

LIFE EXPECTANCY

Adewale Taofiq Ariwo-Ola

Strategic Thinking

Capstone CA2

INTRODUCTION

The aim of the capstone project is to apply skills obtained during my studies for Higher Diploma Data Analytics for Business to evaluate a possibility to predict a chosen target variable based on other independent variables. I will use data exploratory analysis (EDA) and statistical methods to understand patterns and trends of a dataset and apply a predictor machine learning model on the processed data.

HYPOTHESIS

Can socioeconomic, health, environmental and geographical location criteria be used to predict if a populations life expectancy is low or high?

GENERAL GOAL

The goal of this project is to analyze the dataset and generate a synthetic data to be used to construct a machine learning model that can predict life expectancy in regions of countries around the world based on factors that are represented in the dataset.

DEFINING THE PROBLEM STATEMENT

I am creating a categorical machine model which can predict life expectancy of a population in regions of countries around the world.

SUCCESS CRITERIA/INDICATORS

- I was able to establish correlation between life expectancy and socioeconomic, geographical location and health conditions.
- Random Forest Model was used to predict if a populations life expectancy is low or high with an outcome of 100% accuracy.
- Independent features used for predicting country status for the machine learning model were extracted based on their importance.
- SHAP (Shapley Additive exPlanations) was used to interpret the machine learning model used and understand the impact of individual features on the model's prediction.

TECHNOLOGIES USED

CODINGS: PYTHON LIBRARIES


- NumPy
- Matplotlib
- Pandas
- Seaborn
- Sklearn
- SciPy



MACHINE LEARNING MODELS:

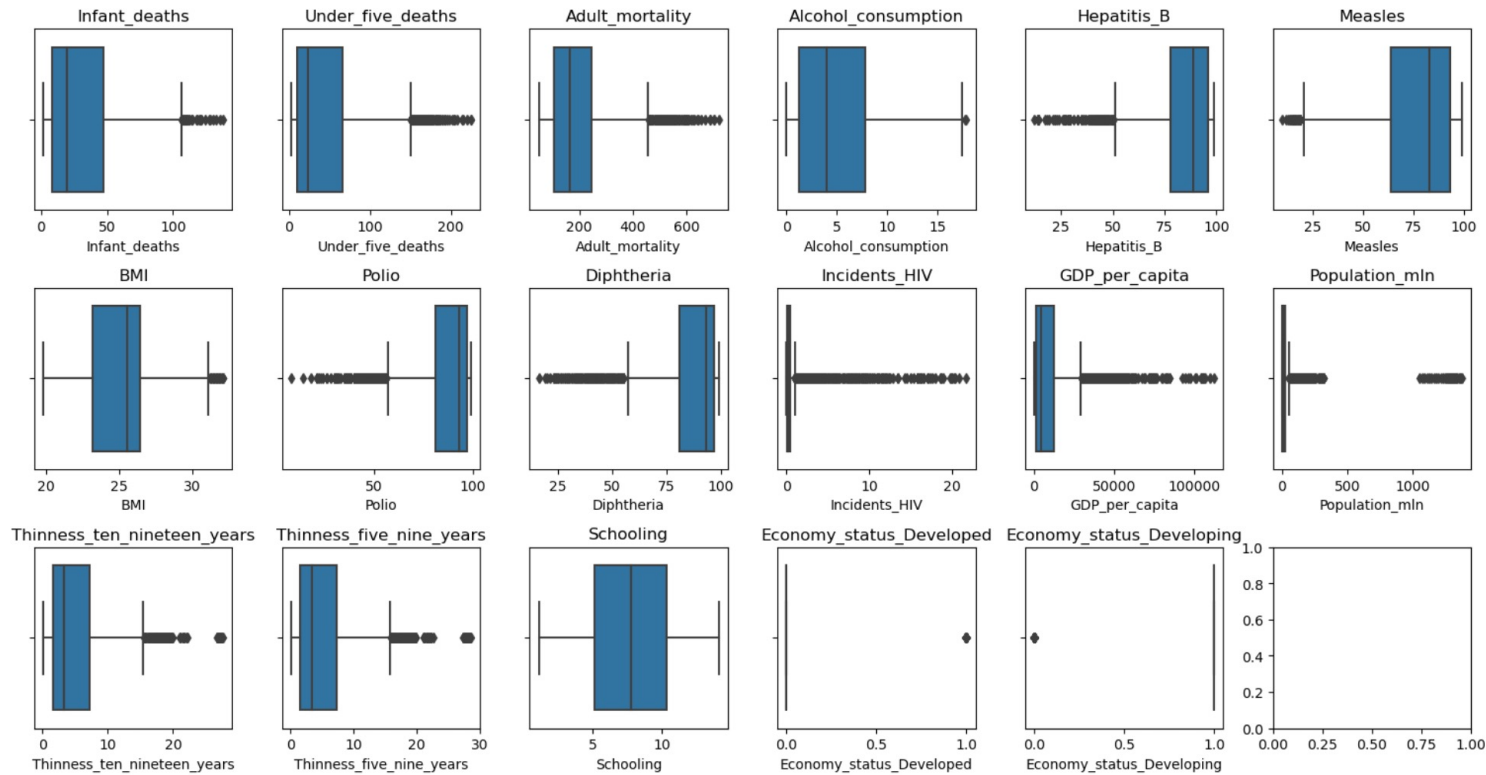
- RANDOM FOREST CLASSIFIER
- DECISION TREE

