Prediction of Stroke Based on Logistic Regression and K-Neighbour's Classifier

ECS784U/ECS784P – Coursework 1- Thaarik Ahamed Ahemed Ali -220207760

ABSTRACT

This report will present the prediction model of stroke based on various health stats of a individuals. The two machine learning models used are Logistic Regression and K-Neighbour's Classifier which are implemented using scikit-learn. The chosen dataset has four features of type int64, five features of type object, and three features of type object. The target feature has type of object int64 containing 0 and 1. The feature correlation's visualization is determined using the heatmap of seaborn visualization package. The report will give further details on steps taken on pre-processing the data, feature engineering techniques used with feature selection, normalization of data, modelling the data with the two machine learning algorithms, testing the prediction with test dataset and further feature engineering steps for better performance of the model. The literature review is also given for this respective machine learning field.

1. INTRODUCTION

The aim of this coursework is to diagnose the stroke attack on the given continuous data of all health stats and the lifestyle of every individual such as glucose level, Body Mass Index value, smoking habits, profession etc. A Stroke is a health condition that causes damage by tearing the blood vessels in the brain. It can also occur when there is a halt in the blood flow and other nutrients to the brain. According to the World Health Organization (WHO), stroke is the leading cause of death and disability globally. [1] Thirteen million individuals get strokes annually, according to the World Stroke Organization. However, approximately 5.5 million patients die as an outcome. Stroke is the leading cause of disability and death worldwide, making its imprint critical in all aspects of life. Stroke impacts the client's workplace, family, and social environment. [2] Stroke Prediction Using Machine Learning Classification Algorithm. Available from: The symptoms of stroke include paralysis or numbness on any parts of the body, headaches, trouble in speaking and understanding etc. There are some health conditions that cause strokes. Some are irregular glucose level, high cholesterol, irregular heartbeats, and high blood pressures.

In the current fast-paced society, lifestyle changes are one of the key factors contributing to stroke development. Migration of human population from rural to urban region leaded the way of development in societies. But these developments heavily impact on humans with stress, work life imbalance, activities that spoils human health such as consumption of alcohol, junk food, smoking etc.

This report analyze the datasets containing health statistics and lifestyle information of the individuals. The data is explored with various visualization tools that helps us understand the relation between each feature that contributes the stroke condition. The report also analyze the outcome of two machine learning techniques such as logistic regression and K-Neighbours classification that identifies the most relevant risk factors for stroke development. This provides a valuable insight to the healthcare professionals.

2. DATA PROCESSING

2.1. Dataset Information:

The dataset is collected from Kaggle source. Each row in the dataset contains relevant information of the patient. There are twelve features present in this dataset which is described briefly in below table:

S.No	Feature Name	Value	Type	Description
1.	id	Unique identifier	int64	Used as unique identifier for each patient.
2.	Gender (gender)	Male, Female, Other	object	Gender of the patient.
3.	Age (age)	0.08 - 82	float64	Age of the patient
4.	Hypertension (hypertension)	0,1	int64	0–Patient with no hypertension. 1 - Patient with hypertension
5.	Heart Disease (heart_disease)	0,1	int64	0–Patient without heart disease. 1 - Patient with heart disease

6.	Marital status (ever_married)	Yes, No	object	Yes – Married, No - Single
7.	Profession (work_type)	Children, Govt_job, Never_worked, Privte, Self-employed	object	Occupation of the patient.
8.	Residence (Residence_type)	Rural, Urban	object	Type of residence.
9.	Average Glucose Level (avg_glucose_level)	55.12 – 271.74	float64	Average of blood glucose level.
10.	Body Mass Index (bmi)	10.3 – 97.6	float64	Body mass index of the patient.
11.	Smoking status (smoking_status)	Formerly smoked, never smoked, smokes, Unknown	object	Smoking status of the patient. Unknown – data unavailable.
12.	Stroke (stroke)	0,1	int64	0 – Patient didn't have a stroke. 1 – Patient had stroke.

Table 1 - Dataset Attribute Information

The dataset is first read through panda library. The shape of the dataset is (5110,12) i.e., 5110 rows (patient record) and 12 columns(features).

The feature 'id' contains unique identifier of all patients. This may give rise to inconsistent performance. So, this column is dropped. Among all the features, 'bmi' feature contains 201 null values. The null values are replaced with the mean value of all bmi value.

