DIAGNOSTIC AND PREDICTIVE STATISTICAL ANALYSIS OF CONTRIBUTORY RISK FACTORS FOR CEREBROVASCULAR ACCIDENT (CVA)

Nilambari Arvind Jagtap
Tuljaram Chaturchand College
9 PUBLICATIONS **1** CITATION

DIAGNOSTIC AND PREDICTIVE STATISTICAL ANALYSIS OF CONTRIBUTORY RISK FACTORS FOR CEREBROVASCULAR ACCIDENT (CVA)

N. A. Jagtap Ph.D. Research Scholar Department of Statistics Shri Jagadishprasad Jhabarmal Tibrewala University Jhunjhunu, Rajasthan India

Dr. A. S. Jagtap Professor Department of Statistics Tuljaram Chaturchand College of Arts, Science and Commerce Baramati, Maharashtra India

Abstract:

The brain's counterpart of a heart attack is a CVA, also referred to as a stroke. In the modern world, it is one of the main causes of death. Every year, 15 million people experience a stroke. Of them, 5 million passes away and a further 5 million become permanently crippled, burdening families and the community. Under 40-year-olds rarely experience strokes, but when they do, high blood pressure is the primary contributing factor. Reduced stroke morbidity and death are still largely dependent on risk factor identification and management. This study successfully models three well-known classification techniques, namely logistic regression, decision trees, and random forests, and also evaluates their performance measures to come to the best predictive model for CVA. The main goal of this research is to examine the behaviours of the risk variables that contribute to CVA and gauge how much of an impact they have on it. The training and test datasets for stroke were carefully examined, and exploratory data analysis was carried out. Through the feature selection process, the most effective and precise variables needed to predict stroke in an individual were collected. Based on the variables obtained, aspects that affect the prognosis of the disease were then identified. This processed data is subjected to predictive modelling using a variety of categorization models, including Random Forest, Decision Tree, and Logistic Regression.

Key Words: Stroke, patient characteristics, Risk factors, Independence tests, Classification models, Survival analysis, logistic regression, decision tree, random forest

Introduction

A cerebrovascular accident (CVA), often known as a stroke, happens when the blood supply to a portion of the brain is cut off or diminished, depriving the brain's tissue of oxygen and nutrients. In minutes, the death of brain cells starts. It is a medical emergency, so getting help right now is essential. CVA is characterised by the sudden onset of a neurologic deficit that can be linked to a specific vascular aetiology. When blood supply to the brain is impaired, a CVA, stroke, or brain attack occurs. The majority of the time, a clot narrows a blood vessel, preventing blood flow to certain areas of the brain. A brain blood vessel bursting occurs less frequently. The brain receives oxygen and nutrients through the blood. Brain cells begin to die within minutes after a blood flow interruption, which results in a stroke. Depending on where in the brain they occur and how much brain tissue is injured, strokes have various impacts. When a clot plugs an artery, it can cause ischemic strokes by preventing blood flow to certain areas of the brain. 87 percent of strokes are ischemic.

Two subtypes exist: Thrombotic: In an artery that delivers blood to the brain, a clot (thrombus) develops. This kind is frequently connected to artery plaque (fatty deposits) (atherosclerosis). The carotid arteries, which go up each side of the neck to the brain, are especially prone to plaque build-up. Embolus: An embolus is a clot that develops outside the brain, frequently in the heart. It travels through the bloodstream to the brain and lodges itself in a blood artery there. Haemorrhagic strokes account for 13 percent of strokes. They happen when a blood artery in or near the brain that is weak bursts. The brain or the region around it becomes splattered with blood. Parts of the brain are deprived of oxygen and nutrients due to the blood pooling. Additionally, the pressure on the brain is damaging. The location of the bleeding (haemorrhage) determines the type of hemorrhagic stroke that occurs.