DATA PROCESSING

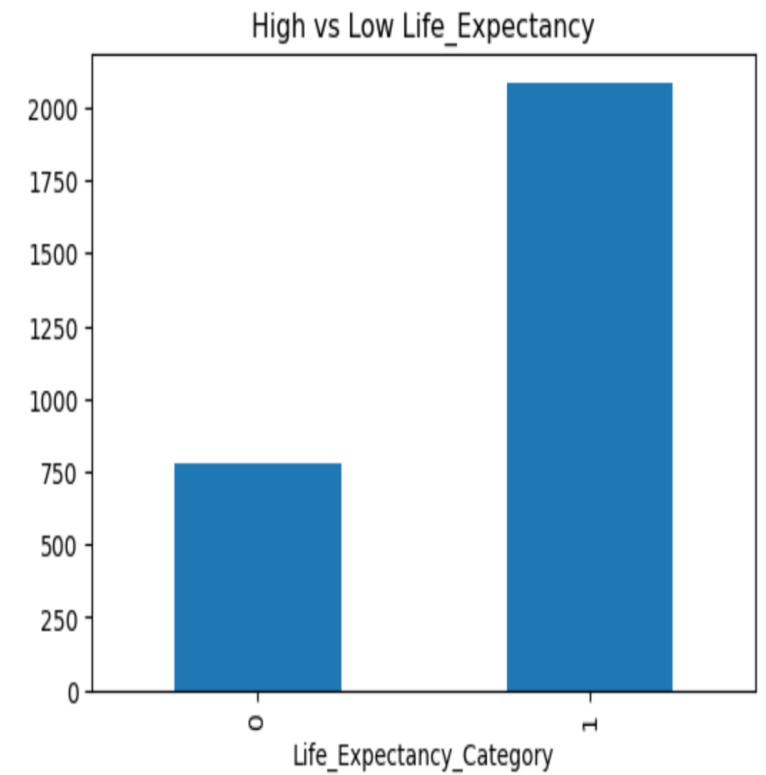
- Number of variables reduction
 - Data points categorisation
 - New column creation
 - Data types adjustment
 - Analysis of duplicates and missing values
 - Data shuffling and Standardization
- 

EXPLORATORY DATA ANALYSIS (EDA)

