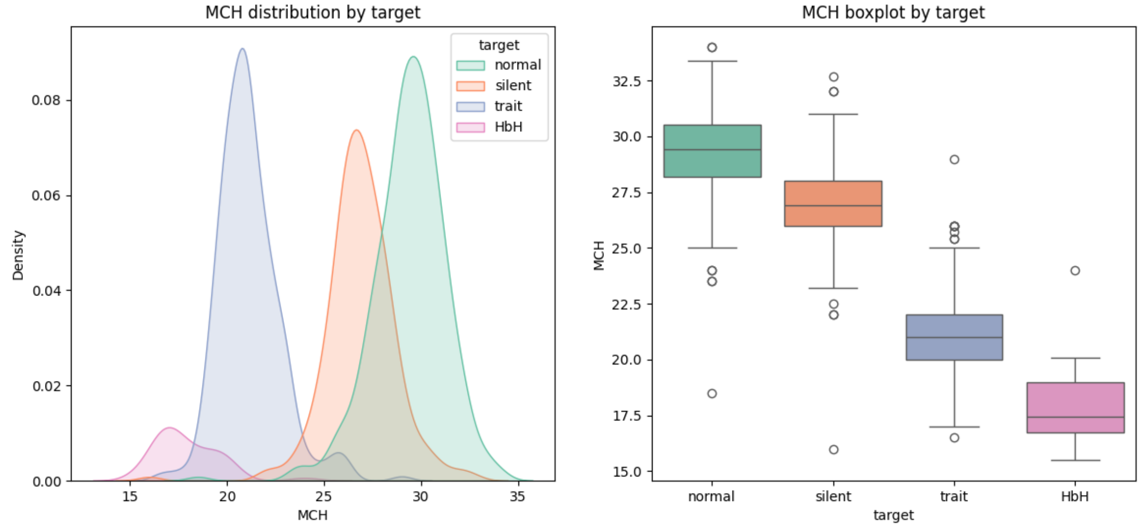
	Shapiro-Wilk (p-value)	Skewness	Kruskal-Wallis
🟩	4.1e-05	0.21	H-Value: 65.963 P-value: 3.1e-14
🟧	0.0009	0.25	
🟦	0.0001	0.12	
🟪	0.0638	0.08	

Initial Data Analysis: MCV & MCH



	Shapiro-Wilk (p-value)	Skewness	Kruskal-Wallis
🟩	0.0003	-0.31	H-Value: 634.62 P-value: 3e-137
🟧	2.7e-08	-0.77	
🟦	6.7e-08	0.83	
🟪	0.01	0.72	

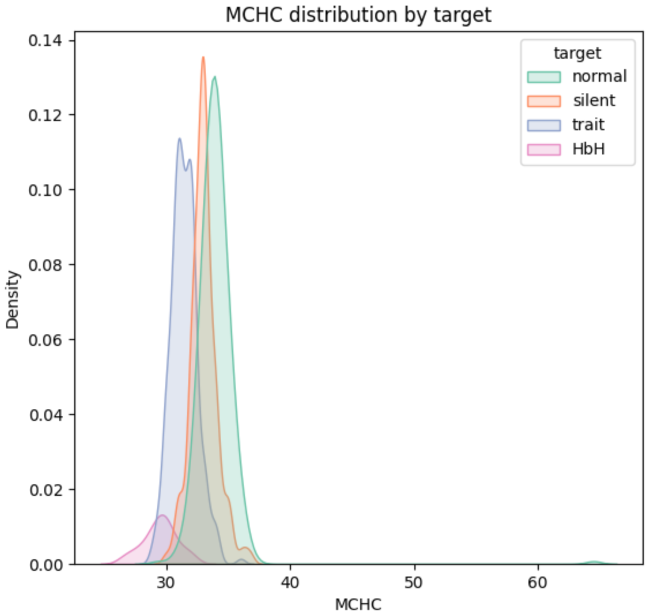
There is a **clear decrease** in **MCV** and **MCH** values across the **different classes**. The presence of outliers highlights greater variability within certain groups. The **distributions** of MCV and MCH appear **similar**, suggesting a possible **strong correlation** between the two variables.

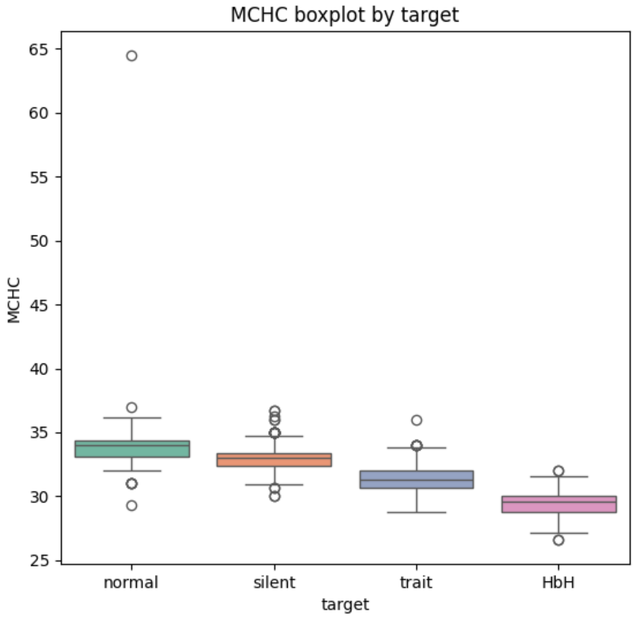


	Shapiro-Wilk (p-value)	Skewness	Kruskal-Wallis
🟩	1.2e-08	-0.93	H-Value: 664.74 P-value: 9e-144
🟧	3.1e-09	-0.70	
🟦	2.1e-09	1.06	
🟪	0.001	1.30	