2.2. Data Visualization:

The dataset is then divided into parts on the basis of numerical value and categorical value. The numerical valued feature consists of age, average glucose level and bmi of every patient. With the help of scatterplot in seaborn package, relationship between these columns and the stroke occurrence is noted. It is found that with the increase in age, the chances of getting stroke are higher. There is not much relation between average glucose level value, bmi value and stroke chances. The probability of getting stroke occurs in all range of bmi and glucose level.

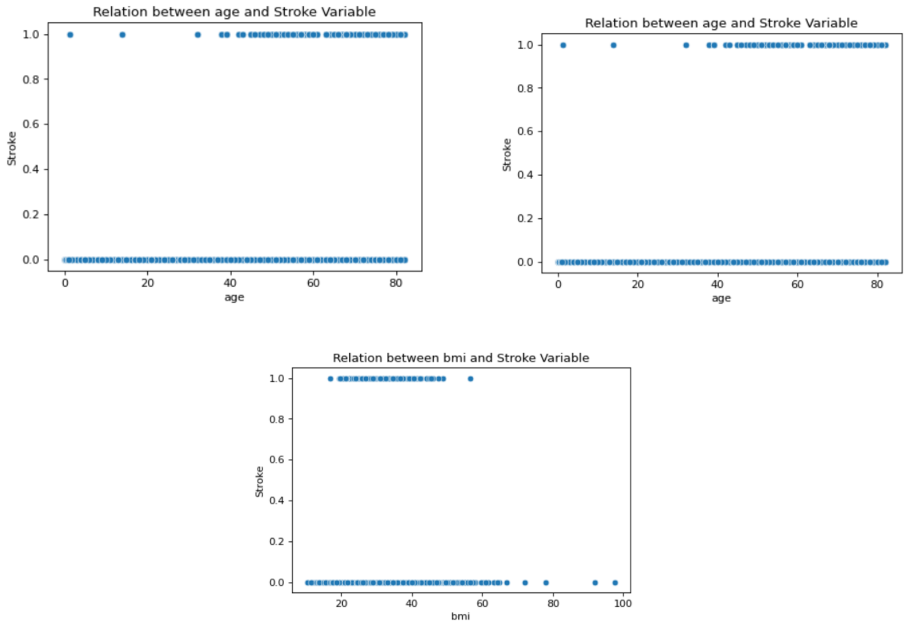


Figure 1- Relation between all numerical variable and target variable

The correlation matrix is generated to find the correlation between all numerical features. It is studied that there is no much higher influence of correlation between all numerical features.

The categorical features in the dataset (second divided part) contains features of gender, hypertension, heart_disease, ever_married, work_type, Residence_type and smoking_status. The gender column has 2994 counts of female patient, 2115 counts of male patient and 1 “other” gender patient. The gender “other” row does not provide significant impact to the model, so this row was dropped. With the help of crosstab function in pandas package, the relationship between all categorical feature and the stroke occurrence is studied. Stroke occurs to all gender. Patients with high hypertension, heart diseases victims and with smoking habits (past and present) have higher chance of Stroke. Surprisingly, patients who are married have higher chances of stroke. Residence type does not impact much in finding the probability of stroke. Self-Employed professionals have higher risk of stroke after the private profession and Government job.

