Predicting Stroke Occurrence: An Applied Machine Learning Approach

School of Science & Technology
Department of Computer Science
City, University of London

***Abstract*—The majority of brain strokes are caused on by an unanticipated obstruction of the heart's and brain's regular operations. researchers can use a variety of machine learning techniques to forecast the likelihood of a stroke occurring. This report studies the use of machine learning techniques to predict the long-term outcomes of stroke victims. We can accurately predict stroke, which is important for early treatment, by developing and analyzing several machine learning models. The dataset included 5110 observations of patients who had suffered a stroke and their modifiable risk factors.**

Keywords*—*Stroke, predictive analytics, brain, machine learning, data analysis

I. INTRODUCTION

In the UK strokes occurs every five minutes, and 100,000 people suffer a stroke each year [1]. Stroke is regarded as the second leading cause of death. The World Health Organization (WHO) estimates that 17.7 million people died from cardiovascular illnesses in 2017, with 6.7 million of those deaths attributable to stroke [2]. Age, body mass index, smoking status, average glucose level, hypertension, heart illness, and body mass index are all risk factors for stroke [3]. Physicians and medical consultants can predict when a stroke might happen by looking at risk factors. Clinical tests support this approach. This early diagnosis helps with disease detection and therapy [4]. A stroke prediction dataset from Kaggle called "healthcare-dataset-stroke-data" has been used to achieve or goals, containing 5110 observations (rows) and 12 attributes (columns). Each patient is represented by each observation, and the attributes are details about each patient's state of health. Id, gender, hypertension (yes/no), heart disease (yes/no), marital status, type of employment (children, government job, never worked, private, self-employer), type of residence (urban, rural), smoking status (formerly smoked, never smoked, smokes), and history of stroke are specifically the categorical variables. We have a person's age, BMI, and average blood glucose level as quantitative variables.

II. ANALYSIS PLAN

Blood pressure is the primary risk factor for stroke. High blood cholesterol, smoking, obesity, end-stage kidney disease, and atrial fibrillation are additional risk factors. Although there are other, less frequent causes of ischemic strokes, blood vessel blockage is frequently the reason. Either bleeding into the brain itself or into the space between the membranes of the brain might result in a haemorrhagic stroke. An aneurysm in the brain that has ruptured may cause bleeding. A physical examination is often used to make a diagnosis, which is then confirmed by imaging tests like a CT or MRI. Although ischemia, which early on often does not show up on a CT scan, can be ruled out by a CT scan, bleeding can still be ruled out by the scan.

It's no doubt the brain stroke is a very serious case to deal with and any studies conducted should be accurate, in this study, we will analyse each of the stroke symptoms and causes and visualize every possible link between attributes to conclude a positive outcome. Thus, the below questions will be the scope of the analysis.

- what are the factors related to a brain stroke?
- Which factors lead the most to a brain stroke case?
- What are the possible correlations between attributes?
- Which model type would best fit this data research?

After finding the analytical question helping us visualize more our goals for this research, we've set our objectives and planned them.

- Find a dataset which features consist of the key risk factors of stroke.
- Investigate the dataset and start making some hypothesis.
- Develop and apply an AI solution using Machine Learning tools and framework.
- Critically evaluate different models for the most accurate predictions.

III. DATA

It is crucial to use the appropriate data and incorporate essential features when building the necessary model for stroke prediction. Datasets are collections of information that are organised so that they are very simple to manipulate and modify. They can be found in many databases and online sources. A dataset that was obtained from Kaggle was used for this study. Popular machine learning and data science community Kaggle offers access to a vast collection of public data and source code [5]. Data collected contained data on 5110 patients with 12 common attributes. Eleven of the twelve attributes are input features such as age, gender, marital status, patient identifier, work type, residence type, body mass index, smoking status, glucose level, heart disease condition and binary attribute hypertension indicator, the last attribute represents the predicted result which is the output attribute indicating a patient is suffered stroke or not.

We want to construct a prediction model that can determine if a person has a high chance of having a stroke or not based on 11 distinct variables after we have thoroughly examined and cleaned the dataset.

IV. ANALYSIS

A. *Data preparation*

Finding the meaning and purpose of the data is the goal of this stage. evaluating the effectiveness of the maximum values, the mean, the median, and the standard deviation. The mean values can be affected if a datapoint has a large number of missing values. The stage after that doesn't perform a null check. Before exploring the data, we need to make sure it's ready and clean enough to have accurate results.