01
02
03
04
05

Initial Data Analysis: MCHC

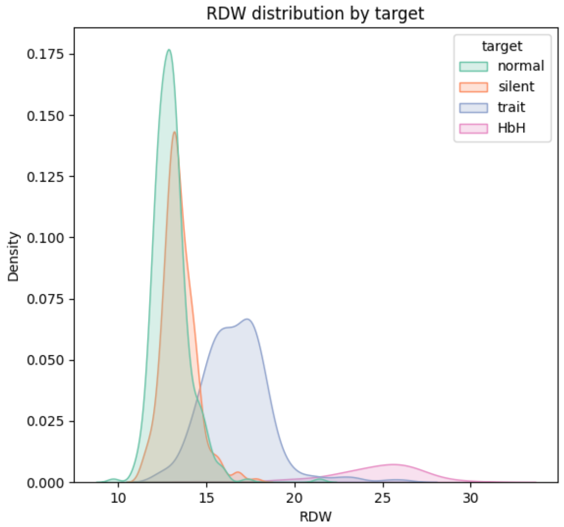


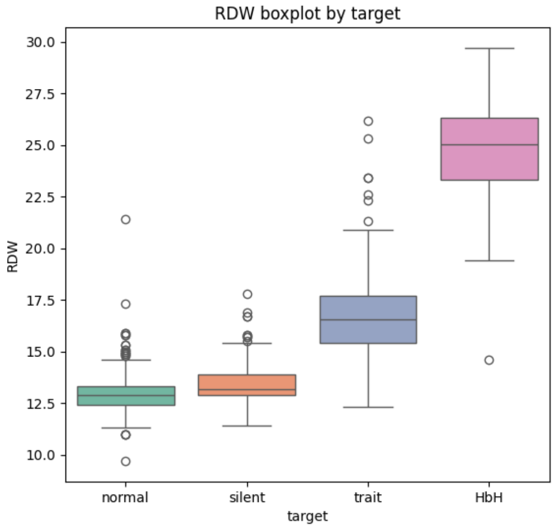


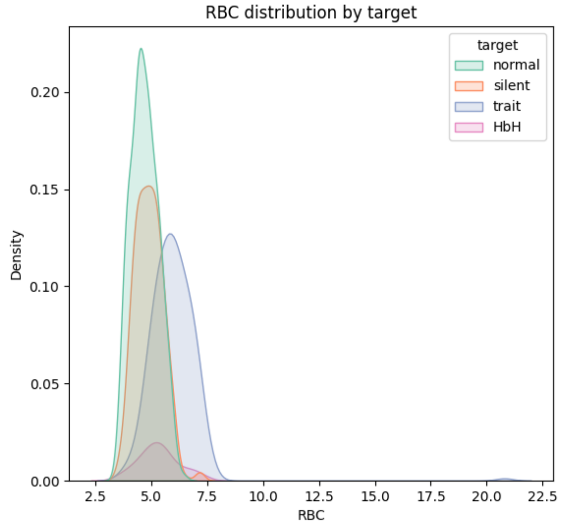
	Shapiro-Wilk (p-value)	Skewness	Kruskal-Wallis
🟩	1e-30	11.07	H-Value: 495.15 P-value: 5e-107
🟧	6.2e-07	0.49	
🟦	0.473	0.46	
🟪	1.044	-0.19	

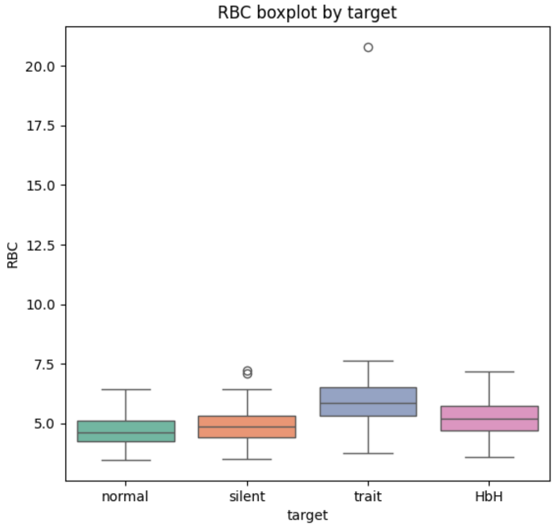
The classes have relatively close values, with a slight decrease form the ***normal*** group to the ***HbH*** group. Additionally, several outliers are noticeable in all groups. In the normal class there is a strange outlier. In the next section the relevant outliers will be handled.

