Contents.

Development of Stroke prediction machine learning model and Bias investigation using protected character (Gender)

Eze, Jonas Chinweike

Q2156726

World count: 2620

1. Introduction

According to Jiang & Nachum (2020), machine learning algorithms are now omnipresent, influencing various aspects of our lives, such as suggesting movies, products, and even potential partners. They are also increasingly used in critical areas like loan approvals and hiring processes. Mehrabi *et al*. (2021) stated that one major advantage of machine learning algorithms in decision making is their impartiality and ability to process vast amounts of data efficiently. Unlike humans, they don't suffer from fatigue or boredom and can consider a multitude of factors simultaneously. Despite their advantages, machine learning algorithms aren't immune to biases. Bias refers to systematic errors in decision-making that can lead to unfair outcomes. Just as humans can be biased, algorithms can also exhibit biases that result in decisions that favor or disfavor certain groups of people. (Alelyani, 2021).

Chakraborty *et al*. (2021) said that the two root sources of bias in machine learning models are :(1) Training Data Imbalance; which refers to situations where the distribution of different classes or categories in the dataset is uneven. For example, in a dataset of medical images, there might be far more images of healthy patients than images of patients with a particular condition and (2) Improper Training Data Labeling; which occurs when the labels assigned to the data are incorrect or inconsistent. For instance, in a dataset of animal images, some images might be labeled as "cat" when they actually depict other animals.

The major effect of bias of machine learning on humans is that protected characters such as race, sex, religion etc. are discriminated against. This discrimination leads to actions or decisions that are damaging or unjust for the people affected and it can happen in social media interactions, image recognition processes, recruitment processes and decisions made within the criminal justice system. (Fuchs, 2018).

2. Data Science Process

Problem identification ⇒ Raw dataset collection ⇒ clean/pre-processing of raw dataset ⇒ Exploratory data analysis ⇒ build model and Analyze result ⇒ presentation of result in a visual way.

3. Machine learning:

Machine learning is a sub field of Artificial Intelligence (AI) that focus on the development of an algorithm and a statistical model that learn from data without being explicitly programed. So, the algorithm is trained using data(features) and it makes prediction(decision) using the model(knowledge) gained during the training.

4. Dataset collection

The dataset is a secondary data that was gotten from Kaggle. It is a health care dataset and it consist of 5110 observations and 12 variables.

4.1 Dataset Attribute Information. (Nate,. 2022)

variables	Observations
Id	Unique identifier
Gender	Male, female or others
Age	age of the patient
Hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
Heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
ever_married	Yes or No
work_type	Children, Govt_job, Never_worked, Private or self_employed
Residence_type	Rural or Urban
Avg_glugose_level	Average glucose level in the blood
Bmi	Body mass index
Smoking_status	Formerly smoked, never smoked smokes or Unknown
stroke	1 if the patient had a stroke or 0 if not

5. Objectives

The main objective of this study and selecting this dataset is to study the dataset, analyze the dataset and use it develop effective and efficient stroke disease prediction machine learning model that can help hospitals and other health care institutions make informed decision.

The specific objective is to investigate bias in the model using protected variable: gender.

6. Features of the dataset.

The dataset is a supervised dataset. This is because it has dependent variable: stroke. It is also a classification task because the target variable (stroke) is category. Based on these features, the model will be developed using a classification machine learning algorithm.

7. Machine learning algorithm.

Machine learning algorithm is an algorithm that learns from historical data without being explicitly programed. There are so many classification machine learning algorithms. Support Vector Machine (SVM) is randomly selected for this study.

7. 1. Support Vector Machine (SVM).

This is a machine learning algorithm that makes prediction by maximizing or finding optimal margin between support vectors and hyperplane. It is used for both classification and regression problem. SVM formulates objective functions using the features of the data point with constraints: to maximize the margin between support vectors and hyperplane while minimizing classification error. This objective function is then converted to optimization problem and solved using quadratic programming problem to fine the hyperplane. SVR has the following hyperparameters:

- Kernel function:

This parameter enables the algorithm to find the optimal hyperplane that fits the data points. There are different types of kernel function and the choice depends on the relationship between the features and the target variable. It includes Linear Kernel for linear relationship and Radial Basis Function (RBF) Kernel for non-linear relationship.

- Class-weight:

This parameter handles class imbalance by assigning weights to each class inversely proportional to their frequency. It helps to ensure that the model does not bias towards the majority class and gives equal importance to minority class.

- Regularization (C):

This parameter is also known as penalty parameter controls the trade- off between maximizing the margin and minimizing the classification error. A small value of C typically leads to a larger margin and potentially more misclassification. A large number of C typically leads to a smaller margin and potentially fewer misclassification.

