Teesside University

MACHINE LEARNING ICA REPORT

STROKE PREDICTION USING MACHINE LEARNING

NAME: OLADEINBO OLUTAYO

STUDENT ID: B1232321

EMAIL: B1232321@live.tees.ac.uk

ABSTRACT

Machine learning is a sub-category of artificial intelligence and effectively automates the process of analytical model building and allows machines to adapt to new scenarios independently(IBM,2021). ML is divided into 2 groups, Supervised and Un-Supervised learning. Supervised learning is where the dataset has a target feature that is to be predicted, this target is called a dependent variable. Unsupervised learning doesn't have a target feature.

In this report, the ML's supervised learning approach is used since the dataset has a target column, it was used to understand the relationships that exist between several columns in the dataset with the target column. To achieve this, Python's (PANDAS and 6 Machine Learning Algorithms) were used to explore the dataset consisting of 5110 rows and 12 columns of different patient parameters.

Using Feature Engineering, several interesting features were selected from the dataset and used in the training of our ML algorithms with high accuracy and low F1 Score. The Model can then be used to predict future cases of diabetes.

INTRODUCTION AND LITERATURE REVIEW

Since the rising of Machine Learning, Deep Learning, and AI, Python programming language has been the most popular language that developers use in this field. This is because it has a lot of libraries that handle data manipulation and also Machine Learning and Deep learning such as Pandas for Data Manipulation, Sklearn for Machine Learning, Keras and Tensor Flow for Deep Learning, Matplotlib, and Seaborn for Data Visualization. The environment can be set up in a script, and executed in an IDE or interactive notebooks such as Jupyter Notebook which was used to perform and present this ICA.

A Stroke is a life-threatening unforeseen medical crisis that occurs when a section of the brain cells is blocked from receiving the normal level of blood supply that brings in oxygen and glucose to the brain cells, thereby causing the cells affected to die(NHS,2021). According to World Health Organisation (WHO) in the year 2019, stroke ranks second on the chart of global causes of death and third in the global causes of disability. It has been noticed over the years that tracking symptoms such as Body Mass Index (BMI), Heart Disease Status, Hypertension status, Average Glucose Level, Age, Smoking Status, Work Type, and Marriage History can help detect stroke patterns early and it can be treated.

Stroke can be avoided by making relevant changes to one's lifestyle and having regular checks on key factors can reduce one's exposure to stroke. Having a model that can help check these factors regularly and alert the individual when to adjust his or her lifestyle can help with prevention and early treatment. Several studies have been carried out to uncover efficient stroke predictors. Studies have been done to discover the main stroke predictors. Because of this research done, Machine Learning algorithms are becoming widely accepted in the medical field to help deal with the early detection of stroke in humans.

The Data was collected from Kaggle's Stroke Prediction Dataset link and consists of 12 features and 5110 rows. The features are:

1. ID
2. Age
3. Average glucose level
4. Body mass index
5. Gender

6. Hypertension
7. Heart Disease
8. Ever Married
9. Work Type
10. Residence Type
11. Smoking Status.
12. Stroke (Target)

Feature Engineering was used to remove some columns and also add new columns, which will be discussed later.

The goal of this ICA is to use the data we have gathered about patients and visualize the relationship between different features, and develop different ML and Deep Learning Models to predict the likelihood of a stroke occurring in patients. The following are the research questions and assumptions:

- Male are more susceptible to stroke than females.
- Does Age have an impact on Stroke? And how the 'Age' parameter is distributed.
- Does BMI level in a person propel stroke?
- Does Glucose level in a person propel stroke?
- Smoking can induce Stroke, is this true?
- A patient with hypertension is prone to stroke.
- Workload resulting from work type could lead to stroke.
- Effect of Marriage and Unmarried people on Stroke.

There have been numerous research works done on stroke. JoonNyung Heo (J.et al, 2019) used 3 ML algorithms (ANN, Random Forest, and Logistic Regression) were used to predict stroke in victims. The models gave different accuracies and were compared to get the best algorithm. Shakiry Alaka (Alaka S. et al, 2020) also worked on the same project using 4 ML algorithms to predict the outcome in stroke victims after endovascular treatment.