Initial Data Analysis: RDW & RBC


RDW distribution by target


RDW boxplot by target


RBC distribution by target

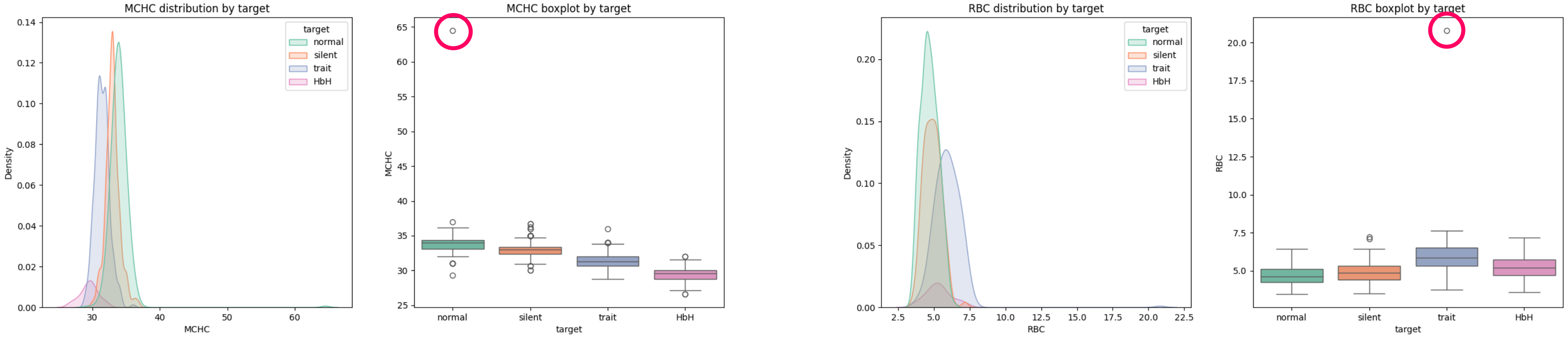

RBC boxplot by target

	Shapiro-Wilk (p-value)	Skewness	Kruskal-Wallis
🟩	8.4e-16	2.16	H-Value: 538.20 P-value: 2e-116
🟧	1.2e-08	1.09	
🟦	4.1e-10	1.17	
🟪	0.001	-1.37	

There is a **clear increase** in **RDW** values across the different groups, especially for the ***trait*** and **HbH** groups. The **RBC** values are similar across the ***normal*** and ***silent*** groups but are higher for the last two groups. There is a strange outlier for the ***trait*** *class.*

	Shapiro-Wilk (p-value)	Skewness	Kruskal-Wallis
🟩	0.001	0.24	H-Value: 298.6 P-value: 1e-64
🟧	0.004	6.89	
🟦	6.1e-24	0.29	
🟪	0.44	1.30	

Initial Data Analysis: outliers



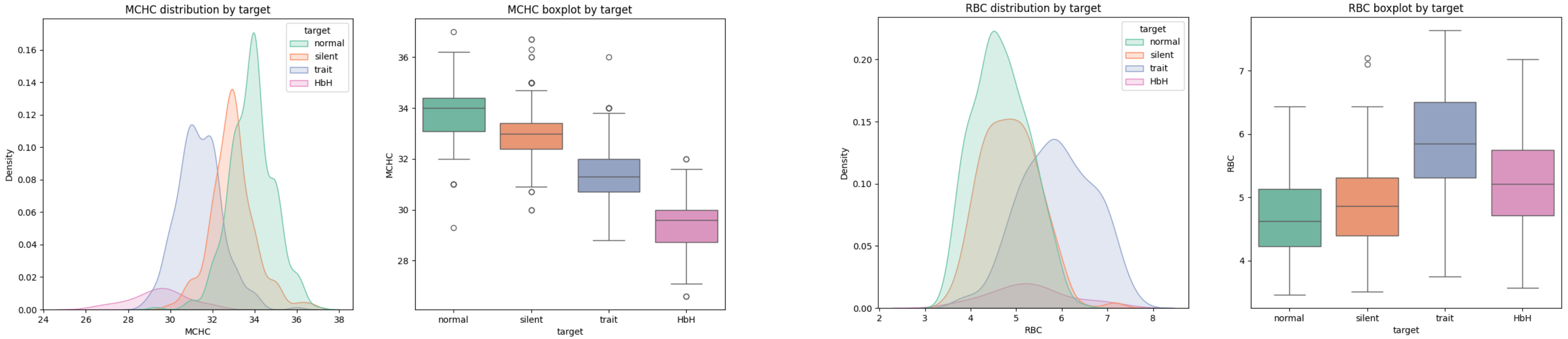
MCHC distribution by target

MCHC boxplot by target

RBC distribution by target

RBC boxplot by target

In the **normal** and **trait** groups, respectively for the independent variables MCHC and RBC, two clearly anomalous values were identified. The values and the corresponding records are removed from the dataset.