- Kernel coefficient:

This parameter determines the smoothness of the SVM decision boundary. There are different types of kernel coefficient. Its choice depends on the type of kernel function used. A small value of kernel coefficient leads to a smoother decision boundary, but if it is too small, it may lead to overfitting. A large

value of it leads to a complex decision boundary that closely fits the training data. This can lead to a model that has high variance and it is more prone to overfitting.

- Tolerance (tol):

This parameter is optimization stopping criteria. The optimization process stops when the difference between two consecutive iterations of the objective function is less than tol.

8. Dataset Importation.
This project was developed using python language and the development environment was Jupyter Notebook. The dataset was imported into the environment using pandas; a python library.

9. Data Cleaning.

Data cleaning is a process in machine learning modeling which focus on detecting and handling corrupt data such as wrong data type, duplicate rows, missing values and outliers to improve the quality of the dataset. First, the dataset was examined for wrong data type and it was observed that the age column was in float type. This is wrong for age should be a discrete value, but not continuous. As a result, it was rounded off and converted to integer.

	id	gender	age
9046		Male	67.0
51676		Female	61.0
31112		Male	80.0
60182		Female	49.0
1665		Female	79.0

Fig.1

	id	gender	age
9046		Male	67
51676		Female	61
31112		Male	80
60182		Female	49
1665		Female	79
...		...	...

fig.2

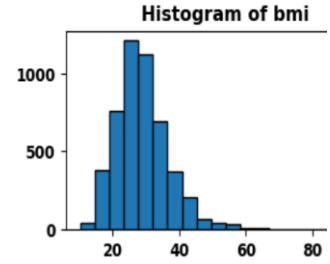
Figure 1 and 2 above shows the age column before rounding off and after rounding off and converted to integer.

Second, the dataset was examined for duplicate rows, but there were no duplicate rows.

Third, it was examined for missing values. It was discovered that bmi column has missing values. Bim column is numeric and continuous values. To handle the missing value, its distribution was examined using histogram. From the examination, it was observed that it is not normally distributed, so the missing values were replaced with the median value.

```
Missing Values:
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
```

Fig.3.



F.4

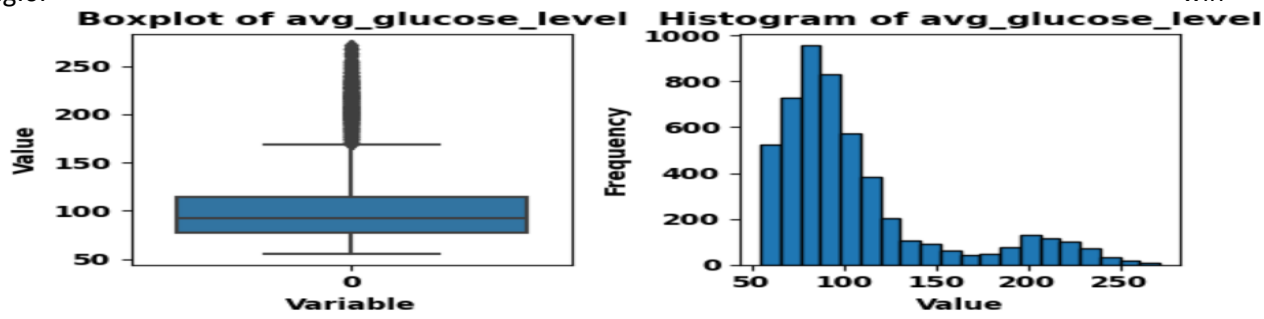
```
Missing Values:
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                  0
smoking_status       0
```

Fig.5

Figure 3, 4 and 5 above show bim column before the missing values were replaced, its distribution and after the missing values were replaced respectively.

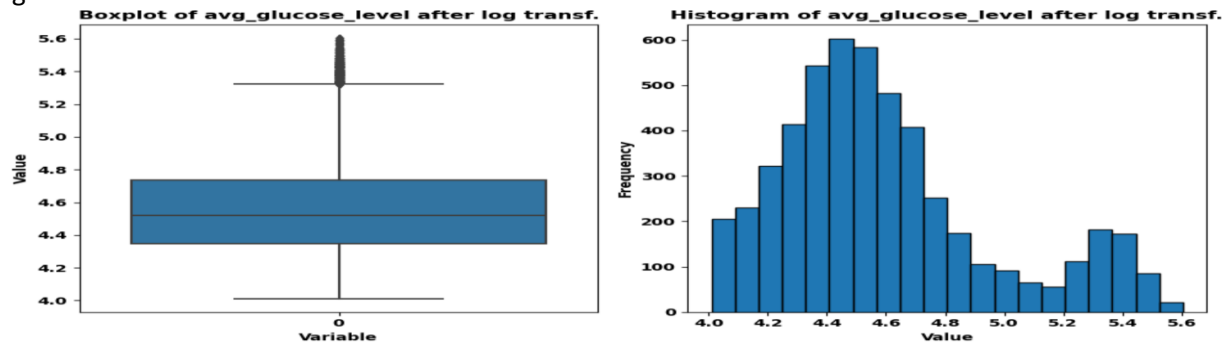
Again, the dataset was examined for outliers using box plot and histogram. It was discovered that there are outliers in avg_glucose_level and bmi columns. The outliers were handled using transformation and winsorization method.