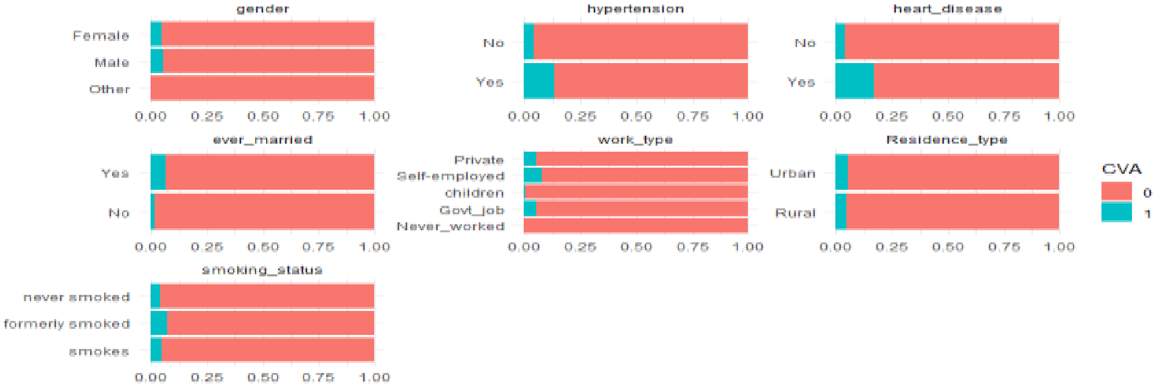
Intracerebral haemorrhage: When a brain artery rupture, blood leaks into the nearby brain tissue. Frequently, high blood pressure is to blame. The most typical type is this one. An artery on the underside of the brain rupture, leaking blood into the space between the brain and the skull, causing a subarachnoid haemorrhage (subarachnoid space). An aneurysm, which is a weak or bulging area in the arterial wall, may rupture and cause this type of bleeding. Symptoms of a cerebral vascular accident include: difficulty walking, light headedness, loss of coordination, and loss of balance. difficulty understanding what others are saying or communicating themselves, a face, leg, or arm that is numb or paralysed, often only on one side of the body, distorted or dimmed eyesight Certain CVA risk factors are uncontrollable. These include family history, age, and gender. However, a lot of CVA risk factors are connected to lifestyle. Making a few small lifestyle modifications can lower the risk of stroke for anyone. High blood pressure, cigarette smoking, diabetes, high blood cholesterol, heavy drinking, lack of regular exercise, and obesity are lifestyle-related variables that raise your risk of CVA.

We have made a sincere effort in this work to statistically analyse the effects of several risk factors (such as age, BMI, average glucose levels, etc.) on CVA. We have employed many machine learning techniques and data mining methods to achieve this. Through the application of comparison analysis and the evaluation of predictive modelling, we have investigated the influence of various risk factors. We were able to depict the underlying trends in the data through data visualisation.

Objective

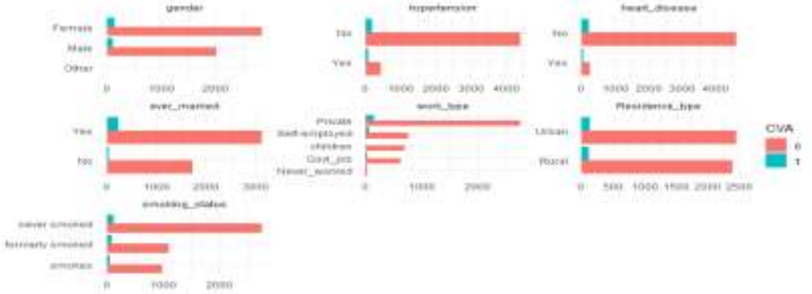
- Analyse a three-dimensional contingency table to determine whether gender, heart disease, and CVA are completely, jointly, and conditionally independent of one another.
- To test for complete, joint, conditional independence of Locality, Hypertension and CVA by analyzing Three-Dimensional Contingency table
- To Build a logistic regression-based classification model for the response variable CVA and evaluate its model performance
- To Construct a well pruned Decision tree-based classification model for the response variable CVA and evaluate its model performance
- To Build a Random Forest based classification model for the response variable CVA and evaluate its model performance
- To Compare the three classification methods namely Logistic regression, Decision tree and Random Forest, to assess their relative performance based on their performance metrics.