MCHC distribution by target

MCHC boxplot by target

RBC distribution by target

RBC boxplot by target

Initial Data Analysis: considerations

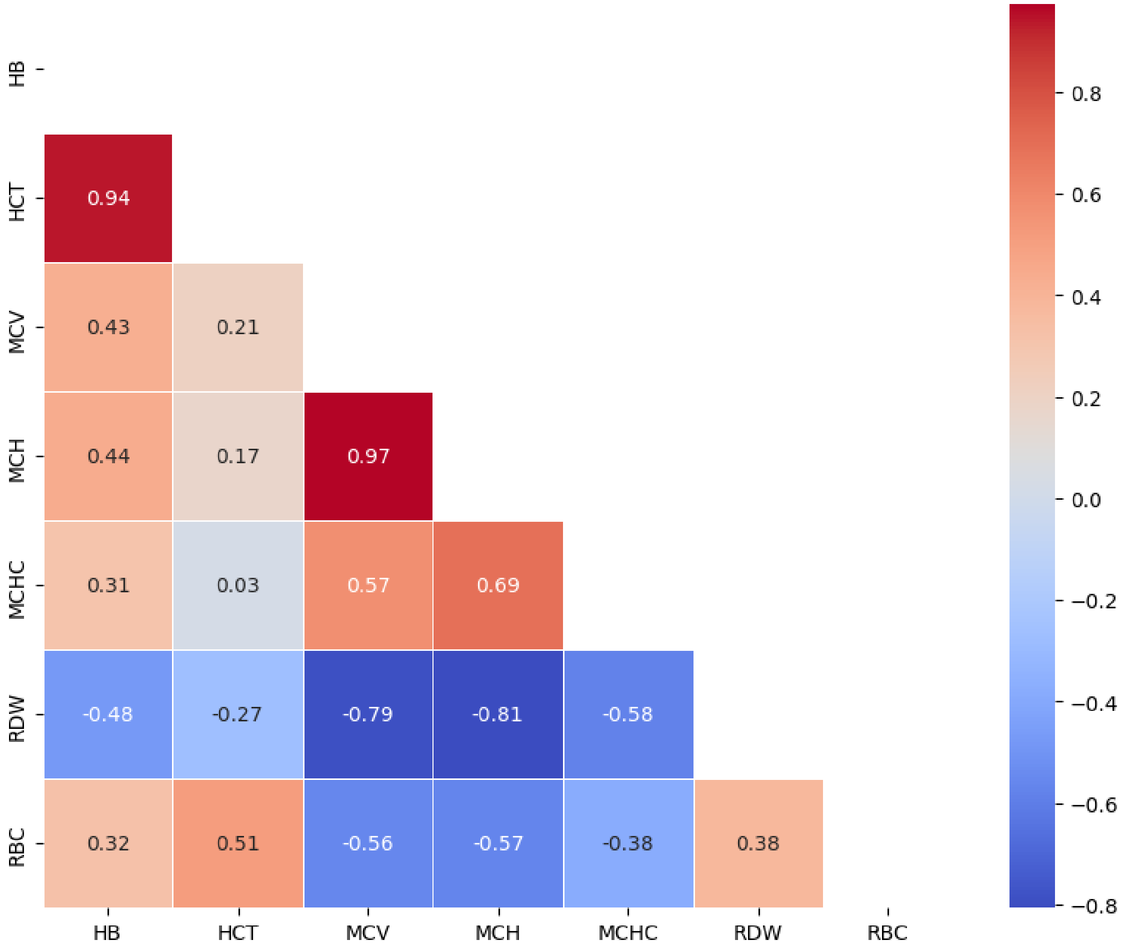
From the analysis conducted, several considerations emerged:

- The variables ***age*** and ***gender*** do not seem to be significantly useful for the construction of a diagnostic model*.
- All the other independent variables obtained from the **CBC** proved to be relevant for the development of a diagnostic model.
- Correlations between some independent variables were identified from the plots and in the next section an accurate analysis is presented.

01
02
03
04
05

* Muncie HL Jr, Campbell J. Alpha and beta thalassemia. Am Fam Physician. 2009 Aug 15;80(4):339-44. PMID: 19678601.

Initial Data Analysis: multicollinearity



Variables	Description
MCV, MCH, MCHC	MCH and MCHC are strongly correlated due to their definitions (0.81). Additionally, both varibles are strongly correlated with MCV, as smaller red blood cells tend to have a lower MCH, while larger red blood cells tend to have a higher value. The combination MCV- MCH ha a Pearson correlation coefficient of 0.97, while the combination MCV– MCHC has a coefficient of 0.67.*
HB, HCT	HB and HCT are closelly correlated (0.94) since almost all the hemoglobin in the blood is contained within red blood cells. Therefore measuring the concentration of hemoglobin (HB) in the blood or the percentage of blood volume occupied by red blood cells (HCT) provides very similar information.**
RDW, RBC, MCV, MCH, MCHC	The strong negative correlation between RDW and RBC and the variables MCV (-0.79, -0.63), MCH (-0.81, -0.64) and, MCHC (-0.69, -0.50) is related, in the case of RDW, to the calculation of the parameter itself, which is directly influenced by MCV. In the case of RBC, the negative correlation is due to the compensatory overproduction of red blood cells in response to the presence of affected red blood cells.***
RBC, HCT	RBC and HCT, by definition, measure in a different way the same phenomenon: tha amount of red blood cells in the blood.****