Fig.6. win



From figure 6 above, the outliers in avg_glucose_level can be seen beyond the box plot whisker's length and from the histogram, it can be seen that it is not normally distributed. As a result, it was first transformed using log transformation method.

Fig.7.



Figgue 7 above shows avg_glucose_level after the application of log transformation.

After the transformation, the outliers were handled using winsorization method.

Fig.8.

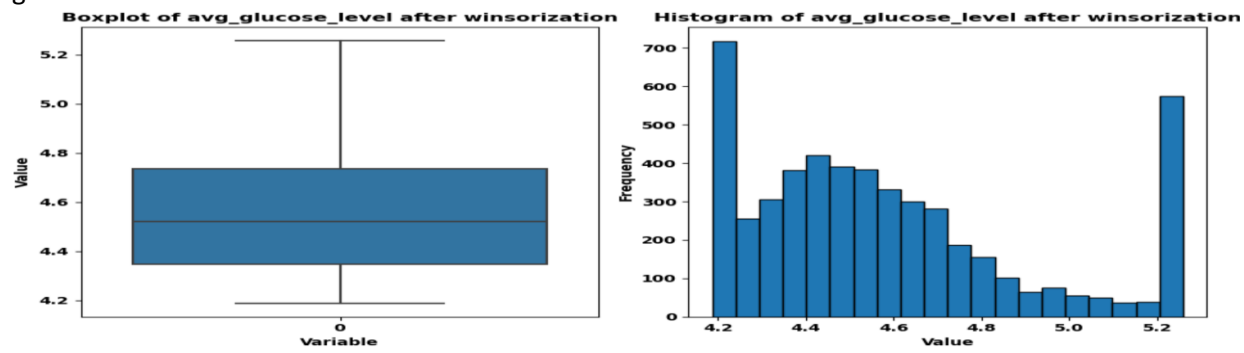
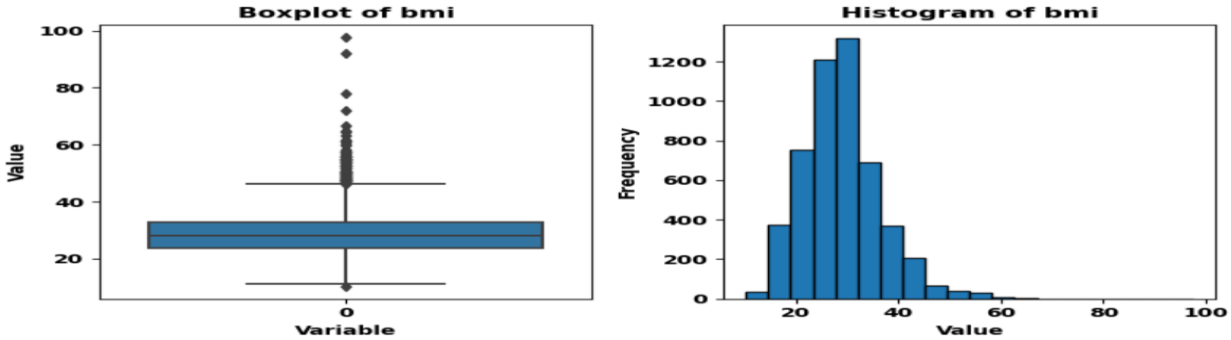


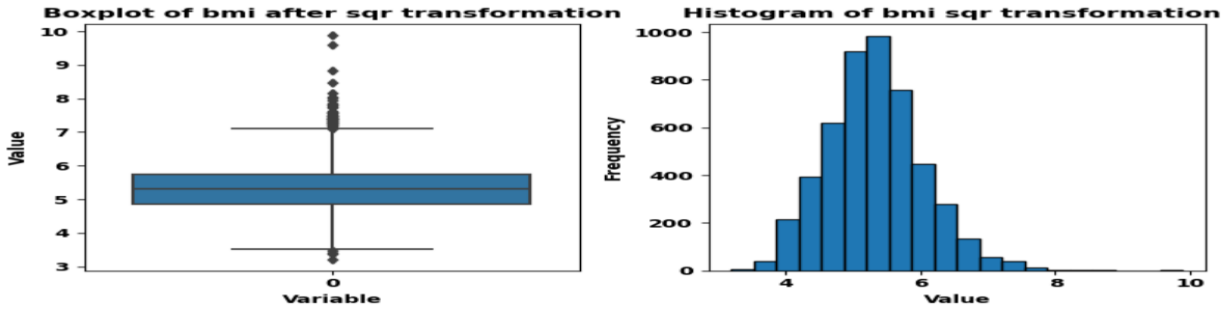
Figure 8 above shows avg_glucose_level after winsorization.

