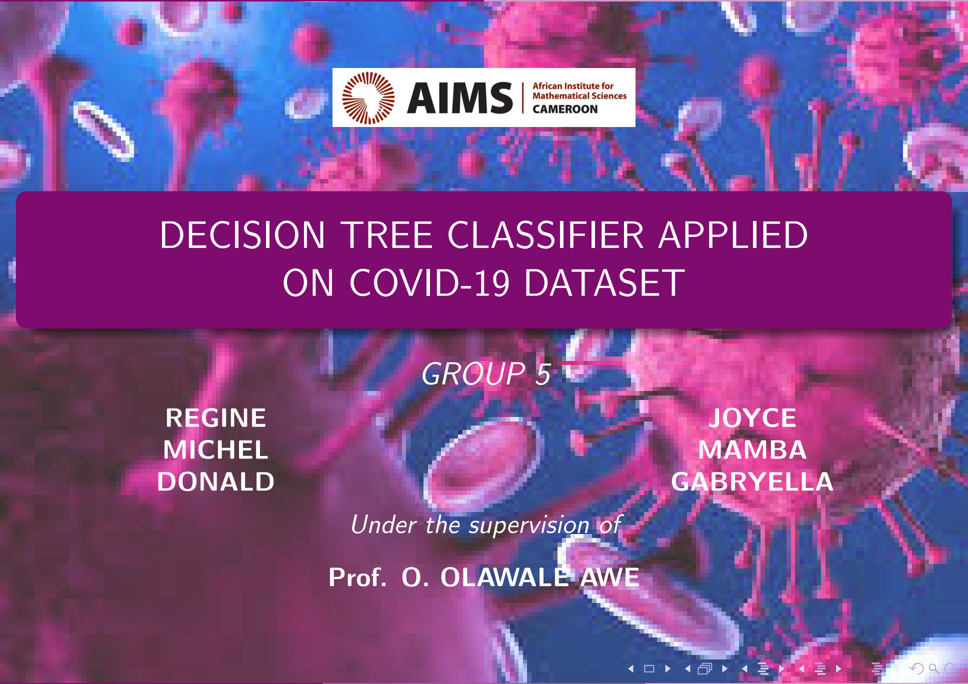


AIMS | African Institute for Mathematical Sciences CAMEROON

DECISION TREE CLASSIFIER APPLIED ON COVID-19 DATASET

GROUP 5

REGINE MICHEL DONALD

JOYCE MAMBA GABRYELLA

Under the supervision of

Prof. O. OLAWALE AWE

Contents

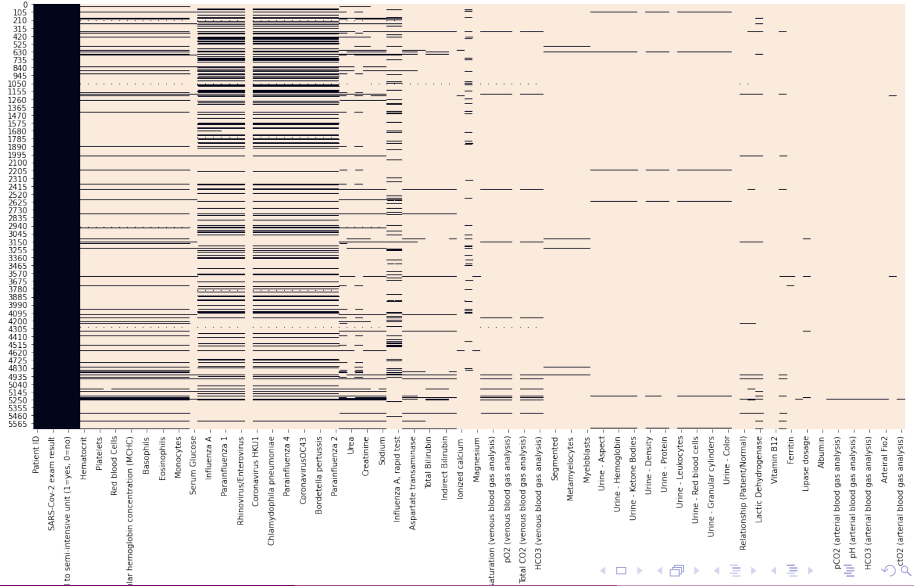
1. INTRODUCTION
2. DATA EXPLORATION
3. TRAINING MODEL
4. MODEL EVALUATION
5. CONCLUSION