Table: Multicollinearity explanation

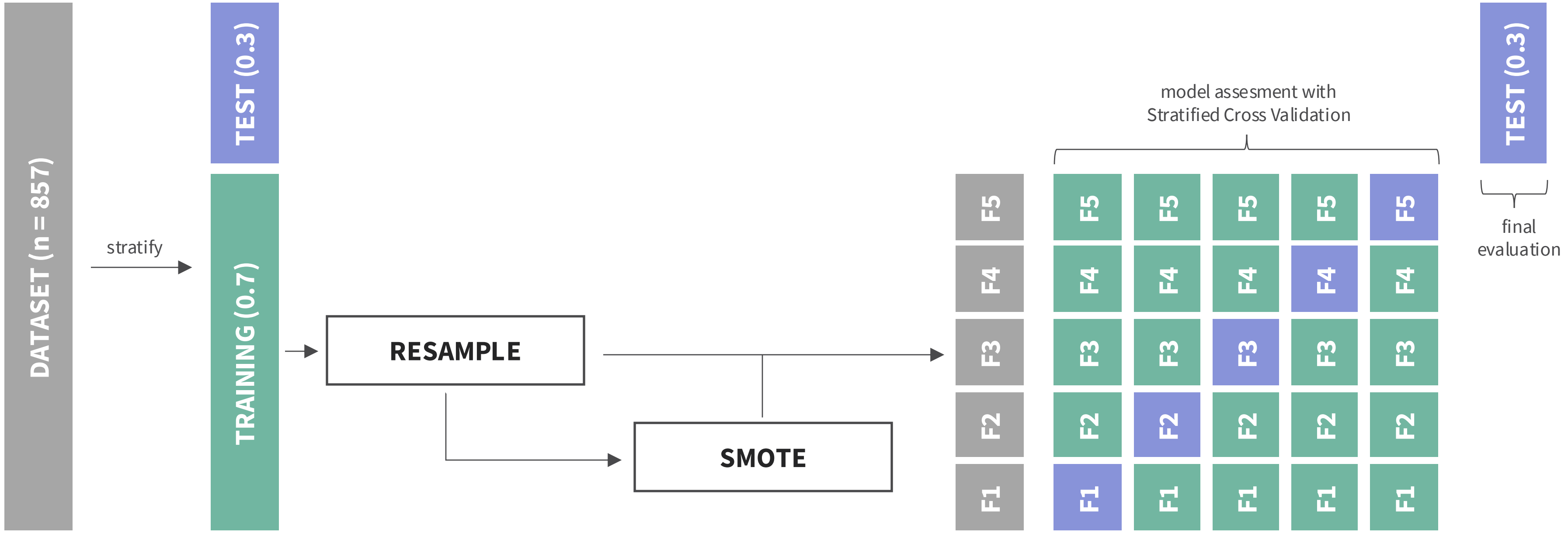
* https://www.valorinormali.com/sangue/mch/
** https://www.crit.cloud/summaries--reviews/hematocrit-and-hemoglobin-get-it-right-once-and-for-all
*** Brzeźniakiewicz-Janus, K., Rupa-Matysek, J., Tukiendorf, A. *et al.* Red Blood Cells Mean Corpuscular Volume (MCV) and Red Blood Distribution Width (RDW) Parameters as Potential Indicators of Regenerative Potential in Older Patients and Predictors of Acute Mortality – Preliminary Report.
**** https://www.nature.com/articles/pr2001213

Diagnostic Models: approach

The goal of this research is to develop a diagnostic model for Alpha-Thalassemia using variables derived from a **CBC**.

Three distinct models were developed using Logistic Regression:

- The first model employs **manual variable selection** based on exploratory analysis
- The second model uses **Recursive Feature Elimination (RFE)** to identify the most relevant variables
- The third model applies **Principal Component Analysis (PCA)** to reduce multicollinearity among the independent variables



Diagnostic Models: logistic regression with manual variable selection

After the exploratory analysis and the assessment of multicollinearity, a manual selection of variables was performed. The following variables were selected: **MCV**, **HB**, **RBC** and the LR model was trained with and without oversampling (SMOTE) on train data.

class	precision	recall	f1-score	n
normal	0.80	0.86	0.83	224
silent	0.72	0.66	0.69	166
trait	0.89	0.92	0.91	182
HbH	0.89	0.59	0.71	27
w. avg	0.81	0.81	0.81	599

Table: Stratified Cross Validation without Oversampling

class	precision	recall	f1-score	n
normal	0.80	**0.84**	**0.82**	96
silent	**0.68**	0.69	0.68	72
trait	0.95	**0.90**	**0.92**	79
HbH	1.00	0.73	0.84	11
w. avg	**0.84**	**0.83**	**0.83**	258

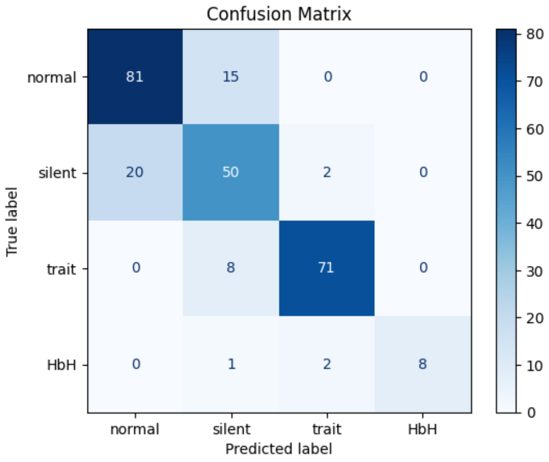
Table: Test Set without Oversampling

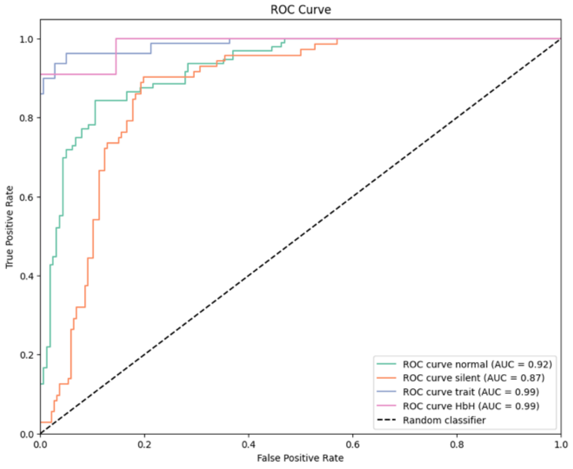
class	precision	recall	f1-score	n
normal	0.78	0.79	0.79	224
silent	0.72	0.73	0.72	224
trait	0.87	0.88	0.87	224
HbH	0.94	0.92	0.93	224
w. avg	0.83	0.83	0.83	869