Fig.9.



From figure 9 above, the outliers in bmi column can be seen beyond the box plot whisker's length and from the histogram, it can be seen that it is not normally distributed. As a result, it was first transformed using square root transformation method.

Fig.10.



Figgue 10 above shows avg_glucose_level after the application of square root transformation. After the transformation, the outliers were handled using winsorization method.

Fig.11.

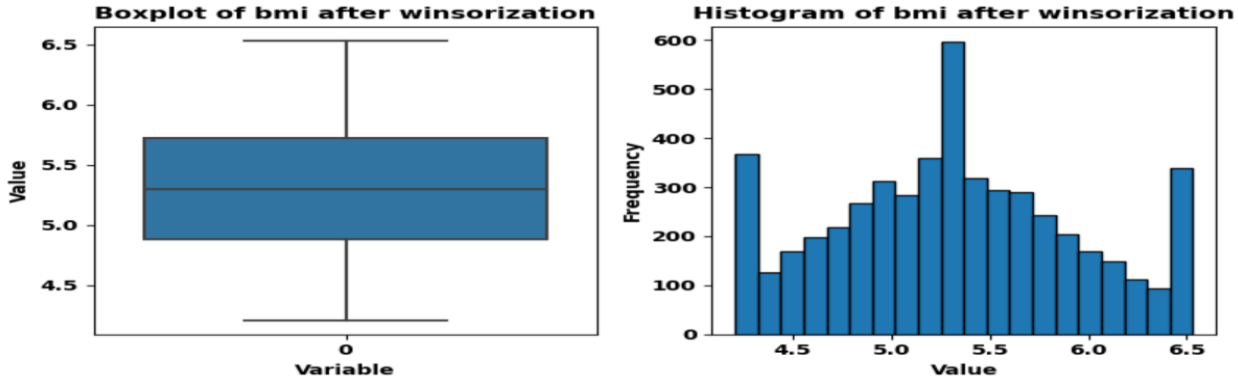
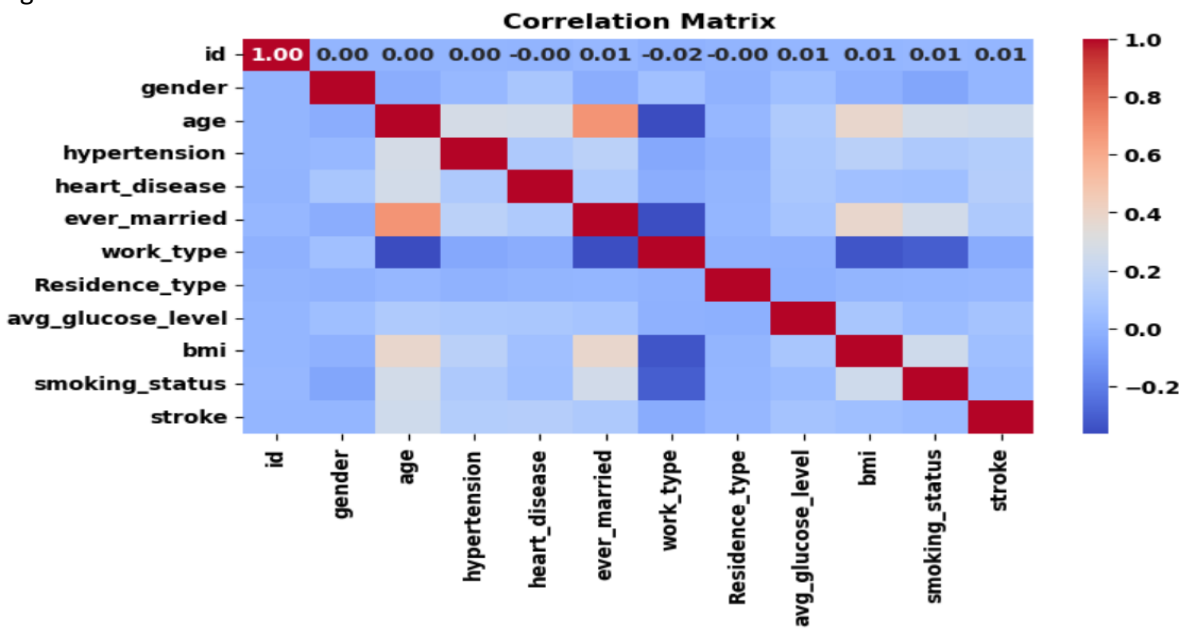


Figure 11 above shows bmi column after winsorization.