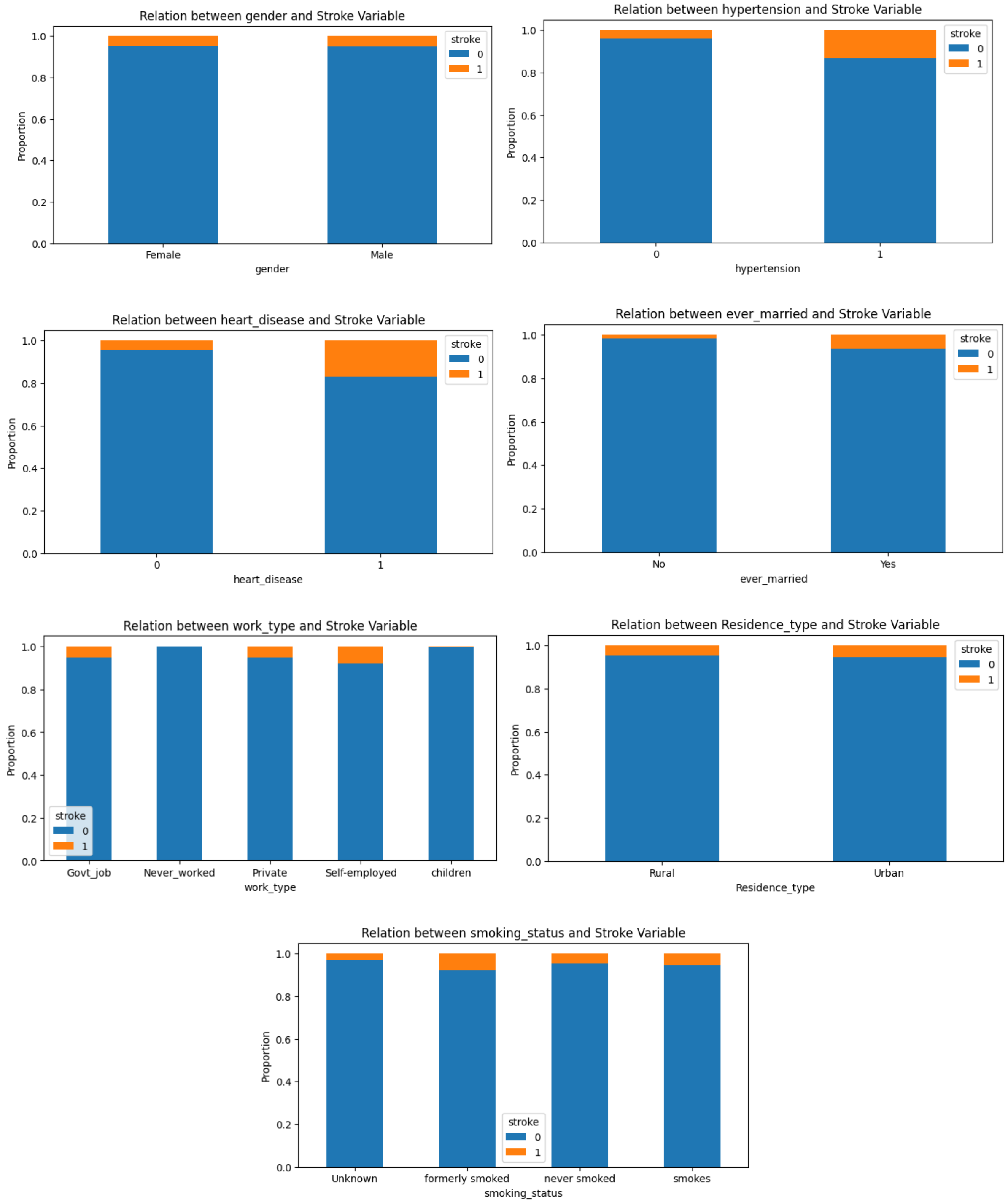


Figure 2 - Relation between all categorical variable and target variable

In the dataset, the count of patient data having stroke (249) is much lesser than the count of patient data with no history of stroke (4860). This shows that the dataset is heavily imbalanced. To handle this, necessary actions must be taken before using the dataset for prediction. Before structuring the machine learning model, there are further pre-processing steps required.

2.3. Normalization of Data

Some data in the numerical dataset are skewed and contains outliers. To solve this issue, normalization must be done before feeding it into the prediction model. ‘avg_glucose_level’ and ‘bmi’ columns require normalization. 'np.log()’ function from numpy library is used on all the dataset values. This makes the data more normalized in the positive range.

2.4. Label Encoding

The categorical columns contains object type which must be changed into numerical value of int64 type. This process is called Label Encoding. This is done on all categorical columns using ‘LabelEncoder()‘ from scikit-learn library.

3. FEATURE SELECTION

3.1 Correlation check between all features

Correlation helps in better understanding of relationship different features in a dataset. This is denoted through correlation coefficients between the pairs of features in the form of matrix. After looking at the magnitude of the coefficients, there are few higher near to 1 positive correlations between the pair of avg_glucose_level – avg_glucose_level_log and bmi-bmi_log. These two pairs contain the original value and the normalized value. So the original valued features are dropped. There are some negative correlation coefficients in some pairs of variables but they are negligible.