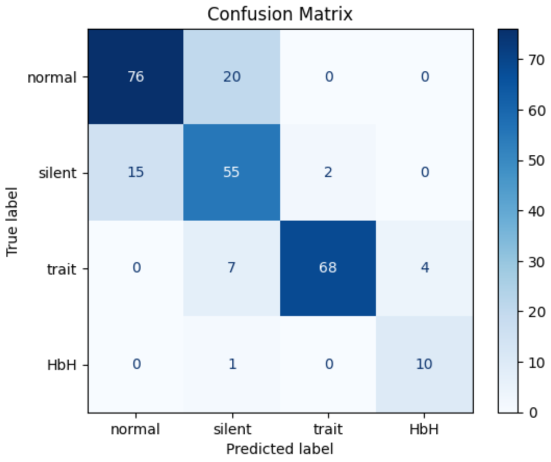
Table: Stratified Cross Validation with Oversampling (SMOTE)

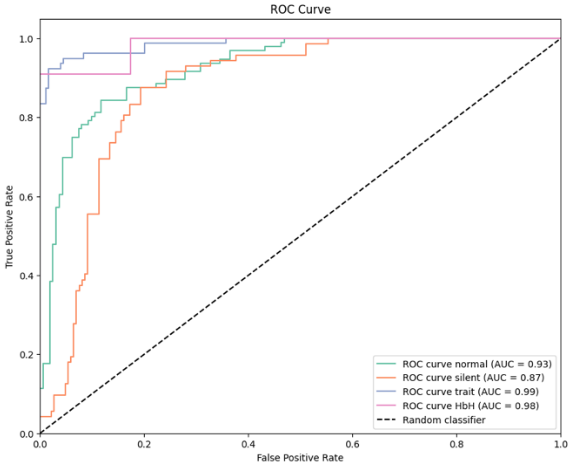
class	precision	recall	f1-score	n
normal	**0.84**	0.79	0.81	96
silent	0.66	**0.76**	**0.71**	72
trait	**0.97**	0.86	0.91	79
HbH	0.71	0.91	0.80	11
w. avg	0.82	0.81	0.81	258

Table: Test Set with Oversampling (SMOTE)









Diagnostic Models: logistic regression with recursive feature elimination

For the second model, Recursive Feature Elimination (RFE) was used. This technique allows the selection of the most relevant variables for the model by progressively eliminating the less important features. RFE returned the following variables in order of importance: MCH, RDW, MCV, HB, MCHC, RBC and HCT.

class	precision	recall	f1-score	n
normal	0.79	0.86	0.83	224
silent	0.74	0.64	0.69	166
trait	0.90	0.93	0.92	182
HbH	0.88	0.81	0.85	27
w. avg	0.82	0.82	0.82	599

Table: Stratified Cross Validation without Oversampling with first 2 variables **MCH** and **RDW**

class	precision	recall	f1-score	n
normal	0.79	**0.84**	0.82	96
silent	**0.73**	0.71	0.72	72
trait	0.99	**0.92**	**0.95**	79
HbH	**0.83**	**0.91**	**0.87**	11
w. avg	0.84	0.83	0.83	258

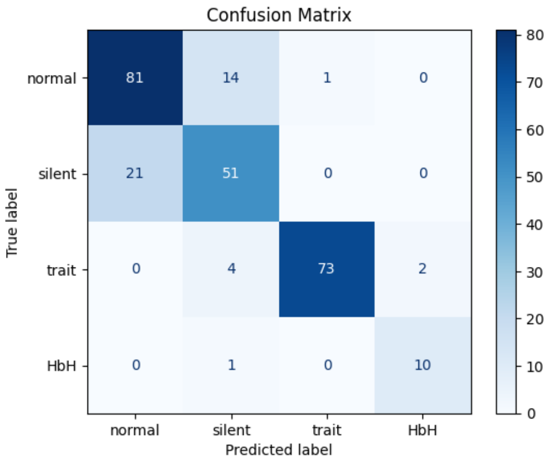
Table: Test Set without Oversampling with first 2 variables **MCH** and **RDW**