EXPLORATORY DATA ANALYSIS AND FEATURE ENGINEERING

EDA is the process of analyzing a dataset to summarize its main characteristics. The best way to go about it is by using visualization. In this ICA, the libraries used for Visualization are Matplotlib, Seaborn, Plotly, and Pywaffle and Data Structures.

The Data contains 12 Features (7 categorical features and 3 continuous Features). The categorical features are Gender (Male, Female, Others), Hypertension (1 for yes and 2 for No), Heart Disease (1 for yes and 2 for No), Ever Married (1 for yes and 2 for No), Work Type (Never Worked, Private, Government Job, Children, Self-Employed).

The dataset has 201 missing data from the BMI and since that accounts for less than 5% of the dataset, the missing columns were dropped.

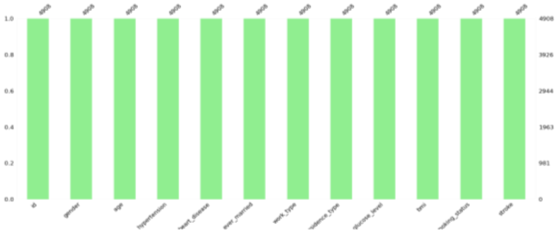


Fig 1: Cleaned Dataset

DEEPER DIVE INTO THE COLUMNS

The Gender column was categorized into 3, Female (2838), Male (1991) and Others (1). The others category was dropped because it only has 1 entry and doesn't have enough information for our model to consider that.

Looking at the distribution in this column, it was observed that the percentage of Male (96% Healthy, 4% Stroke) and females (96%

Healthy, 4% Stroke) are the same percentages. Therefore, the assumption that Male are more susceptible to stroke than females is wrong.

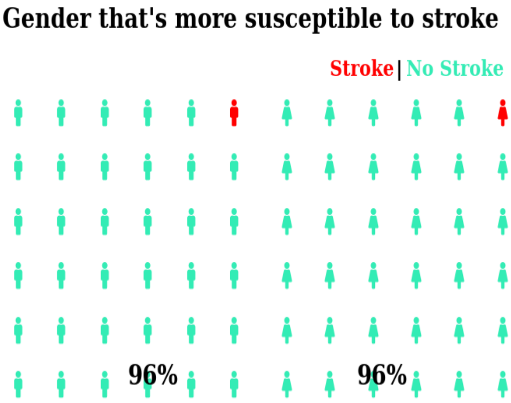


Fig 2: Gender Distribution of Stroke and Healthy Patients

The Age column records the age of every instance in the dataset. The data collected was analyzed and it was observed that older patients are more susceptible to stroke than younger patients, which answers our 2nd research question. The ages were further broken down into different categories to show how the number of Stroke cases increased as the age group increased.

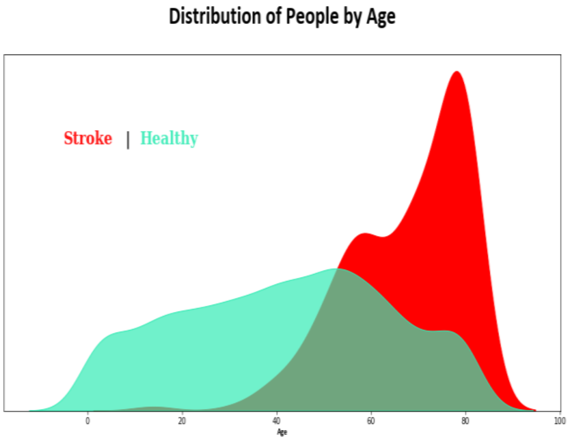


Fig 3.1: Age Distribution of Patients with Stroke

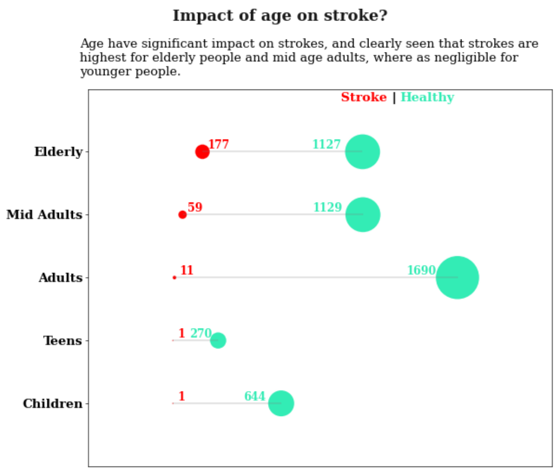


Fig 3.2: Impact of Age on Stroke

Hypertension

The Hypertension Column is a categorical column that records if a patient has hypertension or not. It was also observed that there is a high possibility of people who have hypertension having a stroke.