10. Exploratory data analysis (EDA).

EDA is a process in machine learning modeling that aims at visualizing the distributions and relationships among the variables. In this project, the distribution of the variables was visualized using histogram when looking for the presence of outliers while the linear relationship among the variables was visualized using correlation matrix. This relationship is very important in machine learning modeling because it determines which variable should be dropped or kept. If two variables have strong relationship; a relationship with correlation coefficient of -1 or +1 or very close to these values, one of the variables would be dropped. This is because training the model with the two variables will lead to collinearity. Under this condition, the model would not be able to differentiate their effect and this could lead to over fitting. This means that the model would be influenced by random fluctuation (occasional errors) and would not be able to generalize well on unseen data point. For the purpose of correlation plot, the categorical variables were label encoded, the correlation coefficient of the variable was calculated and the correlation matrix was plotted using heat map as shown below.

Fig.12.



From the correlation matrix above, it was observed that none of the variable is strongly related to one another. So, all of them were used for the modeling.

After examining the linear relationship among the variables from the correlation matrix, the dataset was reimported. This was done because the categorical variables that are not ordinal were label encoded in order to calculate the correlation coefficient. So, the reimported dataset was cleaned using the same

process, but the categorical variables were one hot encoded to form binary variable.

11. Feature engineering analysis.

Feature engineering analysis is a phase in machine learning modeling that focus on the transformation of the original dataset for effective learning. In this study the following feature engineering analysis were carried out.

11.1. One hot encode.

The categorical variables were one hot encoded to form binary columns.

Normalization.

Avg_glucose_level and bmi column were normalized using min-max normalization to have uniform or relatively uniform scale; a scale between 0 and 1.

11.2. Up sample.

The target variable classes were not balanced, so they were balanced by up sampling using smote method. figure 12 and 13 bellow show the target variable before and after up sampling respectively.

fig.12.

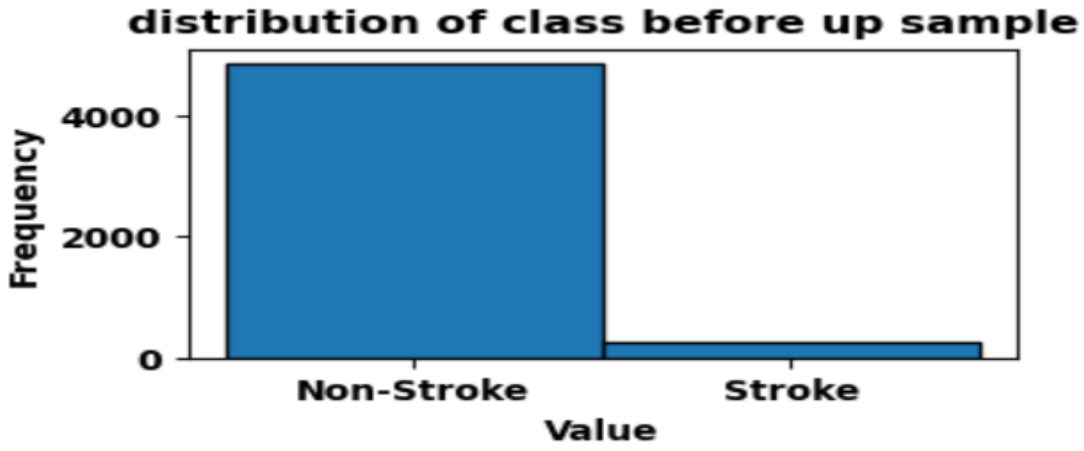
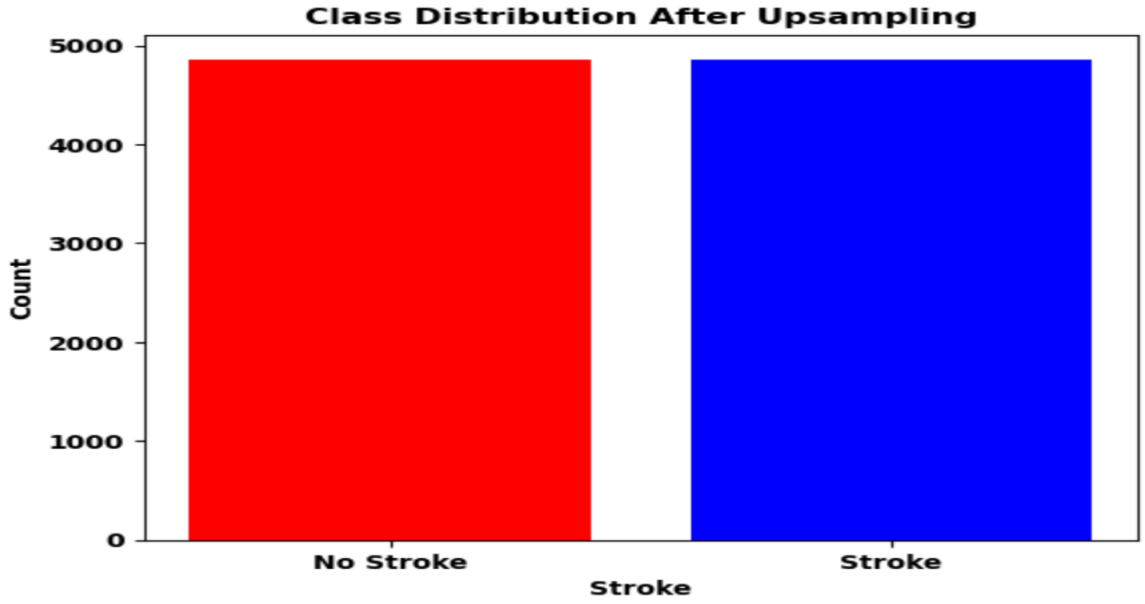