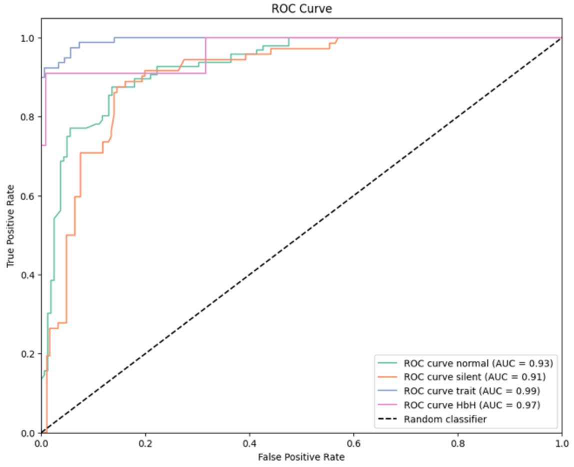
class	precision	recall	f1-score	n
normal	0.80	0.80	0.80	224
silent	0.75	0.75	0.75	224
trait	0.94	0.91	0.92	224
HbH	0.97	1.00	0.98	224
w. avg	0.86	0.86	0.86	869

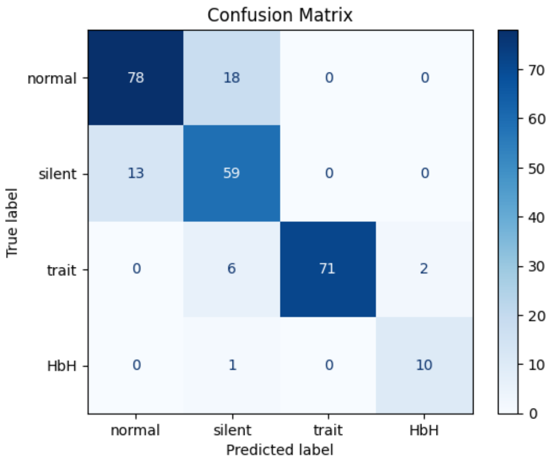
Table: Stratified Cross Validation with Oversampling (SMOTE) with first 5 variables **MCH**, **RDW**, **MCV**, **HB** and **MCHC**

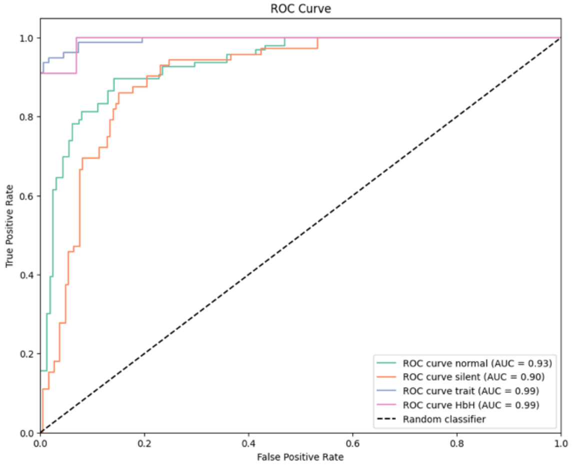
class	precision	recall	f1-score	n
normal	**0.86**	0.81	**0.83**	96
silent	0.70	**0.82**	**0.76**	72
trait	**1.00**	0.90	**0.95**	79
HbH	**0.83**	**0.91**	**0.87**	11
w. avg	**0.86**	**0.84**	**0.85**	258

Table: Test Set with Oversampling (SMOTE) with first 5 variables **MCH**, **RDW**, **MCV**, **HB** and **MCHC**









Diagnostic Models: logistic regression with principal component analysis |

For the third model, Principal Component Analysis (PCA) was used with the goal of reducing multicollinearity among the independent variables.

01
02
03
04
05

class	precision	recall	f1-score	n
normal	0.80	0.84	0.82	224
silent	0.72	0.67	0.69	166
trait	0.91	0.92	0.92	182
HbH	0.84	0.78	0.81	27
w. avg	0.81	0.81	0.81	599

Table: Stratified Cross Validation without Oversampling and 3 Principal Components

class	precision	recall	f1-score	n
normal	0.78	**0.83**	**0.80**	96
silent	**0.67**	0.68	0.68	72
trait	**0.99**	**0.91**	**0.95**	79
HbH	**1.00**	0.82	**0.90**	11
w. avg	**0.82**	**0.81**	**0.82**	258

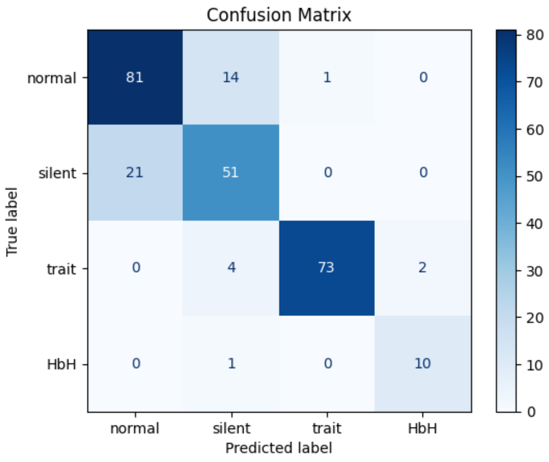
Table: Test Set without Oversampling and 3 Principal Components