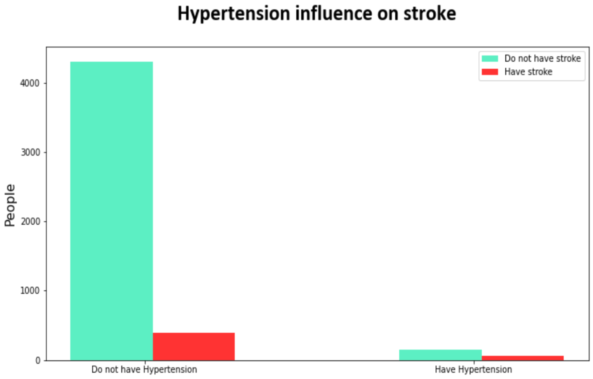


Fig 4: Hypertension influence on stroke.

Marriage

In the Ever_Married column, it is a categorical column that shows the marriage status of the patient. It is observed that being married influences stroke.

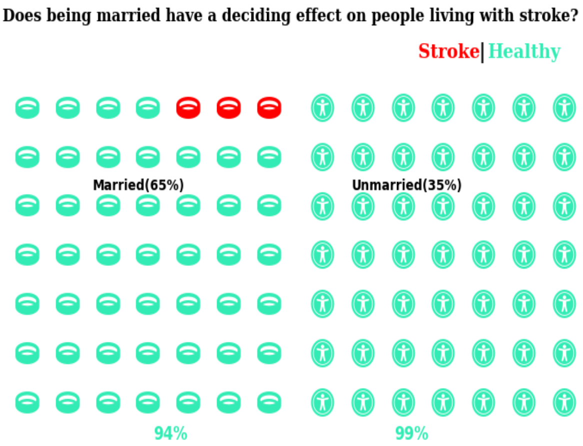


Fig 5: Distribution of Married and Unmarried people with Stroke

Work Type

The WorkType column is divided into 5 categories that a patient can be in. Further analysis of this column with the stroke column shows that patients who have never worked are least susceptible to stroke while self-employed people are more susceptible to stroke. This may be because of the high volume of work and responsibilities of a self-employed individual.

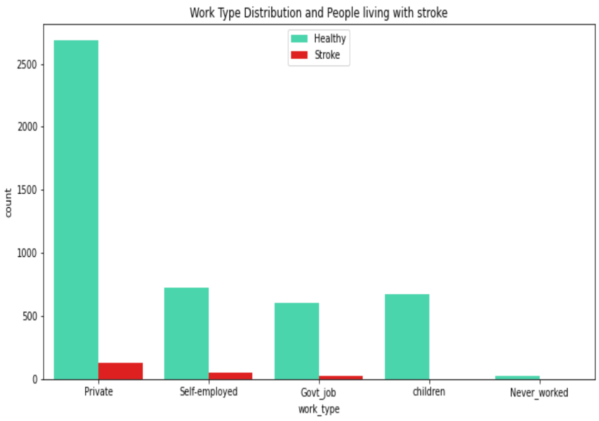


Fig 6: Distribution of Work type with Stroke and Healthy Patients

Smoking

The Smoking Status column has 4 categories that a patient can fall into. It was further analyzed and the distribution shows that Smoking does not have a significant impact on Stroke.

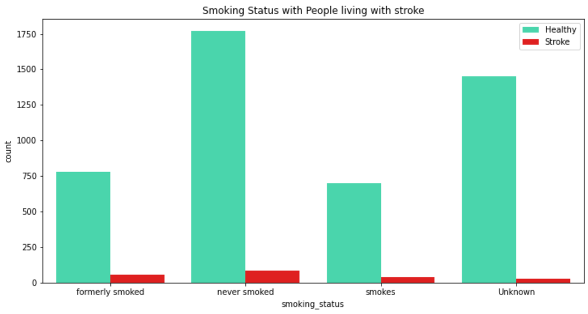


Fig 7: Distribution of Smoking Status on Stroke

Body Mass Index

BMI refers to the Body Mass Index of a person, the BMI column shows the data distribution of people’s BMI, and the distribution shows that people with higher BMI are more susceptible to stroke.