Graphical Statistical Analysis

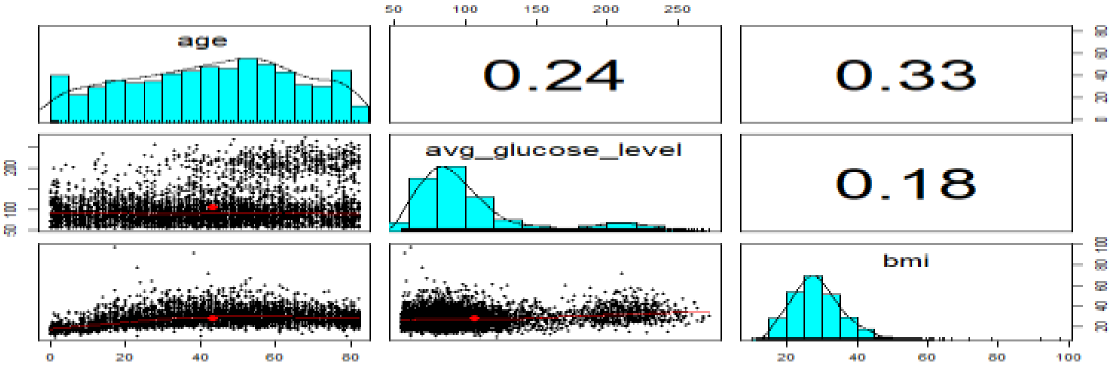


The stacked bar graphs shown above allow us to compare totals and identify sudden changes at the response class level, which are most likely to have an impact on category totals. As compared to the other characteristics in their respective categories, we see that the proportion of CVA positive class is much greater in adults with hypertension, heart disease, marriage, self-employment, and a history of smoking.

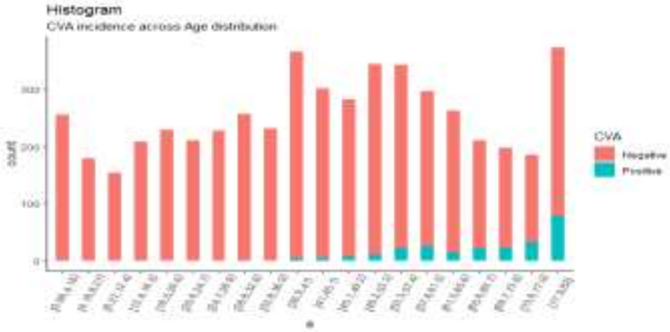
ISSN: 2278-4632
Vol-12 Issue-11 No.03 November 2022



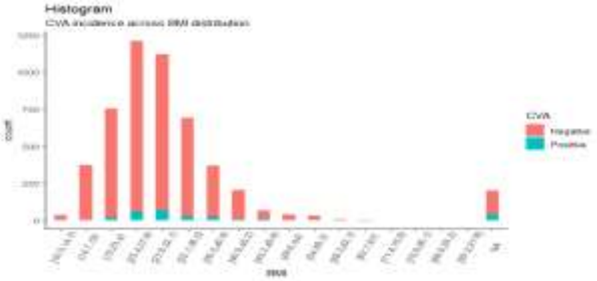
We can get the overall picture of our dataset by referring to the grouped bar charts above for all the categorical variables in our dataset. The total number of observations for each attribute of each categorical variable may be seen.



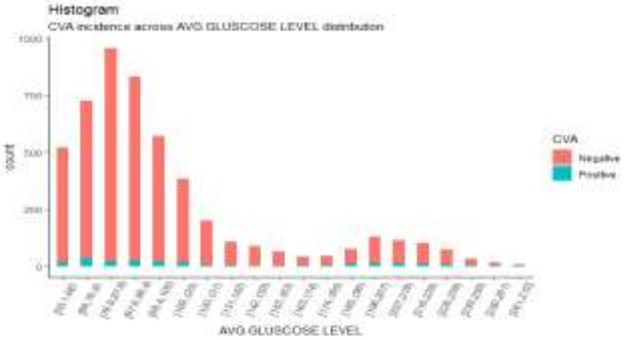
Age, BMI, and average blood glucose levels are three continuous variables that are shown in the above paired scatter plot along with their distributions and associations. We see that whereas age and average blood glucose level have somewhat skewed distributions, BMI has a normal distribution. These three variables do not correlate very well with one another.

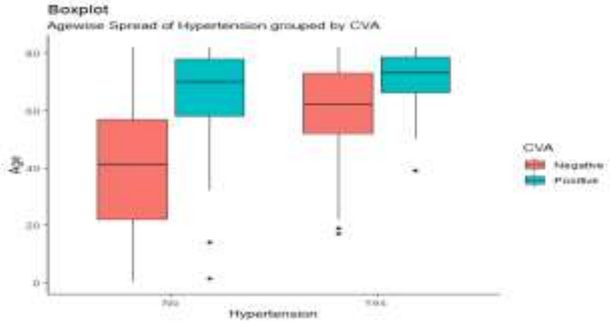


The proportion of CVA positive cases is seen to increase in the stacked bar graph above after the age of 41, and a sharp increase is visible from the age range [49.2, 53.3]. For age groups under 50, the percentage of positive instances is extremely small.