1. Box plots visualisation of the data



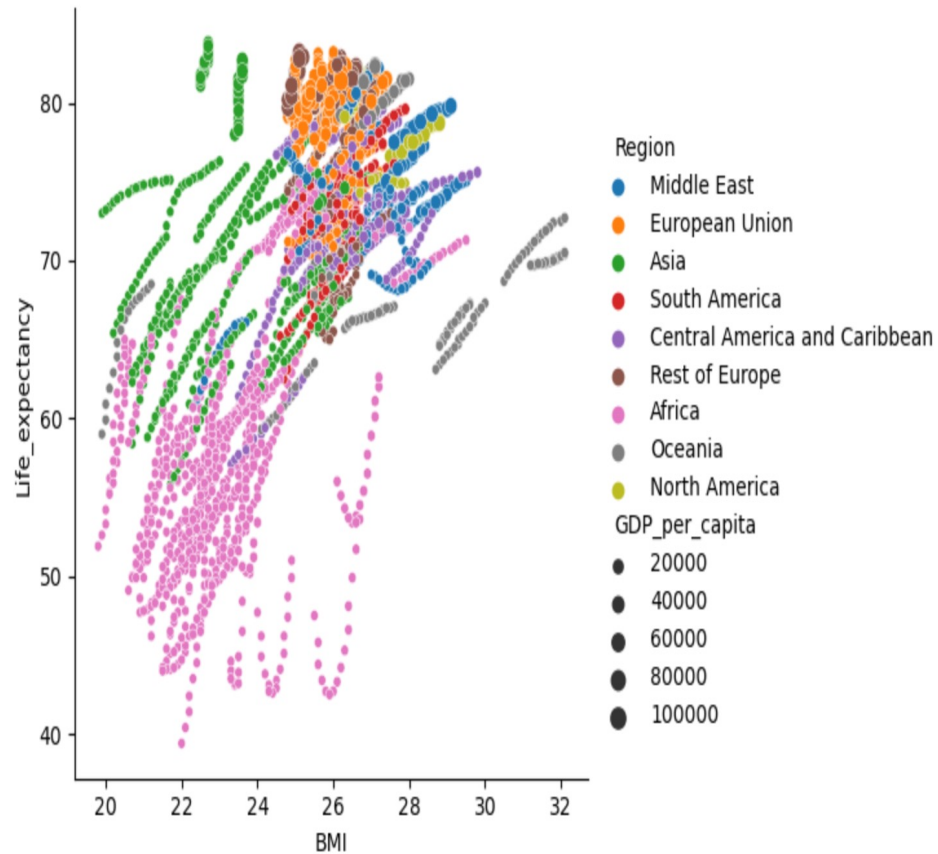
2. Distribution of the target variable



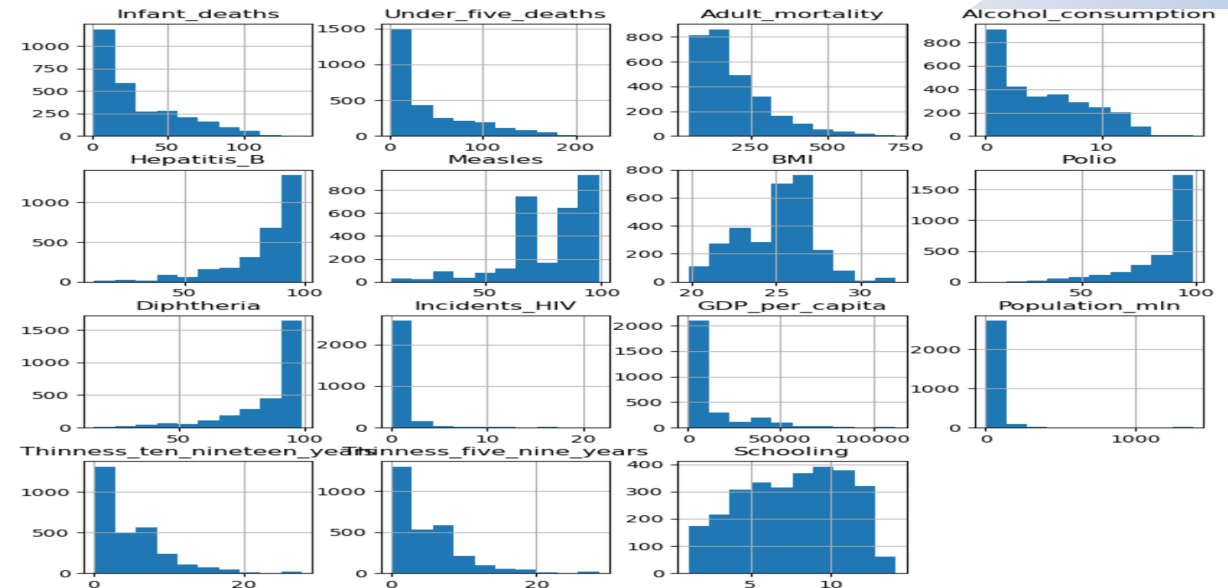
VISUAL EXPLORATORY DATA ANALYSIS

2.Bar chat distribution of categorical columns

Life expectancy correlation with BMI, Region and GDP per capital.1.



