BANL-6625-01

GROUP 3

Submitted by:

SHASHIKUMAR GADUSU AND RENUKA CHOWDARY PARVATHANENI

INTRODUCTION:

The objective of the dataset is to predict the outcome of the upcoming election by employing various regression methods. Within the dataset, there are variables like State, voter id, gender, family size, political party, age, salary and voted last election. Multiple models are trained using this data, and their effectiveness is assessed through a performance comparison.

DATA PREPROCESSING:

Let's now explore the data and understand how the data is distributed. Let's visualise it by using histogram. In the visualization we can observe the correlation among different variables present in the vote dataset.

We have omitted the duplicate variables and then we considered only three variables that are age, salary and voted last election. We next change the categorical values and put dummy variables or numerical values into the dataset for further analysis. To ensure integrity of our analysis, we divide the dataset into (70%) training test and (30%) test set by Data Partition function.

MODEL BUILDING:

After reviewing the code, we had used a different classification techniques Logistic regression, SVM, KNN, Decision tree and Random Forest. We had trailed all the classification techniques to predict the optimum which can be obtained the desired design parameters and best set of operation conditions to satisfy our output. A depart of variables we had considered out only Age, salary because those two variables are solidly correlative for predicting the output which is Voted_Last_Election.

MODEL EVALUATION:

The effectiveness of the analysis is determined by the performance of the model. To analyses our performance, we are using different metrics such as accuracy, precision, F1_Score, recall and roc.

Models	Accuracy	Precision	Recall	F1_Score	Roc
Logistic Regression	0.5163442	0.5221239	0.7612903	0.6194226	0.5077170
SVM	0.5010007	0.5112033	0.7948387	0.6222222	0.4906514
KNN	0.4896598	0.5062500	0.5225806	0.5142857	0.4885003
Decision Tree	0.5170113	0.5170113	1.0000000	0.6816183	0.5000000
Random Forest	0.4909940	0.5082192	0.4787097	0.4930233	0.4914267

- Accuracy compares the predicted values with true values which results in overall accuracy rate in our model Decision tree performs best with higher accuracy with 0.5170113 and lowest being KNN model with 0.489
- Precision means the percentage of positive prediction that were correct.in our case Decision tree performs higher precision with 0.517 and lowest being KNN model with 0.506
- Recall is the percentage of actual positives that were correctly predicted. Here Decision tree performs higher Recall with 1.000 and Random Forest being lowest with 0.478.
- F1_Score combines precision and recall into a harmonic mean. In our case Decision Tree performance is higher with 0.681 and lowest with Random Forest 0.493
- Logistic Regression performs higher roc with 0.507 and lowest roc with KNN model 0.488.

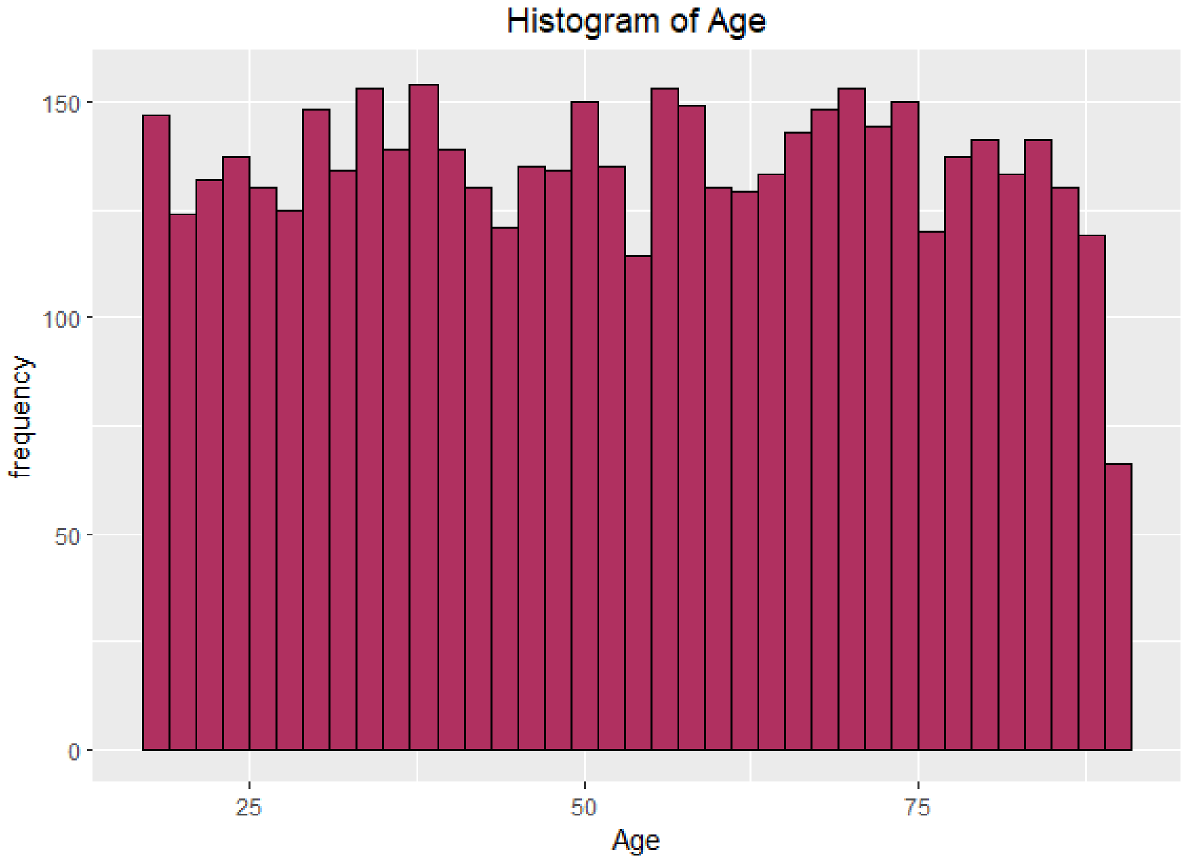
Identification of Best Performing Model:

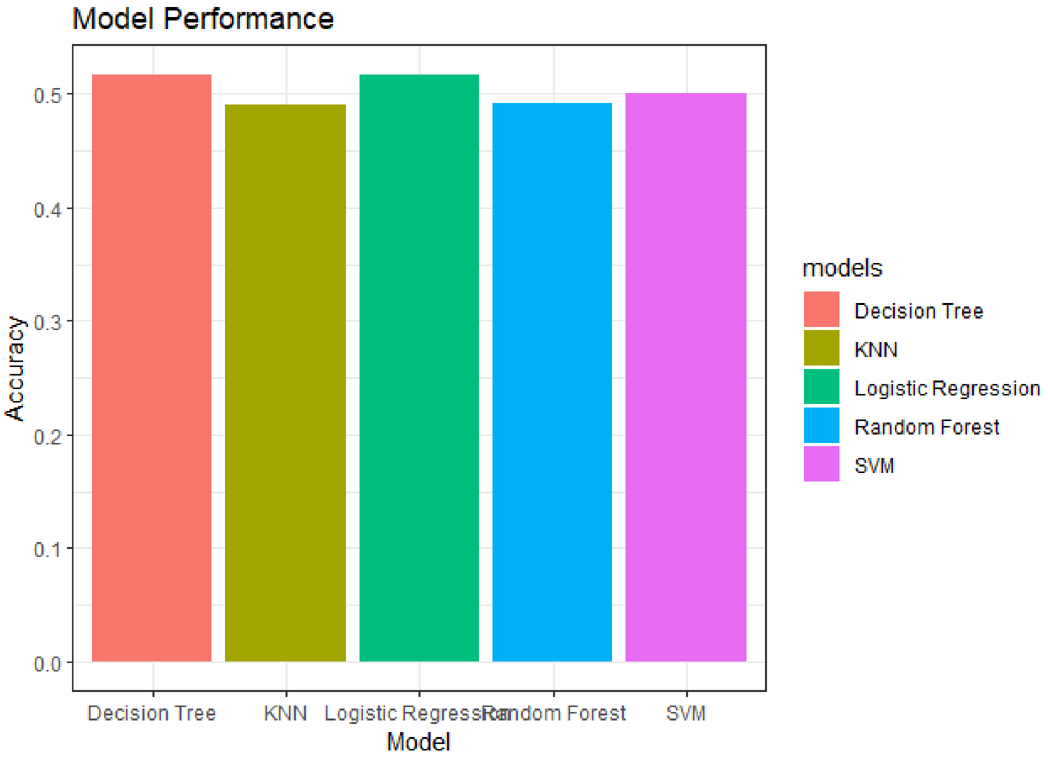
Keen observation from the model evaluation Decision Tree model performs well in all the metrics being with accuracy 0.517 and precision with 0.517 & Recall with 1.000 & F1_Score with 0.68 & roc with 0.50.

CONCLUSION:

In conclusion after analysing a performance metrics resulted in decision tree classification model with attributes age and salary significantly influenced Voted_Last_Election which determines who would likely vote in the future elections. Evaluation process and visualizations helped in determining the best performance model.

VISUALIZATION:


Histogram of Age
frequency
Age


Model Performance
Accuracy
Decision Tree
KNN
Logistic Regression
Random Forest
SVM
Model
models
Decision Tree
KNN
Logistic Regression
Random Forest
SVM