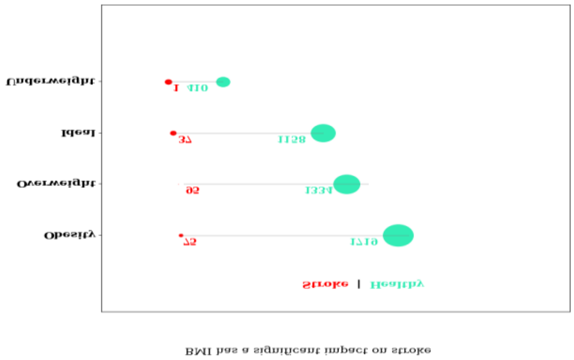


Fig 8: Distribution of BMI levels in patients with stroke.

Glucose Level

The Glucose Level column shows the level of Glucose in a person’s body. It was divided into different ranges from low to high. It shows that glucose Level in a person determines if a person will have a stroke or not.

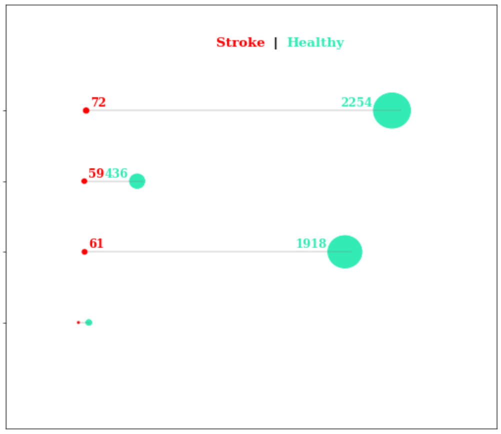


Fig 9: Distribution of Glucose Levels with stroke

FEATURE ENGINEERING

This is the process of extracting important features from raw data and transforming them into formats that are suitable for our model. There are a lot of advantages Feature Engineering gives to our model such as

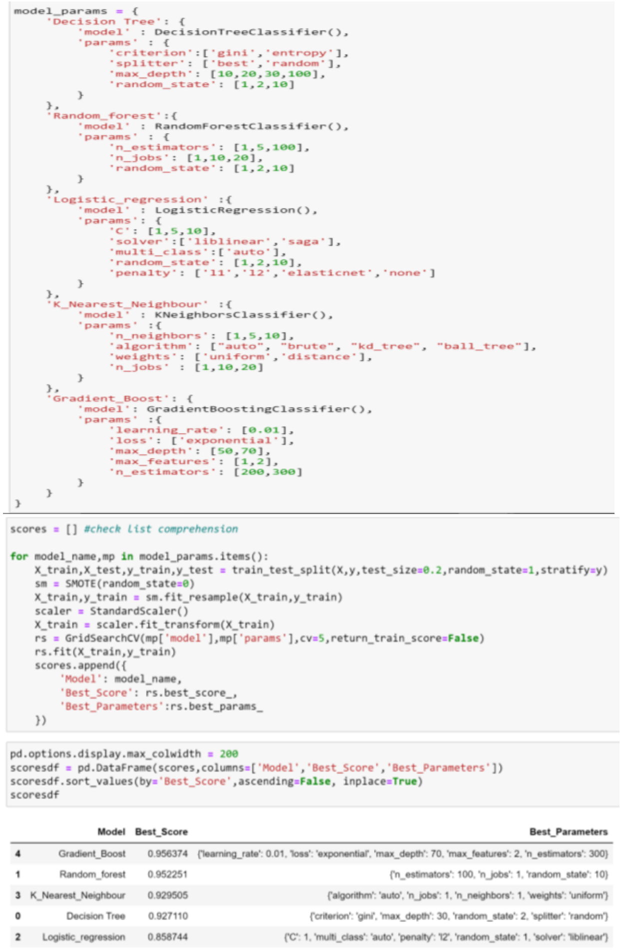
- Less Complexity and easy to interpret.
- Improved Accuracy
- Reduced overfitting
- It reduces model training time.