3.Histogram distribution for Continuous variables

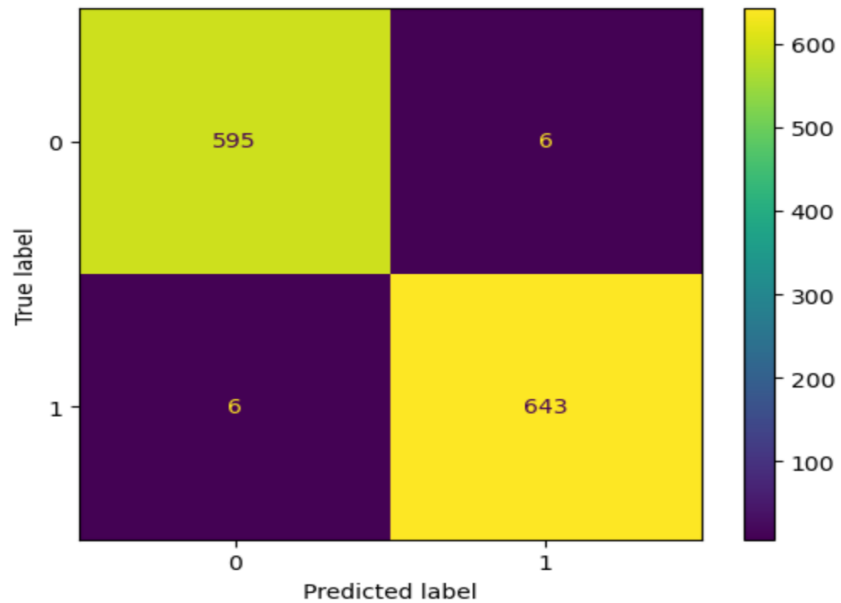


MODELLING PHASE 1

Decision Tree

Results:

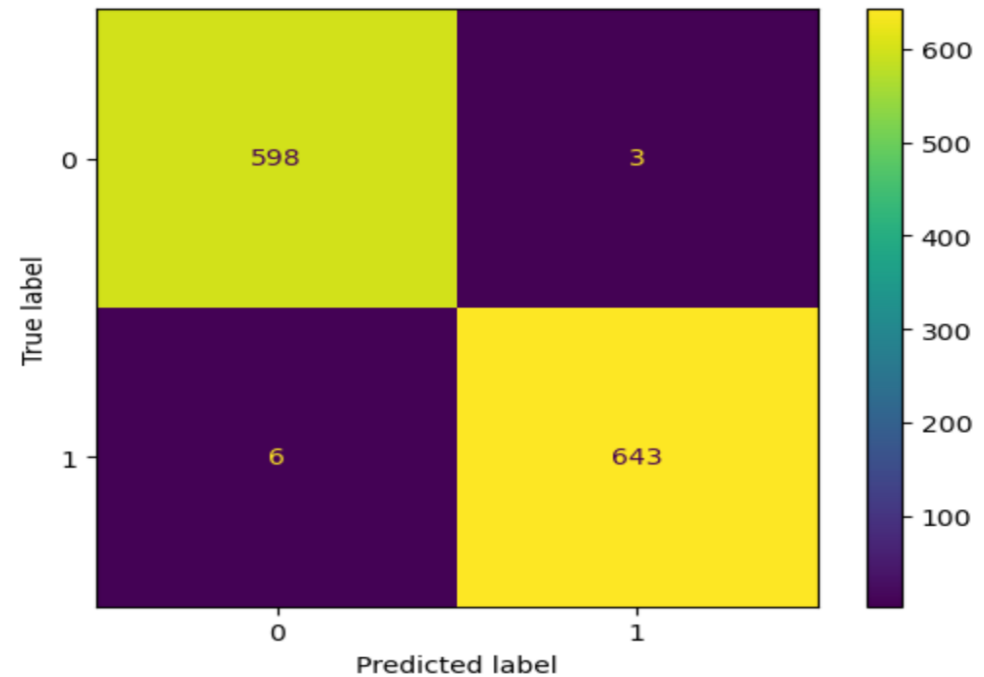
- Training Accuracy 100%
- Testing Accuracy 99%



Random Forest

Results:

- Training Accuracy 100%
- Testing Accuracy 99%

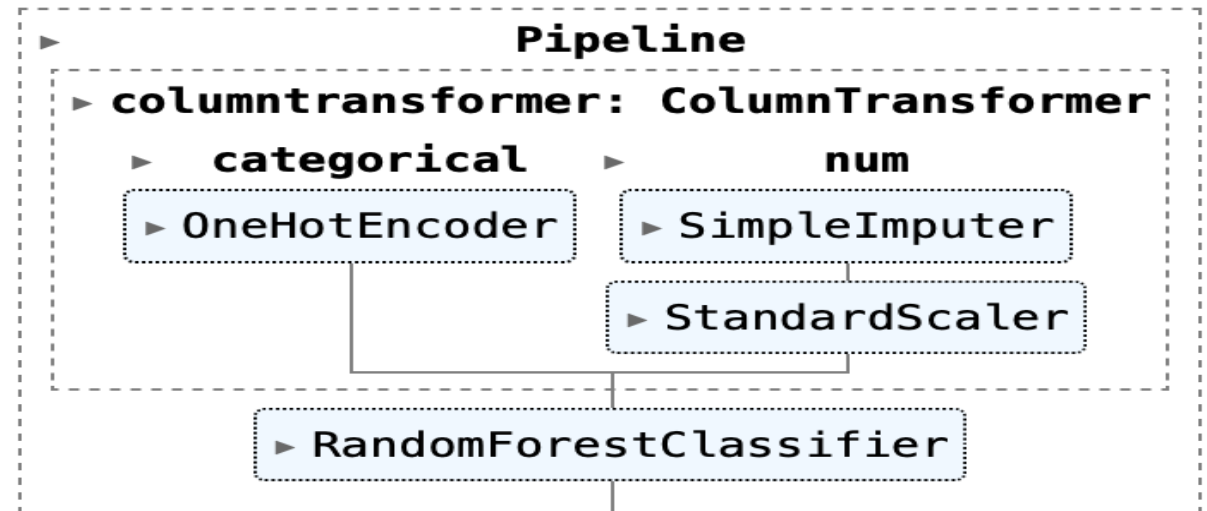


MODELLING PHASE 2

1. Score evaluation of models

	model	accuracy	precision	recall	f1-score
0	Logistic_Regression	0.99	1.00	0.98	0.99
1	SVC	0.98	0.99	0.98	0.99
2	Random_Forest	0.99	0.99	0.99	0.99
3	Decision_Tree	0.99	0.99	0.99	0.99

2. Pipeline for processing

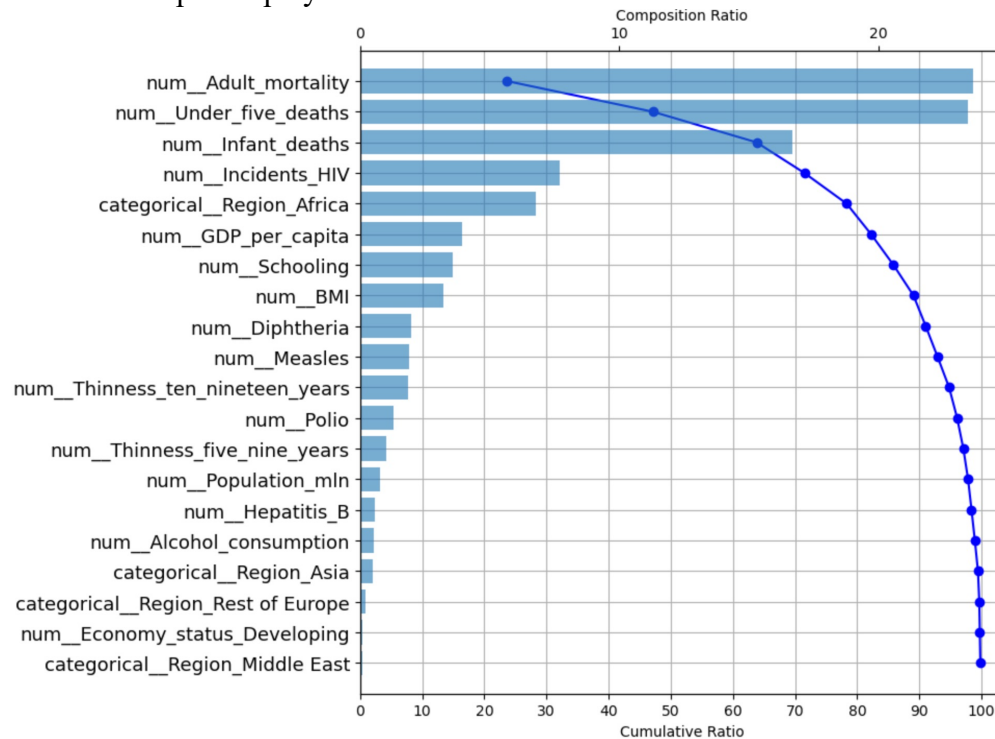


The Pipeline class is a class that allows ‘gluing’ together multiple processing steps into a single scikit-learn estimator