Introduction

Covid-19 is a globally known, highly contagious and quite deadly respiratory illness caused by SARS-CoV-2, short for severe acute respiratory syndrome coronavirus 2. Due to the fact that this virus is highly contagious and can also lead to death, it is of utmost importance that it is detected in its early stages so that treatment can be be started promptly before it progresses and cause complications There are a number of tests that have been developed to determine whether a person is infected or not but these test are quite costly and can sometime take too long before the results can be receved.

In light of this information, this project aims to build a machine learning model that, given a patients current symptoms, age and medical history, predicts whether the patient is infected with the virus or not with the highest precision. This model was trained on a dataset provided by the Mexican government that contains anonymized patient related information including pre-conditions

Visualization of Dataset before tiding



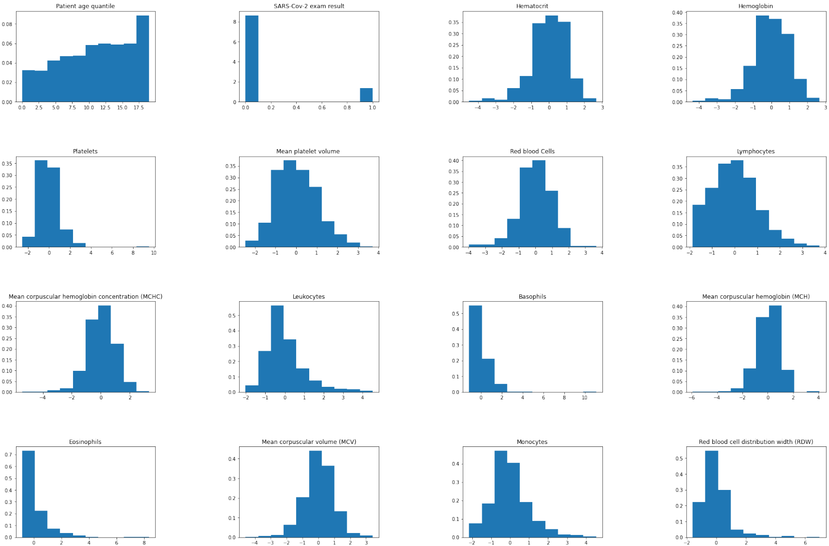
Preprocessing and Data cleaning

Shape of the data before: $(5644, 111)$

1. **Handling missing values**: Drop columns which have proportion of missing values greater than 0.7, and drop line where there is missing values.
2. **Label encodage**: Encode negative = 0, positive = 1, non detect= 0, detect = 1.
3. **Features engineering**: Create a new column that groups all the columns of diseases.

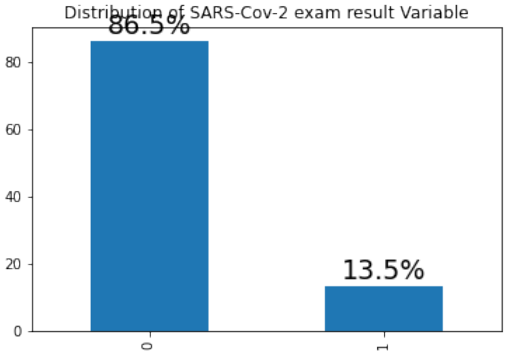
Shape of the data after: $(598, 16)$

Visualization of each variable after preprocessing



Method

1. Splitting
2. Standardization
3. Training and optimization(GridSearchCV)
4. Handling imbalanced data
5. Evaluation of model with best hyper-parameter
6. ROC- AUC curve, confusion matrix and learning curve
7. Interpretation