Machine Learning Models do not work with categorical features, so we have to convert our categorical columns into numeric data. Data can be encoded in different ways but One Hot Encoding was used in this ICA.

Normalizing and Scaling your data also helps in increasing your model's accuracy. In this ICA, the Standard Scalar library was used to improve the accuracy of our model.

HYPERPARAMETER TUNING

Hyperparameter tuning was tested on 5 algorithms to get the best parameters using Grid search as shown below:



```
model_params = {
    'Decision Tree': {
        'model' : DecisionTreeClassifier(),
        'params' : {
            'criterion':['gini','entropy'],
            'splitter': ['best','random'],
            'max_depth': [10,20,30,100],
            'random_state': [1,2,10]
        }
    },
    'Random_forest':{
        'model' : RandomForestClassifier(),
        'params' : {
            'n_estimators': [1,5,100],
            'n_jobs': [1,10,20],
            'random_state': [1,2,10]
        }
    },
    'Logistic_regression' :{
        'model' : LogisticRegression(),
        'params': {
            'C': [1,5,10],
            'solver':['liblinear','saga'],
            'multi_class':['auto'],
            'random_state': [1,2,10],
            'penalty': ['l1','l2','elasticnet','none']
        }
    },
    'K_Nearest_Neighbour' :{
        'model' : KNeighborsClassifier(),
        'params' :{
            'n_neighbors': [1,5,10],
            'algorithm': ["auto", "brute", "kd_tree", "ball_tree"],
            'weights': ['uniform','distance'],
            'n_jobs' : [1,10,20]
        }
    },
    'Gradient_Boost': {
        'model': GradientBoostingClassifier(),
        'params' :{
            'learning_rate': [0.01],
            'loss': ['exponential'],
            'max_depth': [50,70],
            'max_features': [1,2],
            'n_estimators': [200,300]
        }
    }
}
```

```
scores = [] #check list comprehension

for model_name,mp in model_params.items():
    X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=1,stratify=y)
    sm = SMOTE(random_state=0)
    X_train,y_train = sm.fit_resample(X_train,y_train)
    scaler = StandardScaler()
    X_train = scaler.fit_transform(X_train)
    rs = GridSearchCV(mp['model'],mp['params'],cv=5,return_train_score=False)
    rs.fit(X_train,y_train)
    scores.append({
        'Model': model_name,
        'Best_Score': rs.best_score_,
        'Best_Parameters':rs.best_params_
    })
```

```
pd.options.display.max_colwidth = 200
scoresdf = pd.DataFrame(scores,columns=['Model','Best_Score','Best_Parameters'])
scoresdf.sort_values(by='Best_Score',ascending=False, inplace=True)
scoresdf
```

	Model	Best_Score	Best_Parameters
4	Gradient_Boost	0.956374	{'learning_rate': 0.01, 'loss': 'exponential', 'max_depth': 70, 'max_features': 2, 'n_estimators': 300}
1	Random_forest	0.952251	{'n_estimators': 100, 'n_jobs': 1, 'random_state': 10}
3	K_Nearest_Neighbour	0.929505	{'algorithm': 'auto', 'n_jobs': 1, 'n_neighbors': 1, 'weights': 'uniform'}
0	Decision Tree	0.927110	{'criterion': 'gini', 'max_depth': 30, 'random_state': 2, 'splitter': 'random'}
2	Logistic_regression	0.858744	{'C': 1, 'multi_class': 'auto', 'penalty': 'l2', 'random_state': 1, 'solver': 'liblinear'}

Fig 10: Hyper-parameter Tuning using Grid Search

EXPERIMENTS AND RESULTS

The best ML models to use on this data depends on the correlation between the different columns in the dataset and how the data was pre-processed. Using the StandardScaler library and model tuning and also choosing several algorithms for training is also advised. The dataset used was encoded during pre-processing and scaled, the dataset shows that we have to apply supervised learning and classification models as our target column is present in the dataset and it is a categorical column. In this ICA 5, Machine Learning algorithms including the ANN algorithm were used to train on the dataset.