Fig.13.



12. Modeling:

12.1. Dataset split:

The dataset was split into three sets: 70% for training, 15% for validation and 15% for testing. A machine learning models were developed using Support Vector machine (SVM), algorithm. It was trained, validated, cross - validated using k- fold and tested. Also, its Receiver Operating Characteristic (ROC) curve and area under it was plotted.

12.2. Model Training:

In this phase, the algorithm was trained with 70% of the dataset to learn the relationship and underlying patterns among the variables and the data points.

12.3. Validation:

In this phase, 15% of the dataset and grid search method were used to tune the hyperparameters of the algorithm.

12.4. Cross - validation:

In this phase, the dataset was split into difference subset or folds using k-fold cross-validation method. This was done to examine how the model preforms across different subset.

12.5. Testing

In this phase, the 15% of the dataset was used to test the model for this determines how it will perform or generalize on unseen data points.

During the development, the model had three configurations and the hyperparameters configuration with best performance was printed. The developed model and its results are shown below.

13. Performance evaluation matrix;

Accuracy, Precision, Recall and F1- score. These are used to examine the overall performance of a model

- Accuracy is a model's performance evaluation matrix that measures the proportion of correct predications (TP and TN) among all the predictions (TP, FP, TN and FN) made by the model.

Mathematically, Accuracy = TP + TN/(TP+TN+ FP + FN)

Where TP = TRUE POSITIVE

TN = TRUE NEGATIVE

FP = FALSE POSITIVE

FN = FALSE NEGATIVE

- Precision is a model performance evaluation matrix that measures the proportion of trues positive (TP) predictions among all the positive predictions (TP and FP) made by the model

Mathematically, Precision = TP/(TP+FP)

- RecaLL is a model's performance evaluation matrix that measures the proportion of trues positive (TP) predictions among all actual positive prediction (TP and FN) made by the model.

Mathematically, Recall = TP/(TP+FN).

- F1-Score is the harmonic mean of precision and recall.

Mathematically, F1-Score =2x ((precision x recall)/(precision + recall)).

Fig.14.

```
Fitting 5 folds for each of 54 candidates, totalling 270 fits
Best parameters of SVM:
{'C': 10, 'class_weight': 'balanced', 'gamma': 1, 'kernel': 'rbf', 'tol': 0.001}

Performance on training set:
Confusion Matrix:
 [[3373   31]
 [  10 3391]]
Accuracy: 0.9939750183688464
Precision: 0.9909409701928696
Recall: 0.9970596883269627

Performance on validation set:
Confusion Matrix:
 [[679  29]
 [ 30 720]]
Accuracy: 0.9595336076817559
Precision: 0.9612817089452603
Recall: 0.96

Performance on test set:
Confusion Matrix:
 [[719  30]
 [ 36 674]]
Accuracy: 0.9547635366689513
Precision: 0.9573863636363636
Recall: 0.9492957746478873
```

Figure 14 above, shows the performance evaluation matrix of the model during training, validation and test.

14. Receiver Operating Characteristic (ROC)

In order to examine the model's performance more especially on its ability to discriminate between the classes (stroke and No stroke) of the target variable, I applied Receiver Operating Characteristic (ROC).

Fig.15.