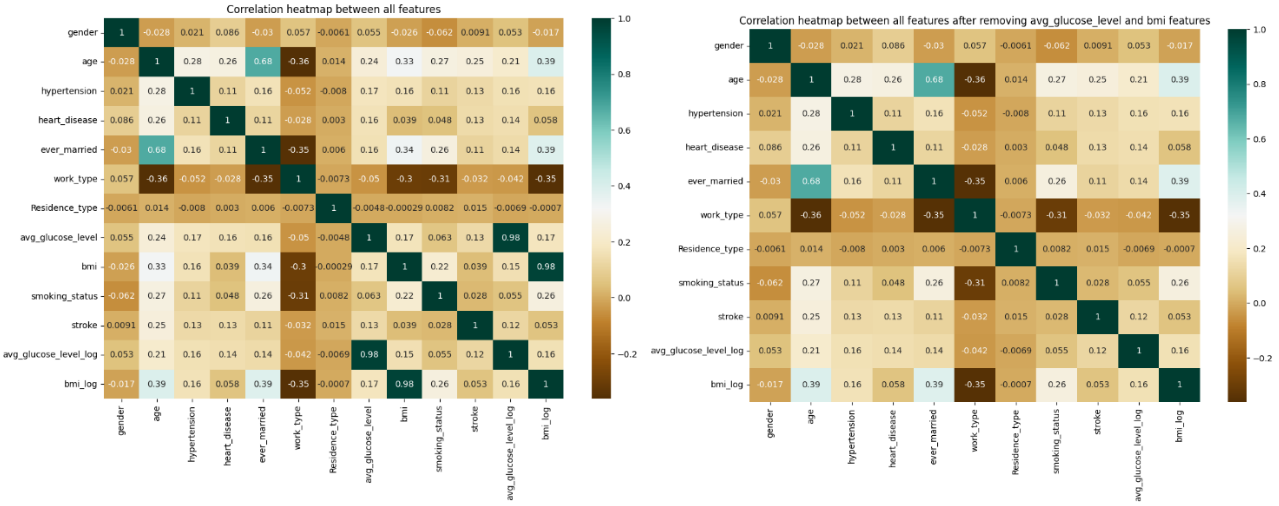


Figure 3 - Correlation Heatmaps - Before and after removal of highly correlated features

After this step, the pre-processed dataset is divided into feature variables and target variable.

3.2 Selecting KBest Feature

To select best K variable features, chi-squared test and f test are used. Chi-squared test is used to find the frequency of each variables in association with target variable. The ftest is used to compare the results of mean value of all the features with target variable. These two tests are used on finding out how much the features creates an impact onto the target variable. The results of both tests show that except age, there was less common in between the features to the target variable. So, there was no dropping of features.

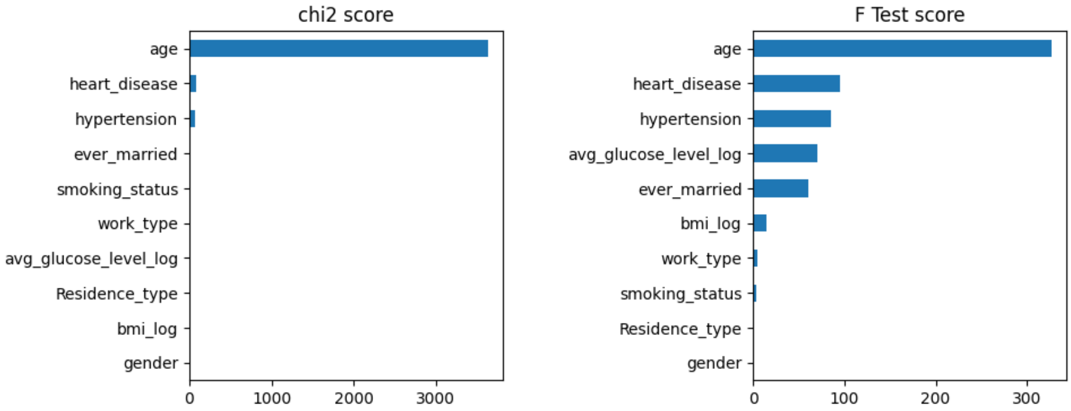


Figure 4 - Selecting KBest feature using chi2 and FTest score

4. LEARNING METHODS

The aim of this project is to predict the stroke attack and the dataset used contains labelled features. So, supervised learning method is applied here. This method helps in training the labelled data and finds the dependence between every feature and its predicted output in the form label. When the model is given an input features, it should predict the correct output labels. Two supervised machine learning approaches are applied in this project. They are Logistic regression and k-Nearest Neighbours (KNN).