- Random Forest Classifier
- K Nearest Neighbours Classifier

- Gradient Boosting Classifier
- Decision Tree Classifier
- Logistic Regression

After pre-processing the data, Feature Selection was used to select dependent and independent features from the dataset, Hyperparameter tuning was used on the 5 ML algorithms to get the best parameters with the highest accuracy score. A function was then created to take in our X and Y dataset, specific model, and the best parameter from our tuning process, the function splits this data into training and test set (80% and 20% respectively), and handles the imbalance in the data by oversampling the minority class on the train set using SMOTE library, then fits it on the Model and parameters supplied, model, is evaluated and displayed on several metrics such as:

Model Accuracy: is the percentage of rightly classified outputs with wrong classifications of output.

$$Accuracy = \frac{Correct\_Predictions}{All\_Predictions}$$

Precision Score: is the percentage of the rightly identified positive outputs from all the predicted positive outputs.

$$Precision = \frac{True\_Positives}{True\_Positives \quad + \quad False\_Positives}$$

Recall Score is the percentage of the rightly identified positive output from all the actual positive output. Important when the cost of False Negatives is high.

$$Recall = \frac{True\_Positives}{True\_Positives \quad + \quad False\_Negatives}$$

F1 score is the harmonic mean of Precision and Recall and gives a better measure of the incorrectly classified cases than the Accuracy. It may be a better measure to balance Precision and Recall if there is an uneven class distribution.

$$F1\_Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The results were summarised and visualized with a bar chart. Several algorithms achieved over 80% but had a lower F1 score. Models were explored more using the SHAP library which helps us understand the relationship between features.

DISCUSSION

Results

Five classification ML algorithms and the ANN algorithm were used on this dataset. Random Forest Algorithm performed better than other algorithms, while Logistic Regression performed the lowest although getting an 83.60% accuracy. We got a very low F1 score, Precision Score, and Recall score because of the imbalance in the test set. The values obtained from the confusion matrix show the classifier’s performance.

The high accuracy gotten was a result of the hyperparameter tuning, we were able to fit the best estimators and parameters directly into our model. Random Forest Algorithm had the highest accuracy, although ANN always tends to give us better accuracy than traditional ML algorithms when we have a lot of data. Overall the models are very good at predicting unforeseen data but cannot be used at an organization or Government level because the data we trained our model on is very small.

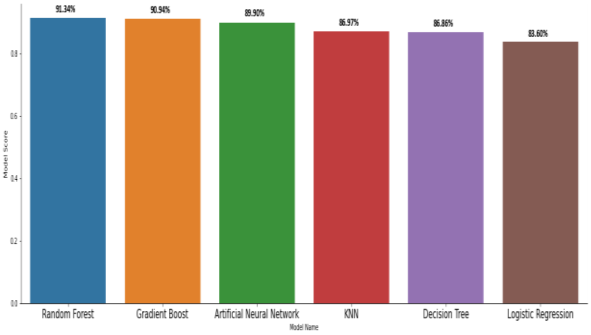
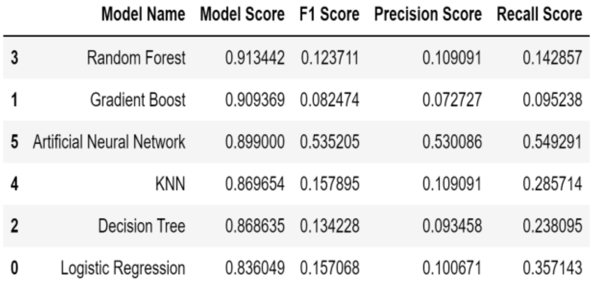


Fig 11: Model Accuracy



	Model Name	Model Score	F1 Score	Precision Score	Recall Score
3	Random Forest	0.913442	0.123711	0.109091	0.142857
1	Gradient Boost	0.909369	0.082474	0.072727	0.095238
5	Artificial Neural Network	0.899000	0.535205	0.530086	0.549291
4	KNN	0.869654	0.157895	0.109091	0.285714
2	Decision Tree	0.868635	0.134228	0.093458	0.238095
0	Logistic Regression	0.836049	0.157068	0.100671	0.357143