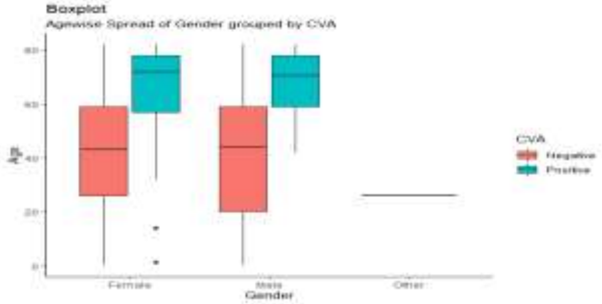


In the above stacked bar chart, we observe the proportion of CVA positive cases tend to be highest for the people with BMI in the range [27.8, 32.1] which corresponds to obesity interval.

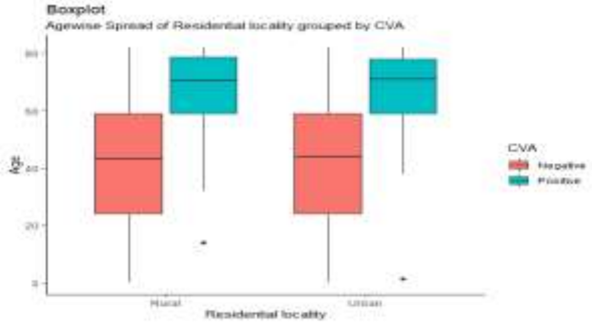




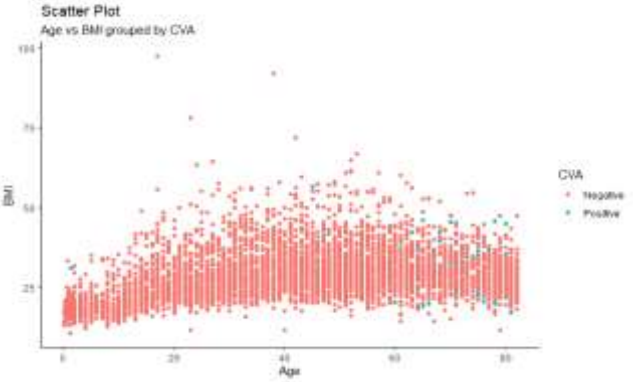
The majority of those who have hypertension are over 50 years old, according to this boxplot. In addition, the individuals who were diagnosed with CVA are well past the age of 60, regardless of their level of hypertension.



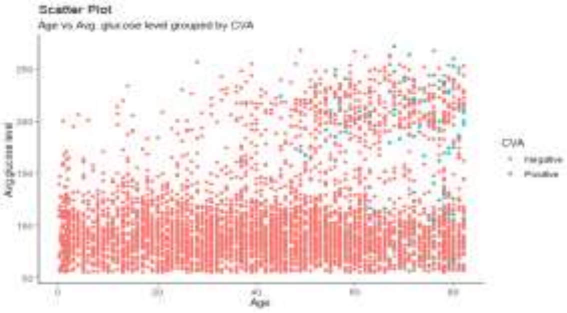
The negative class is seen to be distributed across the same age range in the boxplot shown above for both genders. Positive classifications are also distributed similarly in the older age ranges. In the female CVA positive class, there are two outliers. Only one observation belongs to the other category, and it is unfavourable.



The negative class of both places is shown in the aforementioned boxplot to be distributed across the same age range. Positive classifications are also distributed similarly in the older age ranges. In each of the CVA positive classifications, we observe one outlier.



According to the aforementioned scatter plot, the majority of CVA positive cases are found in those over 50 with BMIs under 60.



In this scatter plot we see that people with older age groups and abnormal avg. blood sugar levels are prone to CVA.

Statistical Analysis

Analysis of Three-Dimensional Contingency table to test for complete, joint, conditional independence of Gender, Heart disease and CVA.

Gender	Heart disease	Disease(CVA=1)	Disease(CVA=0)
Female	Yes	19	94
	No	122	2730
Male	Yes	28	134
	No	80	1819

A. To test if Gender, Heart disease and disease (CVA) are completely independent.

Hypothesis:

H_0: Gender, Heart disease and CVA are completely independent

Vs

H_1: Gender, Heart disease and CVA are not completely independent.

Statistics:

```
                        X^2          df       P(> X^2)
Likelihood Ratio     97.10303      4        0
Pearson              142.16948    4        0
```

Here **the p-value < 0.05**, hence there is strong statistically significant evidence to reject H_0. Hence, we **Reject H_0** at 5% l.o.s

Thus, Gender, Heart disease and CVA are not completely independent.