4.1. Logistic Regression

Logistic Regression (also called *Logit Regression*) is commonly used to estimate the probability that an instance belongs to a particular class (e.g., what is the probability that this patient has stroke?). If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class (called the *positive class*, labeled "1"), and otherwise it predicts that it does not (i.e., it belongs to the *negative class*, labeled "0"). This makes it a binary classifier. [3]

$$\widehat{p} = h_{\boldsymbol{\theta}}(\mathbf{x}) = \sigma(\boldsymbol{\theta}^{\intercal}\mathbf{x})$$

Equation 1 - Logistic function [3]

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

Equation 2- Sigmoid function [3]

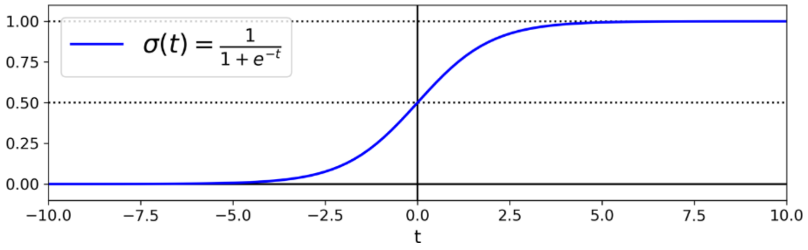


Figure 5 - Logistic Function [3]

This model is estimated by applying the sigmoid function on the features and generates the value in the range of 0 and 1.

The reason of choosing this model is as follows: The model has simple algorithm and yet performs very well with large datasets, since this model is used for binary classification, it is the best approach of classifying our problem to less risk and more prone to risk condition for the given patient data, and it can handle more noisy data. This model can also be used for easy interpretability of identifying overfitting issue and handling it with various regularization methods.

4.2. K-Nearest Neighbours

The K-Nearest Neighbors classifier (KNN) is one of the simplest yet most commonly used classifiers in supervised machine learning. KNN is often considered a lazy learner; it doesn't technically train a model to make predictions. Instead, an observation is predicted to be the class of that of the largest proportion of the *k* nearest observations. For example, if an observation with an unknown class is surrounded by an observation of class 1, then the observation is classified as class 1 [4]. KNN is considered as instance-based classification as it classifies the new data by comparing the instances of previous historical data that closely matches. If we have training dataset of 4 classes C1, C2, C3 and C4 and if we give K=6, then the goal is to find the class that closely matches with majority to the new data X. [5] The majority is decided by the distance function between the new data and the previous data that are labelled with classes. The Figure below explains the functioning of KNN:

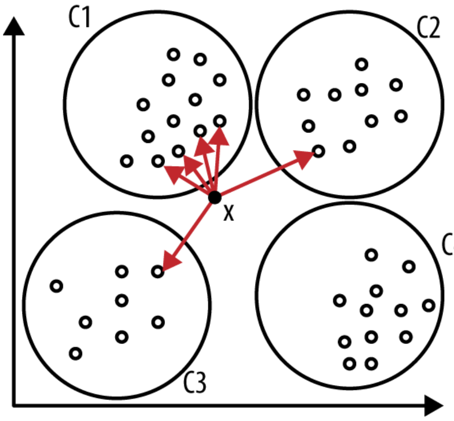


Figure 6- kNN classification with C1, C2, C3, and C4 [5]

The reason of choosing the kNN classification is as follows: The model can handle both distribution of data, continuous and categorical value, without any assumptions, it is useful in this project as the dataset used is comparatively small and easy to handle, it can handle noisy data and finally it can handle binary and multiclass classification if there was any future update on the classification of stroke disease prediction.