First thing to do is to visualise the data by columns and rows in order to get a better insight, we can use the "head()" function for this purpose. Then we will need to use the "describe()" function which shows the count, mean, standard deviation and the minimum and maximum values for the quantiles of the data of the numerical values. When pre-processing data, it is sometimes recommended to replace some missing values with the mean, median, mode, or constant value [6].

Numerous data points have numerous unique values. The performance of the model can deteriorate by having columns of data with a large cardinality [7].

The next stage of becoming acquainted with the data is to look for outliers in the columns with a high cardinality. The table below shows the number of unique values and null values for every attribute of the dataset.

Number of values	unique	null
gender	3	0
Age	104	0
hypertension	2	0
heart_disease	2	0
ever_married	2	0
work_type	5	0
residence type	2	0
avg_glucose_level	3852	0
	418	201
smoking_status	4	0
stroke	2	0

Fig.1 – Number of unique and null values in each column

In this case, we will look in glucose level which had a large number of unique values. We will try to visualize the data using matplotlib to check for outliers

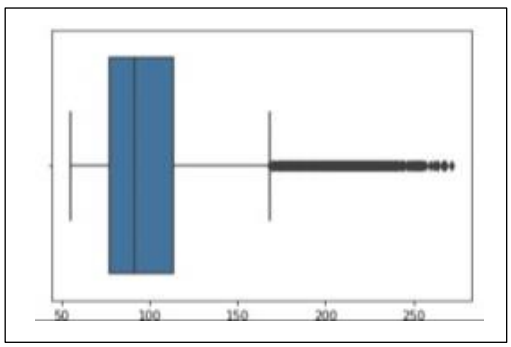


Fig.2.1 – Checking for outliers

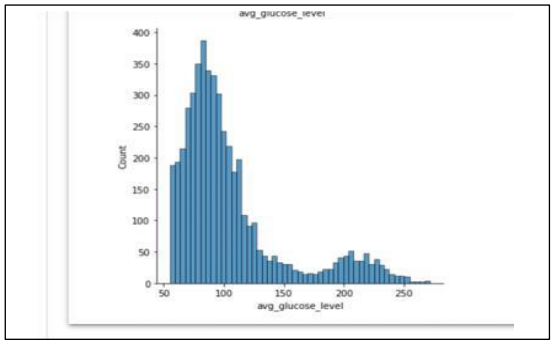


Fig.2.2 – Visualising glucose level

In fact, avg glucose, bmi, and age are related. Age, AGL, and BMI relationships need to be looked into. This explains the high level of uniqueness in the datapoints.