B. To test if Gender and Heart disease are jointly independent.

Hypothesis:

H_0 : Gender and Heart disease are jointly independent

Vs

H_1 : Gender and Heart disease are not jointly independent

Statistics:

```
                     X^2        df        P(> X^2)
```

```
Likelihood Ratio    59.34660     3       8.106849e-13
Pearson           91.04525     3       0.000000e+00
```

Result:

Here **the p-value < 0.05**, hence there is strong statistically significant evidence to reject $H_{0.}$ Hence, we **Reject H_0** at 5% l.o.s

Thus, Gender and Heart disease are not jointly independent.

C. To test if heart disease and CVA given Gender are independent.

Hypothesis:

H_0 : Heart disease and CVA given Gender are independent.

Vs

H_1 : Heart disease and CVA given Gender are not independent.

Statistics:

```
                    X^2        df      P(> X^2)
Likelihood Ratio     58.74312    2      1.754152e-13
Pearson              89.07520    2      0.000000e+00
```

Result:

Here **the p-value < 0.05**, hence there is strong statistically significant evidence to reject $H_{0.}$ Hence, we **Reject H_0** at 5% l.o.s

Thus, Heart disease and CVA given Gender are not independent.

Analysis of Three-Dimensional Contingency table to test for complete, joint, conditional independence of Locality, Hypertension and CVA

Locality	Hypertension	Disease(CVA=1)	Disease(CVA=0)
Urban	Yes	34	213
	No	101	2196
Rural	Yes	32	219
	No	82	2149

A. To test if locality, Hypertension and disease (CVA) are completely independent.

Hypothesis:

H_0 : Locality, Hypertension and disease (CVA) are completely independent

Vs

H_1 : Locality, Hypertension and disease (CVA) are not completely independent

Statistics:

```
                  X^2      df     P(> X^2)
Likelihood Ratio 61.33737    4     1.518452e-12
Pearson            82.43504    4     0.000000e+00
```

Result:

Here **the p-value < 0.05**, hence there is strong statistically significant evidence to reject $H_{0.}$ Hence, we **Reject H_0** at 5% l.o.s

Thus, Locality, Hypertension and CVA are not completely independent.

B. To test if Locality and Hypertension are jointly independent.

Hypothesis:

H_0 : Locality and Hypertension are jointly independent

Vs

H_1 : Locality and Hypertension are not jointly independent

Statistics:

```
                    X^2           df        P(> X^2)
Likelihood Ratio     61.10804       3        3.408385e-13
Pearson             82.37133       3        0.000000e+00
```

Result:

Here **the p-value < 0.05**, hence there is strong statistically significant evidence to reject H_0. Hence, we **Reject H_0** at 5% l.o.s

Thus, Locality and Hypertension are not jointly independent.

C. To test if Hypertension and CVA given Locality are independent.

Hypothesis:

H_0 : Hypertension and CVA given Locality are independent

Vs

H_1 : Hypertension and CVA given Locality are not independent.

Statistics:

```
                     X^2        df          P(> X^2)
Likelihood Ratio     59.74779    2           1.061373e-13
Pearson             81.33811     2         0.000000e+00
```

Result:

Here **the p-value < 0.05**, hence there is strong statistically significant evidence to reject H_0. Hence, we **Reject H_0** at 5% l.o.s

Thus, Hypertension and CVA given Locality are not independent.

LOGISTIC REGRESSION

```
glm(formula = CVA ~ ., family = "binomial", data = trainbal)
Deviance Residuals:
    Min       1Q   Median      3Q      Max
-2.6763  -0.7754  -0.2336   0.8025   2.7715
```