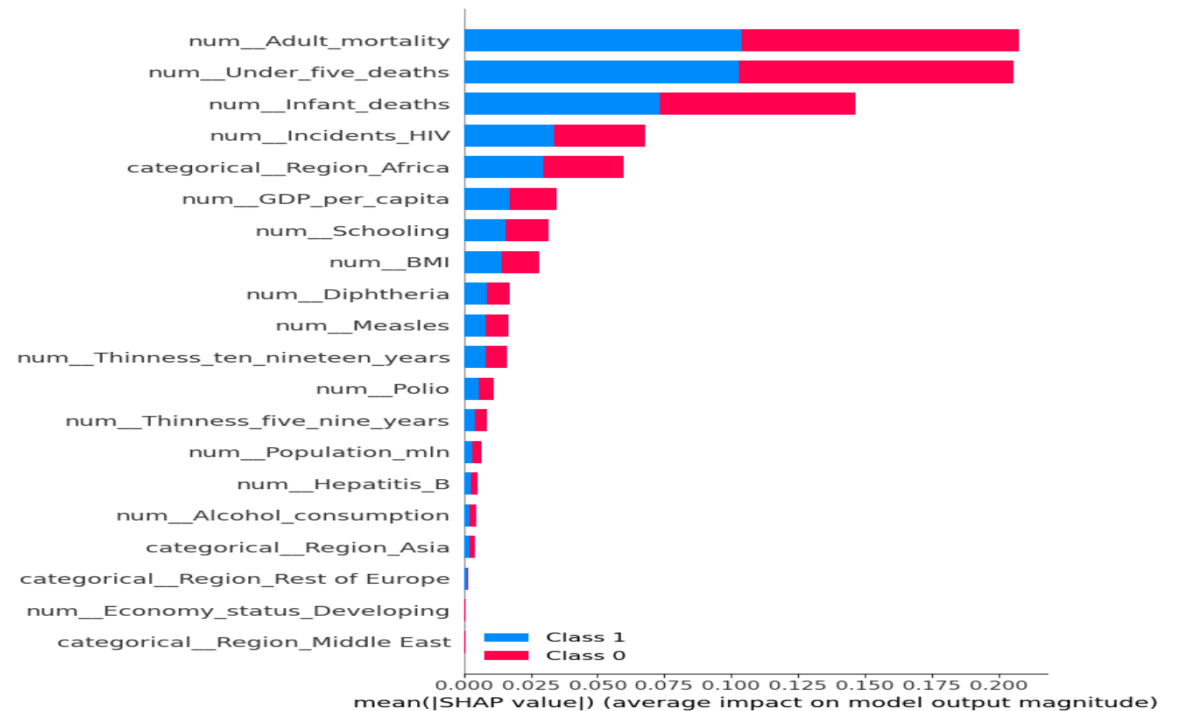
SHAP (SHAPLEY ADDITIVE EXPLANATIONS)

Is there a correlation between chosen socioeconomic conditions and life expectancy based on a statistical data analysis?

1. Waterfall plotdisplay the contributions of individual features



2. SHAP Summary bar plots to visualize features importance



Summary

CHALLENGES ENCOUNTERED AND STRATEGIES USED TO OVERCOME THEM

Inconsistencies in the min and max values in the data



Scale the data by using StandardScaler algorithm

Class Imbalance in Target Variable



SMOTE

CONCLUSION

In this project using CRISP DM methodology, I applied multiple data management techniques on a specific data set to analyse its informational value to support the hypotheses and then successfully developed two Machine Learning models that are able to predict with 100% accuracy the set target feature based on the rest of independent features .

This project is considered a success based on the predictive results obtained. The hypothesis is considered True.

Thank You