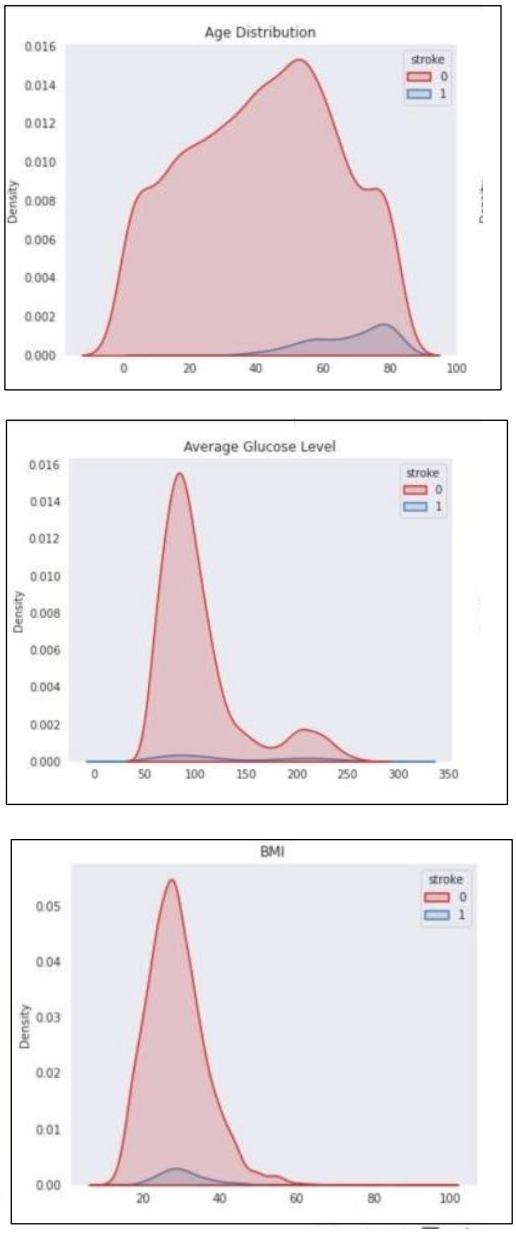


Fig.3 – investigate into Age, Average Glucose

These components appear to have a close connection. The odds of developing diabetes and having a stroke increase with age, as shown by the graphs, which proves the cardinality of those datapoints. Stroke risk appears to be correlated with glucose and BMI. Features that will significantly influence future models are also visible. It was necessary to check this feature for redundancy, null values, corruption, and cleaning.
There are too many null values in the BMI column. Another choice is to use the columns' mean average to replace the null values. The efficiency of determining the mean will first be hampered by missing data. Dropping the values is the most effective strategy because 201 is 10% of 4908 [8].

B. Data exploration and analysis

Correlation matrix is a very efficient way to explore new data. This will determine the correlation between each set of the variable combinations for the relationship analysis. This correlation value will be used to gauge how strongly any two elements of the patient's electronic health records are linearly related. We have used a colourmap in Fig. 4,

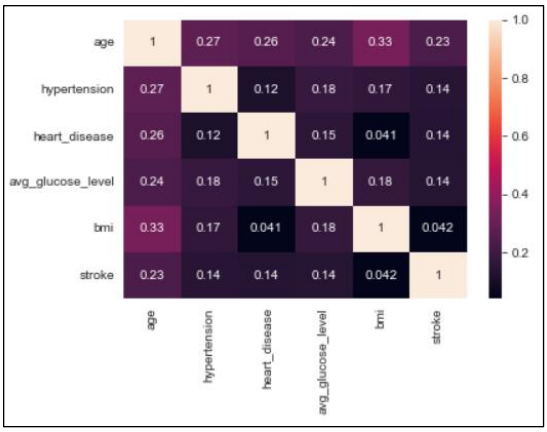
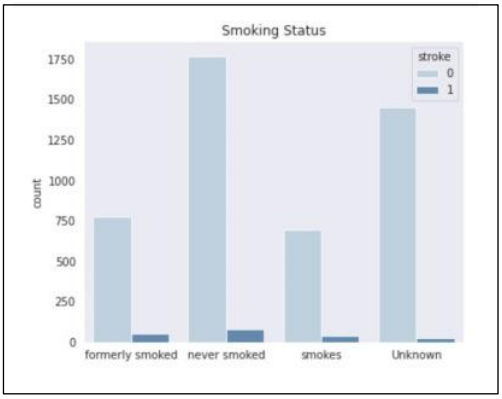
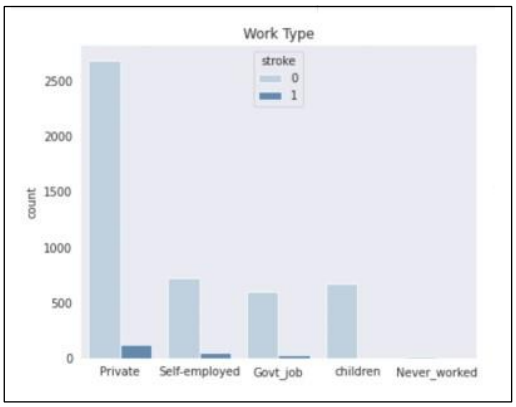


Fig.4 – relationship analysis using correlation matrix

This correlation value will be used to evaluate how strongly any two elements of the patient's electronic health records are linearly related. As we can notice in the fig.4, most values are near 0 which indicates a weak correlation between these attributes.

Now we can start making assumptions on dataset. Work type, residency type, can possibly affect one getting a stroke.