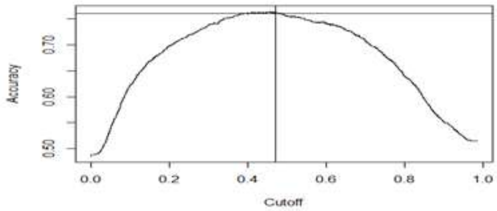
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.634e+01	2.735e+02	-0.060	0.952365
gender1	-1.639e-01	8.607e-02	-1.904	0.056852 .
age	6.468e-02	3.097e-03	20.884	< 2e-16 ***
hypertension1	7.263e-01	1.171e-01	6.201	5.60e-10 ***
heart_disease1	4.473e-01	1.403e-01	3.187	0.001435 **
ever_married1	1.009e-01	1.372e-01	0.735	0.462401
work_type1	1.258e+01	2.735e+02	0.046	0.963299
work_type2	1.130e+01	2.735e+02	0.041	0.967028
work_type3	1.121e+01	2.735e+02	0.041	0.967311
work_type4	1.130e+01	2.735e+02	0.041	0.967045
Residence_type1	4.960e-02	8.406e-02	0.590	0.555132
avg_glucose_level	5.095e-03	7.386e-04	6.898	5.27e-12 ***
bmi	8.047e-03	5.599e-03	1.437	0.150666
smoking_status1	3.699e-01	9.720e-02	3.806	0.000141 ***
smoking_status2	4.130e-01	1.150e-01	3.590	0.000330

```
***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
    Null deviance: 4875.5  on 3518  degrees of freedom
Residual deviance: 3500.4  on 3504  degrees of freedom
AIC: 3530.4
Number of Fisher Scoring iterations: 13
```

The most significant predictors of our model, according to our observations, are age, hypertension, heart disease, average blood sugar levels, and smoking status (0.05 or 5 percent).

The classification model was initially built to classify the probabilities above 0.5 as a positive (1) CVA case and that below 0.5 to be classified as Negative CVA. In order to verify if this classifying criterion was the optimum choice, we evaluate the best cut-off probability at which our model had maximum accuracy.



From the above accuracy vs cut-off probabilities plot we find that the 1923th value is maximum accuracy data point. And 0.7638534 is the maximum accuracy value whereas 0.4569022 is the cut-off value corresponding to the maximum value.

Here, we get to the conclusion that the cut-off value that ensures the greatest degree of accuracy is 0.4569, which, when rounded off, is equal to the probability value that we initially took into account for categorization. As a result, we draw the conclusion that the accuracy of 0.76 at the cut-off value of 0.46 is the best.

The model performance was then evaluated on the test dataset. The confusion matrix obtained is as follows,

	Predictions	
Actuals	FALSE	TRUE
0	1082	351
1	16	58

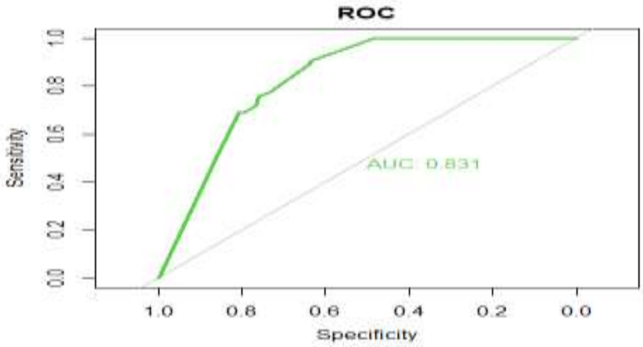
The correct classification rate is observed as **75.65%** whereas the misclassification rate is **24.35%.**

precision: 0.736

recall: 0.776

F: 0.378

Since here we are dealing with medical condition (CVA) dataset, we are more interested in correctly predicting the disease outcome events occurred (CVA=1). For this purpose, we use ROC curve.



We can see from the ROC curve shown above that the ROC curve is far above the intercept, which is a sign of a more effective classification model. Additionally, AUC shows that 83.10 percent of the area is covered by the curve.

Higher misclassification rates are seen in the initial decision tree, which may be the result of overfitting. The tree can be pruned to solve this issue. Decision trees can be pruned to make them smaller by deleting branches that lack the ability to classify cases. The chance of overfitting is highest for decision trees among all machine learning techniques; however, it can be decreased with careful pruning. Finding the ideal value for the pruning parameter, alpha, which regulates how much or how little pruning occurs, is the key to pruning a decision tree.

Age is chosen to be the root node. All data points with 49 percent CVA positive and 51 percent CVA negative cases are included in the root node. The fraction of negative and positive cases is indicated for each node, correspondingly. Individuals who meet the age criterion of 53 are divided, and following this initial division, those who meet the age criteria of 53 are moved on to the division of 42, where 18 percent of positive instances are seen. In contrast, the alternate branch from the root node that has an age greater than 53 is taken to another split with an age greater than 68, where 70% positive cases are seen. A threshold separates the observations at each node. CVA positive instances are indicated by the number 1 and CVA negative cases by the number 0. The category that is most prominently represented in each node is shown at the top. For instance, the root node is labelled 0 since CVA negative cases predominate there in terms of numbers.