Splitting, Standardization and training

X_train = 70 % and y_test = 30%
We use Standard Scaler to scale our Data after split.
We Compare Decision Tree Classifier with Logistic Regression.
We apply GridSearchCv and find the best Hyper-Parameter for each model and the Result is :

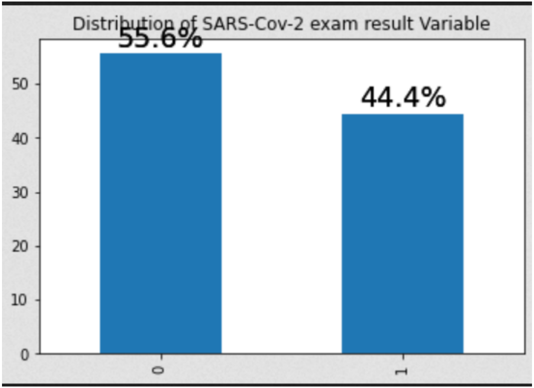
- **Decision Tree Classifier**: criterion =”gini”, ”Max_depth” = 5 , min_sample_leaf =2, min_sample_split=10.
- **Logistic Regression** : C = 10, penalty =l2.

Note : By specifying the Hyper parameters of the decision tree, we are applying Pre-pruning that helps to avoid Overfitting.

Handling Imbalance and evaluation

Our data is imbalanced. To handle the imbalance we use oversampling methods :

1. SMOTE
2. ADASYN
3. BorderlineSMOTE
4. RamdomOversample



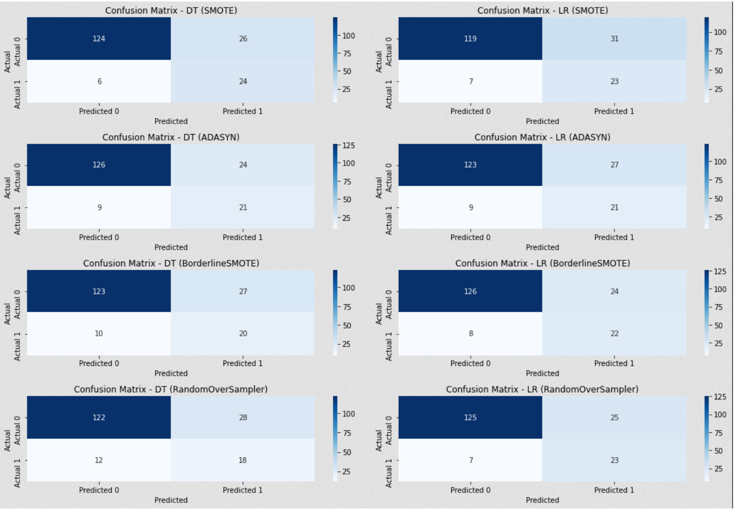
Performance Metrics

Resample	Model	AS	BLC	F1	P	Sensitivity
SMOTE	DT	0.822222	0.813333	0.600000	0.480000	0.800000
ROS	LR	0.822222	0.800000	0.589744	0.479167	0.766667
BSMOTE	LR	0.822222	0.786667	0.578947	0.478261	0.733333
ADASYN	DT	0.816667	0.770000	0.560000	0.466667	0.700000
SMOTE	LR	0.788889	0.780000	0.547619	0.425926	0.766667
ADASYN	LR	0.800000	0.760000	0.538462	0.437500	0.700000
BSMOTE	DT	0.794444	0.743333	0.519481	0.425532	0.666667
ROS	DT	0.777778	0.706667	0.473684	0.391304	0.600000