5. ANALYSIS AND TESTING OF THE RESULT

The dataset used in this project is heavily imbalanced. This kind of dataset are usually expected in the medical field. Stroke is less common than other diseases. So, to handle this imbalance dataset, SMOTE technique is used. SMOTE stands for Synthetic Minority Over-sampling Technique. It creates synthetic examples of the minority class (1 value in our target variable) by selecting and interpolating similar minority class samples randomly. This balances the dataset by increases the size of the minority class distribution. SMOTE is applied using imblearn package which is derived from scikit-learn library.

5.1. Analysis of the result

Logistic regression and kNN model are applied by using the scikit-learn library. The data is splitted into training data (80%) and testing data(20%). The training data is passed into these two models and the following results are obtained:

Metrics	Logistic Regression	kNN
Accuracy on training data	78.0%	92.2%
Accuracy on test data	77.8%	90%
Precision	0.783	0.847
Recall	0.762	0.975
F1-score	0.806	0.906

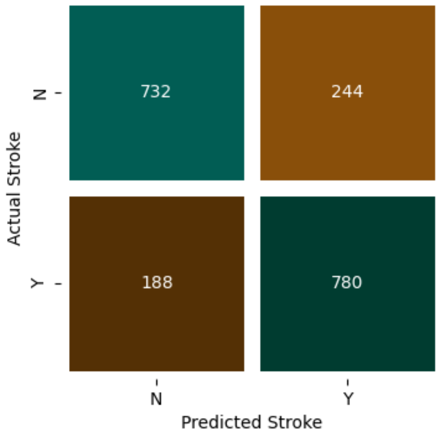
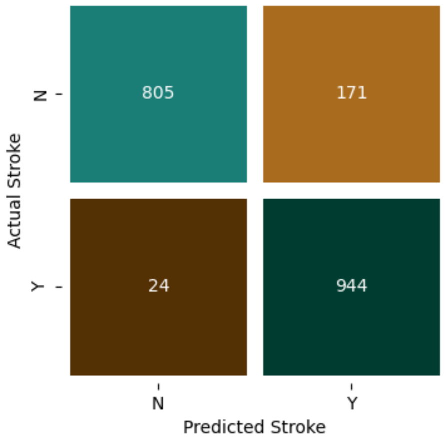
Confusion Matrix		

Table 2 - Analysis and Test Table

5.2. Validation of the result:

The trained models are evaluated with Kfold cross validation with K=10 using test dataset. The kNN model better in both training and test dataset with higher values of accuracy, f1 score, precision and recall. But comparatively, kNN model has lower accuracy difference on test dataset (90%) with logistic regression dataset (77.8%). kNN model can become computationally expensive if large datasets are loaded. So, an analysis on improving the Logistic regression model is done using two methods. With the help of Grid search technique (GridSearchCV from sklearn library), the best hyperparameter for regularizing the Logistic regression model is calculated (C=10), that improved the accuracy value of 79.3%. Principal Component Analysis (PCA) is also done on Logistic regression model, and its found that the model already achieved best performance with the applied number of features. SO now feature reduction is required. Even after these optimizations, kNN model still outperforms Logistic regression model.

6. CONCLUSION

The aim of the project has been achieved by following the analysis of the dataset nature, studying the two best supervised machine learning models and it is concluded that the kNN classification performs well than the logistic regression model.

The limitation of this project is the imbalance dataset. Even though, this issue is solved with resampling technique, it is better to add a greater number of datasets which are required to develop the validity of the dataset and model. The given features are not enough to determine the prediction of stroke. There must be a greater number of underlying medical factors which must be consulted from a medical professional so that we can improve an appropriate dimension in the dataset. With good number of datasets, we can use advance supervised machine learning algorithms such as XGBoost and Neural Networks for better optimized results. The curse of dimensionality and overfitting issue must be foreseen throughout the development.

REFERENCES

[1] M. Wiryaseputra, "Stroke Prediction Using Machine Learning Classification Algorithm," in *International Journal of Scientific & Engineering Research* , 2017.

[2] A. I. M. A. L. I. Mohammed Guhdar, "Optimizing Accuracy of Stroke Prediction Using Logistic Regression," in *Journal of Technology and Informatics (JoTI)*, 2023.

[3] A. Géron, in *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*, O'Reilly Media, Inc..

[4] C. Albon, in *Machine Learning with Python Cookbook*, O'Reilly Media, Inc..

[5] M. Parsian, Data Algorithms, O'Reilly Media, Inc..