With the exception of not having a variable and threshold for dividing the data, the leaves are identical to the nodes. A percentage is used to indicate the total number of observations in the node. For instance, the root node has 100% of the observations, whereas the observations are divided into 40% for ages 53 and 60% for ages >53. The class (0 or 1) that dominates the node gives the nodes and leaves their colours. In this instance, CVA positive (1) is various shades of blue, whereas CVA negative (0) is various shades of green. The amount of skewedness in a node or leaf is indicated by its gini score, which decreases with increasing shadow intensity.

Model performance evaluation:

The performance of the model is evaluated on the test dataset. The confusion matrix is as follows,

```
Confusion Matrix and Statistics
                Reference
Prediction       0          1
     0          1089       18
     1           344        56
```

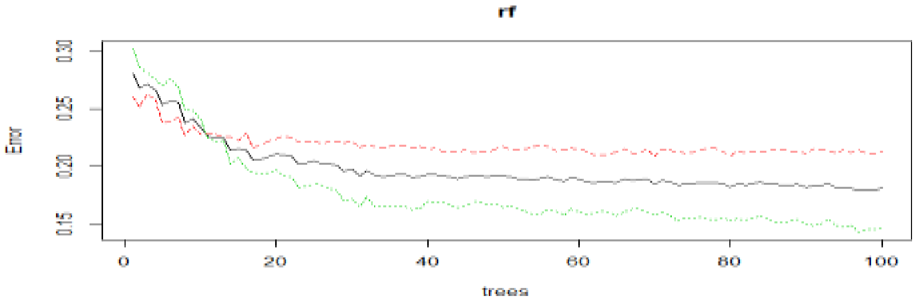
Accuracy: 0.7598, Precision :0.75978, Recall/ Sensitivity: 0.75676, Specificity: 0.75994

Pos Pred Value: 0.14000, Neg Pred Value: 0.98374, Prevalence: 0.04910, Detection Rate: 0.03716, Detection Prevalence: 0.26543, Balanced Accuracy: 0.75835.

RANDOM FOREST

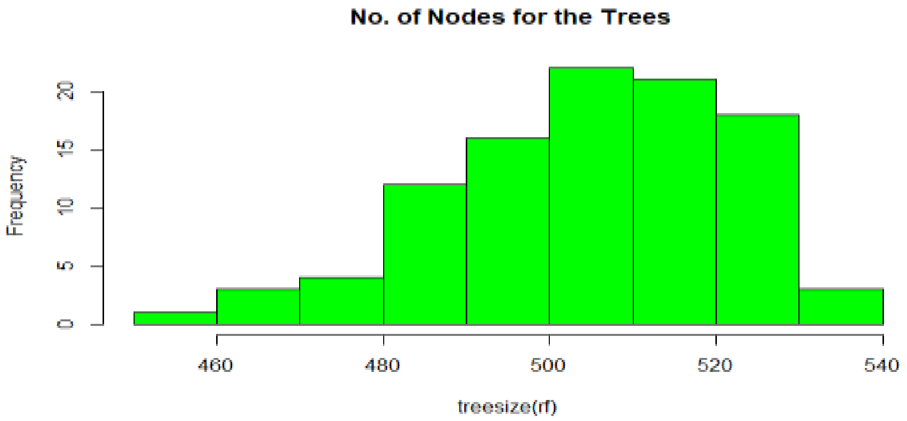
```
Call:
 randomForest(formula = CVA ~ ., data = trainbal, ntree = 100, mtry = 5,      importance = TRUE,
proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 5
        OOB estimate of  error rate: 18.08%
Confusion matrix:
     0        1        class.error
0 1434      387        0.2125206
1  254      1470       0.1473318
```

Error rate of Random Forest model:



As the no. of trees grow, we can see its out-of-bag error initially drops down and then it becomes almost constant. We are not able to improve this error after about 100 trees.