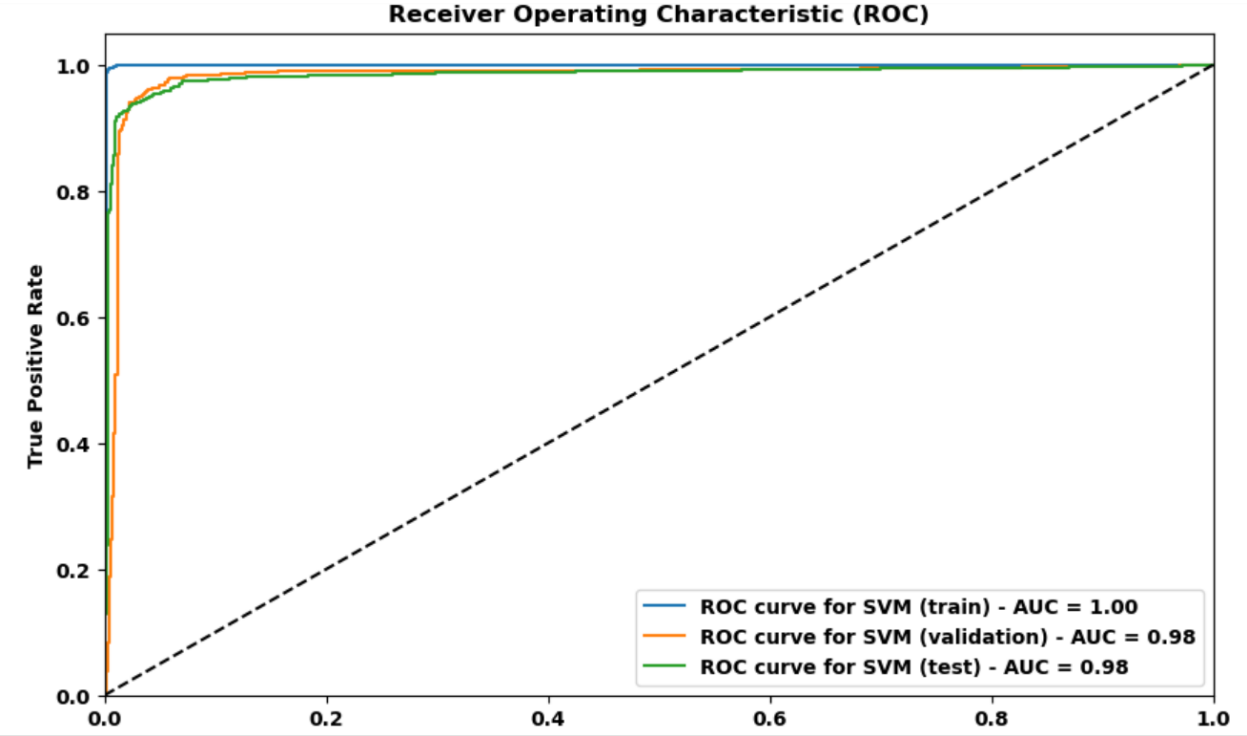


Figure 15 above shows Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) of the model. From the graph above, the AUC during training is 1; this implies that the model had a perfect discrimination during training. Also, during validation and testing it had equal AUC; 0.98. This implies that during validation and testing, out of all the discriminations made by the model, 98% were correct.

15. Bias investigation using gender.

In order to investigate bias in the model using gender, the test confusion matrix was split into men and women and their indices were generated and printed as shown in Figure16 below.

Fig.16.

```
Predicted values for Men: [0 1 0 0 0 0 1 1 1 1 0 0 1 0 0 1 0 0 0 1 1 1 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1
 1 0 0 1 1 1 0 0 0 0 1 0 1 1 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0
 1 0 0 0 1 1 1 0 0 1 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0
 1 1 1 0 0 1 0 0 0 1 1 1 0 0 1 1 1 1 1 1 0 0 0 1 1 1 0 0 0 0 0 0 0 1 0 0 1
 1 1 0 0 1 1 1 0 1 0 0 0 1 1 1 0 0 1 0 1 0 1 1 0 0 1 0 1 1 1 0 0 0 0 1 1 0
 0 1 0 1 0 0 0 1 0 1 0 0 0 1 1 0 1 0 1 0 0 0 1 0 0 0 0 1 0 1 1 0 1 1 0 1 0
 0 0 0 0 1 0 0 1 1 1 0 1 1 1 1 0 0 1 1 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 1
 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0 1 0 1 1 0 1 1 0 0 0 1 0 1 0 0 1 1 0 1 0 1 0
 1 1 1 0 1 0 1 1 0 0 0 1 0 1 1 1 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0 1 1 1 1 0 0
 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0 0 0 0 0 0 1 0 1
 0 1 1 0 1 0 1 1 0 0 1 1 1 0 0 0 0 1 0 1 1 0 1 1 1 0 0 1 1 0 0 0 1 1 1 0 0
 1 0 0 0 0 1 1 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 1
 1 1 0 0 0 0 0 1 0 1 1 1 0 1 0 0 1 1 1 0 1 1 1 1 1 1 1 1 0 0 0 0 0 1 0 1 1
 0 1 0 1 1 0 0 0 0 0 1 1 1 1 1 1 0 1 1 0 0 1 1 1 0 1 1 0 0 1 1 1 1 1 0 1 0
 1 0 0 0 1 0 0 0 0 1 1 1 1 0 1 1 1 0 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0 1 1 0 1
 0 0 1 1 1 0 0 0 1 0 1 1 1 0]
Predicted values for Women: [0 0 1 0 0 0 1 1 0 0 1 1 0 0 0 0 1 1 1 1 1 1 1 1 0 1 1 0 1 1 0 0 1 0 0 1 0
 1 0 1 0 1 1 0 1 0 0 0 0 1 0 0 1 1 1 0 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0 0 0 0
 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 1 1 1 1 0 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 1
 0 1 0 0 0 1 0 0 1 0 0 1 1 0 0 0 1 0 0 1 0 1 1 0 1 0 1 1 0 0 1 1 0 1 0 1 1
 0 1 0 0 0 1 1 1 0 0 0 0 0 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 0 0 1 1 1 0 0 0 0 0
 0 1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 1 1 1 0 0 1 1 0 0 0 0 1 1 0 0 1 1 1 1 0 0
```