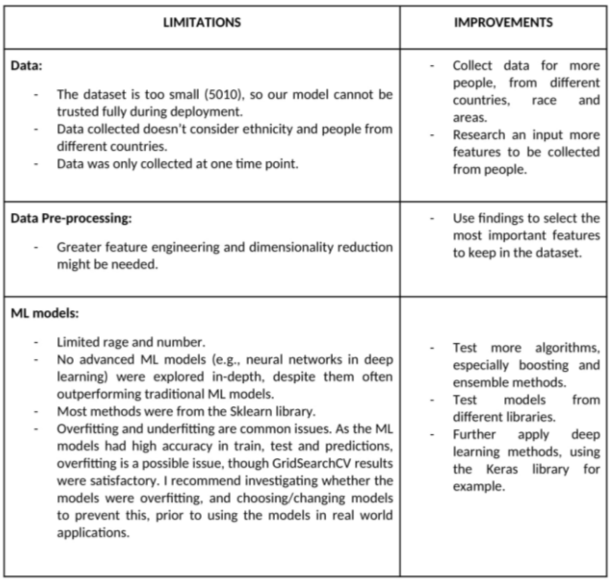
Fig 12: Model Comparison table

ETHICAL ISSUES AND RISKS

Ethical issues that were taken into account during this project include removing personal information about the patient from the data source. The data is to be used for educational purposes only and not used for production.

It is important to note that the predictions made by these models are not 100% accurate, some of these models have between 8-15% possibility of predicting a patient's stroke status wrong, therefore, the need for a human factor to validate the predictions been made by the model is essential.

ICA LIMITATIONS



LIMITATIONS	IMPROVEMENTS
Data: - The dataset is too small (5010), so our model cannot be trusted fully during deployment. - Data collected doesn't consider ethnicity and people from different countries. - Data was only collected at one time point.	- Collect data for more people, from different countries, race and areas. - Research an input more features to be collected from people.
Data Pre-processing: - Greater feature engineering and dimensionality reduction might be needed.	- Use findings to select the most important features to keep in the dataset.
ML models: - Limited rage and number. - No advanced ML models (e.g., neural networks in deep learning) were explored in-depth, despite them often outperforming traditional ML models. - Most methods were from the Sklearn library. - Overfitting and underfitting are common issues. As the ML models had high accuracy in train, test and predictions, overfitting is a possible issue, though GridSearchCV results were satisfactory. I recommend investigating whether the models were overfitting, and choosing/changing models to prevent this, prior to using the models in real world applications.	- Test more algorithms, especially boosting and ensemble methods. - Test models from different libraries. - Further apply deep learning methods, using the Keras library for example.

CONCLUSION

This project has shown that the accuracy of a model varies largely on the problem definition, the nature of the data, how messy the data is, incorrect entries or omitted entries as seen in the 'BMI' column and 'Unknown' category in the Smoking column. It is advised to always test different models on a particular data and in some cases use hyper-parameters to get the best parameters to use in the model training. In this project, I recommend using Random Forest as it has the accuracy highest score.

REFERENCES

1. A. D. Jamthikar et al. (2022) Ensemble Machine Learning and Its Validation for Prediction of Coronary Artery Disease and Acute Coronary Syndrome Using Focused Carotid Ultrasound.

2. Dev, S. et al. (2022) 'A predictive analytics approach for stroke prediction using machine learning and neural networks', Healthcare Analytics, 2, pp. 100032. doi: https://doi.org/10.1016/j.health.2022.100032.

3. G. Fang, P. Xu and W. Liu (2020) Automated Ischemic Stroke Subtyping Based on Machine Learning Approach.

4. G. Joo et al. (2020) Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea).

5. H. Xu et al. (2021) Predicting Recurrence for Patients with Ischemic Cerebrovascular Events Based on Process Discovery and Transfer Learning.

6. M. Monteiro et al. (2018) Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients.

7. A. Subudhi, M. Dash and S. Sabut, "Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier", Biocybernetics Biomedical Eng, vol. 40, pp. 277-289, 2020.

8. C. O. Johnson, M. Nguyen, G. A. Roth, E. Nichols and T. Alam, "Global regional and national burden of stroke 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016", Lancet Neurol, vol. 18, pp. 439-458, 2019.

9. Ching-Heng Lin et al., Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. Computer Methods and Programs in Biomedicine, vol. 190, pp. 105381, 2020, [online] Available: https:// doi.org/10.1016 / j.cmpb.2020.105381.