No of nodes for the trees



The above graph shows the number of nodes plotted against the frequency of trees with that number of nodes. Here, we observe that 20 trees have the 510 nodes which is the highest tree size.

```
           Reference
Prediction    0      1
         0  1076    26
         1   296    89
```

Accuracy : **0.7834**

95% CI : (0.7435, 0.7979)

Precision: 0.7834

Recall /Sensitivity: **0.7842**

Specificity: 0.7786

Pos Pred Value: 0.9764

Neg Pred Value: 0.1161

Prevalence: 0.9527

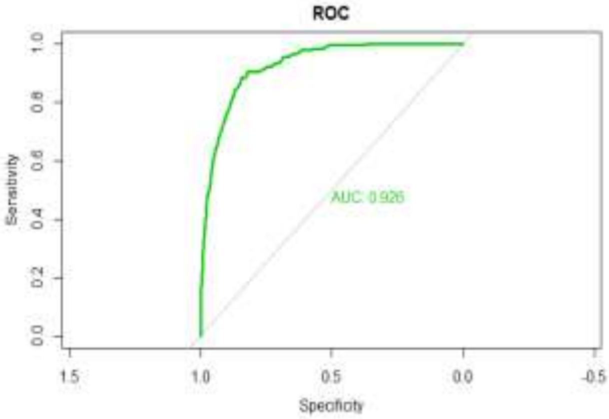
Detection Rate: 0.7265

Detection Prevalence: 0.7441

Balanced Accuracy: 0.7752

'Positive' Class: 0

Conclusion: Recall is high, so our model is good.

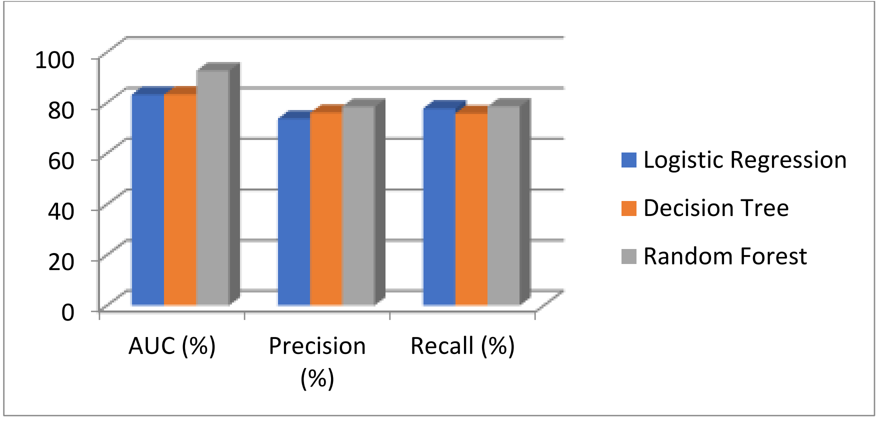


The above ROC curve is well above the intercept, which is an indicator of a better functioning random forest classification model. Also, AUC indicates 92.6 % of area is covered under the curve.

COMPARISON BETWEEN ALL THREE CLASSIFICATION MODELS:

Models	AUC (%)	Precision (%)	Recall (%)
Logistic Regression	83.1	73.6	77.6
Decision Tree	83.2	75.98	75.67
Random Forest	92.6	78.34	78.42

- The classification models in ascending order of their AUC values are as follows, Logistic regression< Decision tree < Random Forest
- The classification models in ascending order of their Precision values are as follows, Logistic regression< Decision tree < Random Forest
- The classification models in ascending order of their Precision values are as follows, Decision tree < Logistic regression< Random Forest



From above graph we can see that Random Forest has the highest AUC, Precision and Recall as compared to other two models.

CONCLUSIONS

- ❖ Gender, Heart disease and CVA are not completely independent.
- ❖ Gender and Heart disease are not jointly independent
- ❖ Heart disease and CVA given Gender are not independent.
- ❖ Locality, Hypertension and CVA are not completely independent.
- ❖ Locality and Hypertension are not jointly independent.
- ❖ Hypertension and CVA given Locality are not independent
- ❖ Random Forest has the highest AUC, Precision and Recall as compared to other two models.

References

1) healthcare dataset stroke data. [cited 2019; Available from: https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data

2) Witten, I.H., et al., Data Mining: Practical machine learning tools and techniques. 2016: Morgan Kaufmann. 27. Bierman, L., Random forests. Machine learning, 2001. 45(1): p. 5-32

3) Stroke Prediction using Distributed Machine Learning Based on Apache Spark

4) “Stroke Statistics,” The Internet Stroke Centre. [Online]. Available: http://www.strokecenter.org/patient s/about-stroke/strokestatistics/.

5) M. Prakash, G. Padmapriy et al.,”A Review on Machine Learning Big Data using R,” Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies, IEEE, 2018

6) An Effective Stroke Prediction System using Predictive Models International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

7) Yonglai Zhang, Wenai Song et al., “Risk Detection of Stroke Using a Feature Selection and Classification Method,” IEEE Access, vol. 6, pp. 31899-31907, 2018.