Then the indices were used to generate confusion matrix for men and women as shown below.

```
Confusion Matrix for men:
[[310  13]
 [ 10 214]]

Classification Report for men:
              precision    recall  f1-score   support

           0       0.97      0.96      0.96       323
           1       0.94      0.96      0.95       224

    accuracy                           0.96       547
   macro avg       0.96      0.96      0.96       547
weighted avg       0.96      0.96      0.96       547
```

```
Confusion Matrix for women:
[[412  14]
 [ 31 332]]

Classification Report for women:
              precision    recall  f1-score   support

           0       0.93      0.97      0.95       426
           1       0.96      0.91      0.94       363

    accuracy                           0.94       789
   macro avg       0.94      0.94      0.94       789
weighted avg       0.94      0.94      0.94       789
```

16. Fairness Criteria

Since performance matrixes (accuracy, precision, recall and f1) are used to evaluate the overall performance of a model but cannot be used to determine if a model is biased in its predication, fairness criteria (Equal Opportunity, Demographic Parity and Equalized Odds) were applied to investigate whether the model is biased.

- Equal Opportunity measures the proportion of true positive (TP) prediction among actual positive predictions (TP and FN)
- Demographic Parity measures positive predictions (TP and FP) among all predictions (TP, FP, TN and FN).
- Equalized Odds measures the ratio of true positive rate (TPR) to false positive rate (FPR) across different demographic groups.

The model's fairness criteria:

Fairness Criteria for Men:
Equal Opportunity: 0.9597523219814241
Demographic Parity: 0.5850091407678245
Equalized Odds: 21.4984520123839

Fairness Criteria for Women:
Equal Opportunity: 0.9671361502347418
Demographic Parity: 0.5614702154626109
Equalized Odds: 11.324852339845524

17. Bias Report.

From the model's fairness criteria above,

- Equal Opportunity for men and women is 95.97% and 96.71% respectively.
- Demographic Parity for men and women is 58.50% and 56.14% respectively
- Equalized Odds for men and women are 21.4984 and 11. 3248

The model's fairness criteria indicate relatively similar performance between men and women in terms of Equal Opportunity and Demographic Parity. However, a significant disparity is observed in Equalized Odds, with men exhibiting a much higher value compared to women (21.50 vs. 11.32). This discrepancy suggests potential bias in the model's predictions, as it indicates unequal treatment or prediction accuracy across gender groups.

Given the significant difference in Equalized Odds between men and women, further investigation into the model's decision-making process and potential sources of bias is warranted. Addressing these biases is crucial to ensuring fairness and equity in the model's outcomes across different demographic groups.

References.

Alelyani, S. (2021). Detection and Evaluation of Machine Learning Bias. *Applied Sciences*, 11(14), 6271. https://doi.org/10.3390/app11146271

Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in machine learning software: why? how? what to do? In ESEC/FSE 2021: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (pp. 429–440). https://dl.acm.org/doi/pdf/10.1145/3468264.3468537

Fuchs, D. J. (2018). The dangers of human-like bias in machine-learning algorithms. *Missouri S&T's Peer to Peer, 2*(1). https://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=1030&context=peer2peer

Jiang, H., & Nachum, O. (2020). Identifying and Correcting Label Bias in Machine Learning. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, PMLR*, 108, 702-712.https://proceedings.mlr.press/v108/jiang20a/jiang20a.pdf

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 115, 1–35. https://dl.acm.org/doi/pdf/10.1145/3457607