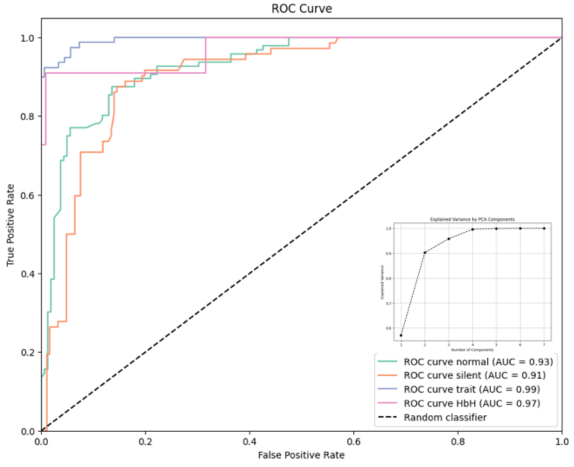
class	precision	recall	f1-score	n
normal	0.79	0.79	0.79	224
silent	0.74	0.75	0.75	224
trait	0.94	0.91	0.92	224
HbH	0.97	1.00	0.98	224
w. avg	0.86	0.86	0.86	869

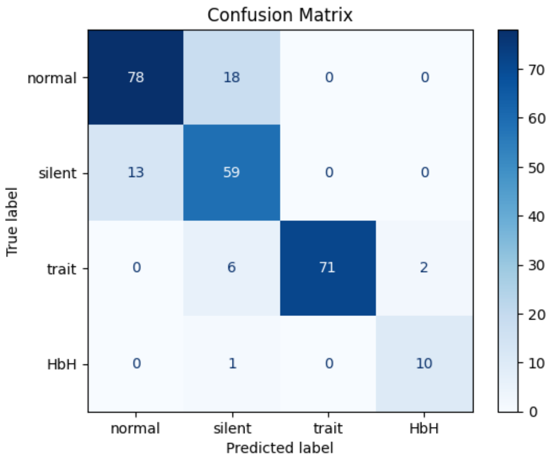
Table: Stratified Cross Validation with Oversampling (SMOTE) and 3 Principal Components

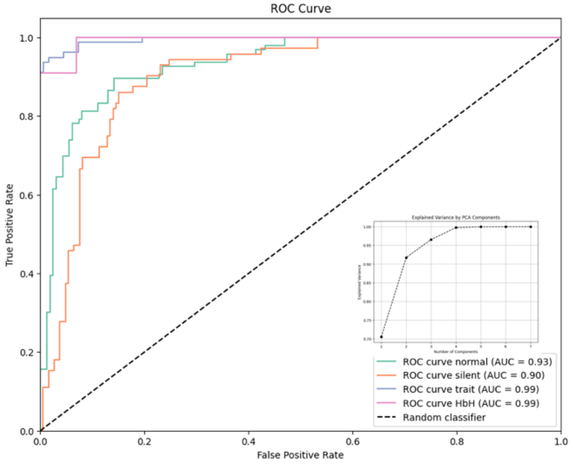
class	precision	recall	f1-score	n
normal	**0.80**	0.79	0.80	96
silent	**0.67**	**0.74**	**0.70**	72
trait	**0.99**	0.90	0.94	79
HbH	0.83	**0.91**	0.87	11
w. avg	**0.82**	**0.81**	**0.82**	258

Table: Test Set with Oversampling (SMOTE) and 3 Principal Components









Results

LOGISTIC REGRESSION (Test Data)	Normal (n = 96)				Silent (n = 72)				Trait (n = 79)				HbH (n = 11)			
Models	prec.	recall	F1 score	AUC	prec.	recall	F1 score	AUC	prec.	recall	F1 score	AUC	prec.	recall	F1 score	AUC
Manual Selection Accuracy = 0.81	0.80	**0.84**	0.82	0.92	0.68	0.69	0.68	0.87	0.95	0.90	0.92	**0.99**	1.00	0.73	0.84	**0.99**
Manual Selection (SMOTE) Accuracy = 0.81	0.82	0.79	0.81	**0.93**	0.66	0.76	0.71	0.87	0.97	0.86	0.91	**0.99**	0.71	**0.92**	0.80	0.98
RFE Accuracy = 0.83	0.79	**0.84**	0.82	**0.93**	**0.73**	0.71	0.72	**0.91**	0.99	**0.92**	**0.95**	**0.99**	0.83	0.91	0.87	0.97
RFE (SMOTE) Accuracy = 0.83	**0.86**	0.81	**0.83**	**0.93**	0.70	**0.82**	**0.76**	0.90	**1.00**	0.90	**0.95**	**0.99**	0.83	0.91	0.87	**0.99**
PCA Accuracy = 0.82	0.78	0.83	0.80	**0.93**	0.67	0.68	0.68	**0.91**	0.99	0.91	**0.95**	**0.99**	**1.00**	0.82	**0.90**	0.97
PCA (SMOTE) Accuracy = 0.82	0.80	0.79	0.80	**0.93**	0.67	0.74	0.70	0.90	0.99	0.90	0.94	**0.99**	0.83	0.91	0.87	**0.99**

Table: Results obtained using Logistic Regression with different techniques

01
02
03
04
05

Conclusions

From diagnostic models' results these considerations emerged:

- All the values of the **CBC** are significant with respect to the target variable.
- The tests on the **HbH** class did not provide significant insights, as the dataset for this class was small.
- The use of **Cross-Validation** showed that **Logistic Regression** effectively discriminates between the classes, achieving an **accuracy above 80%** across all models.
- In general, the results confirm the hypothesis and findings from the exploratory analysis: the ***trait*** and ***HbH*** classes are better discriminated compared to the ***normal*** and ***silent*** classes.
- The **RFE** model trained with **SMOTE** achieves the best performance in discriminating between the **Normal** and **Silent** classes, with a recall for this last class of 0.82.

01
02
03
04
05