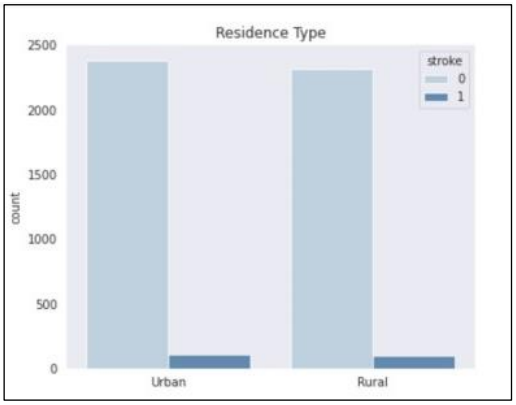
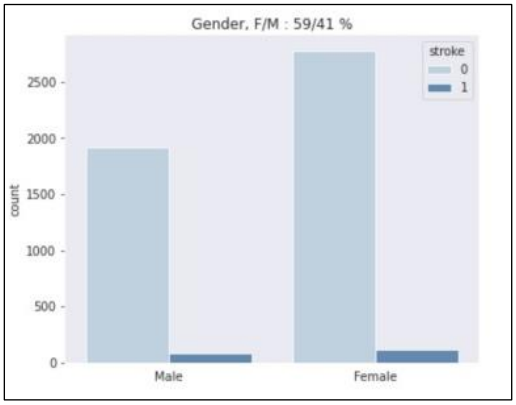
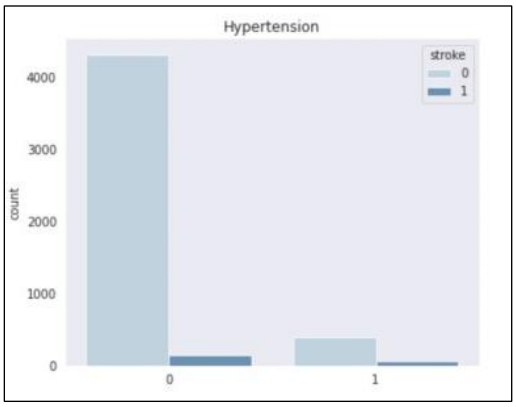


Fig.5.1 – Bar chart for work, type of residency, and glucose level

We couldn't see any relevant connection between stroke and these features. We moved to the Gender, hypertension and heart disease features.





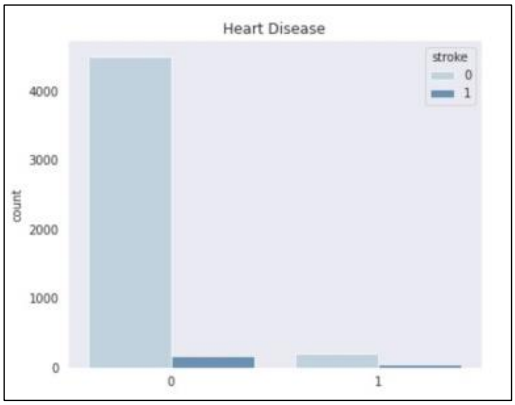


Fig.5.2 – Bar chart for gender, Hypertension, heart disease

Given the ratio between each bar, these bars appear to be about the same length, suggesting that there is some relationship. Additionally, there appear to be inequity in the stroke bar feature.

C. Data pre-processing

It was discovered during the exploratory phase that the data is "unbalanced." Only 209 people in the stroke column experienced a stroke, leaving 4699 people without one. As a result, the model's accuracy was 96% after it was built and tested. This is an unusually high accuracy because the model correctly identified the majority class, which is 0 (no stroke). So, we took care of imbalance data.

```
Patient that dont stroke ratio: 0.04258353708231459
Patient that have stroke ratio : 0.9574164629176855

0    4699
1     209
Name: stroke, dtype: int64
```

Next step is to encode categorical data using OneHotEncoder from sklearn library. We use this technique because machine learning models cannot compute correlations between columns containing string values. Splitting was necessary for the training and testing sets.

Feature scaling was also applied which scales all variables to make sure they all take values in the same scale preventing one feature to dominate the other.

D. Machine learning models

After preparing the data, it's time to deploy models and compare results in terms of accuracy. Starting with classification models, Logistic Regression, Random Forest, Decision Trees, Naïve Bayes, K-Nearest Neighbors and Support Vector machine.

We tried to use the gradient boosting algorithms, known as XgBoost and CatBoost, which are two automatic machine learning accomplishes predictive modelling using high-performance machine learning techniques.

V. Results and Reflexions

After training and testing each model, we came up with this table below which shows results for all models.