Table: Performance Metrics for Different Resampling Methods and Models

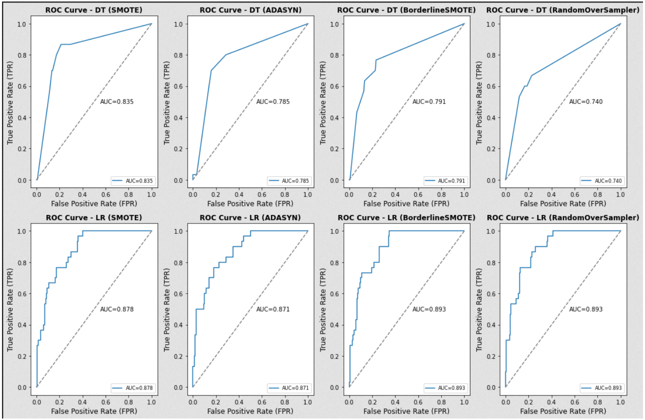
The best resampling Method for Decision Tree is SMOTE.

Confusion matrix



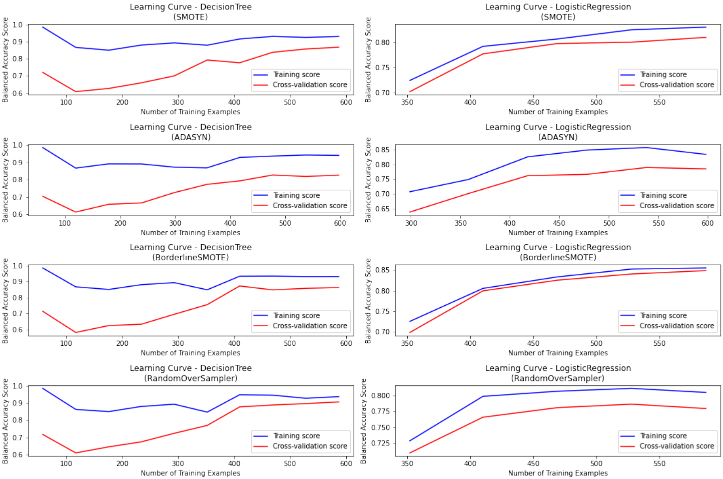
The Confusion Matrix for each model with different resampled method.

ROC-AUC curve



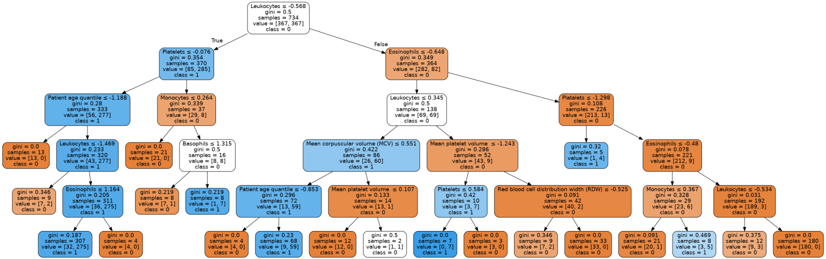
The ROC-AUC Curve of LR for all the resampling method is better than the DT.

Learning curve



The training and validation curve tends to be convergent and this is an indication that the model is learning from the underlying patterns in the data very well.

Tree with SMOTE resampling method



Our tree has depth = 5, 2 leafs and the the criteria for splitting is Gini. The Variable with the best Gini Impurity is a node root(Leukocytes).

Conclusion

- Based on the results of this work, we can conclude that decision trees can be used for the diagnosis of Covid-19 with a better performance compared to logistic regression when using SMOTE and ADASYN oversampling methods.
- As return time of test results is of utmost importance in the diagnosis of Covid-19, this work provides strong motivation for the adoption of Machine Learning methods, particularly the use of decision trees to provide prompt and trustworthy results for Covid-19 diagnosis which could help save lives.

Thank you!