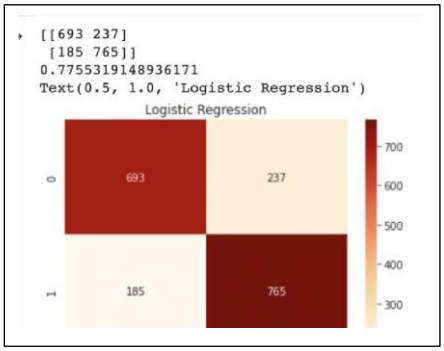
- Precision: how close two or more measurements are to each other.
- Recall: The model capacity to find all the relevant cases within a data set.
- F1-score: the harmonic means of precision and recall.
- Cross Validation: uses data in different portions to test and train a model on different iterations.

	precision		recall		f1-score		Cross Validation	
	class 1	class 0	class 1	class 0	class 1	class 0	Acc %	SD%
LR	0.77	0.79	0.80	0.75	0.78	0.77	78.01	1.78
RF	0.98	0.96	0.96	0.99	0.97	0.97	97.31	0.64
DT	0.95	0.96	0.96	0.96	0.96	0.96	95.45	1.08
NB	0.57	0.99	1.00	0.24	0.61	0.38	62.93	1.10
SVM	0.94	0.90	0.90	0.95	0.93	0.92	92.59	1.34
K-NN	0.90	0.99	0.99	0.88	0.94	0.93	93.18	1.20
Xgb	0.99	0.96	0.99	0.96	0.97	0.97	97.17	0.55
Catb	0.99	0.96	0.96	0.99	0.97	0.97	97.27	0.74

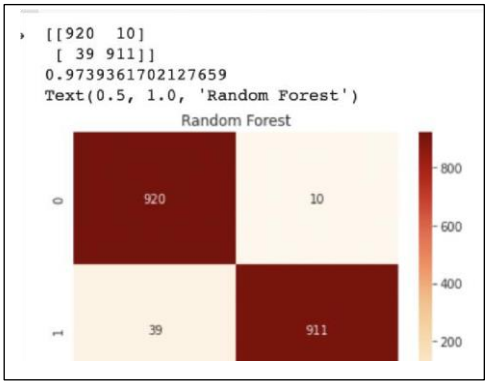
Fig6. – Bar chart for gender, Hypertension, heart disease

Our machine learning classifiers perform fairly well compared to prior research (0.97 AUC). The performance of our model can be influenced by a number of different factors. We may spend a lot of time investigating and cleaning up our data, which could be the first contributing component. Supervised machine learning attempts to optimize its features. Needing a significant amount of labelled data. We carefully labelled and cleaned this dataset. The dataset's quality was the second determining criterion; it was well-labeled and free of outliers. Several data preprocessing procedures were carried out to address this, including balancing dataset feature scaling and OneHotEncoder.

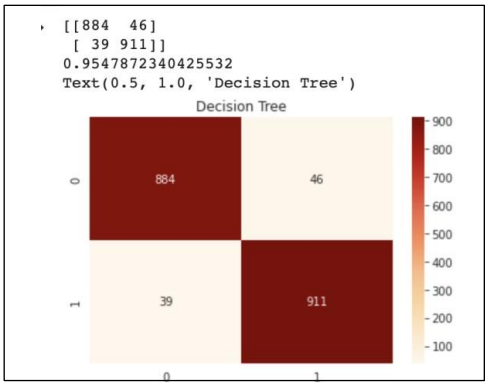
Confusion Matrix:
Logistic regression:



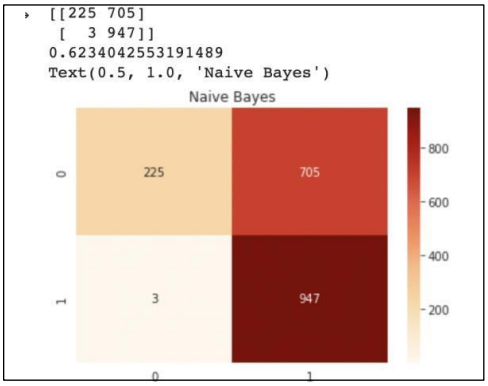
Random forest:



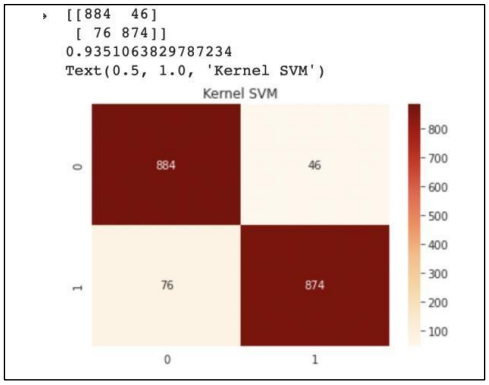
Decision Tree:



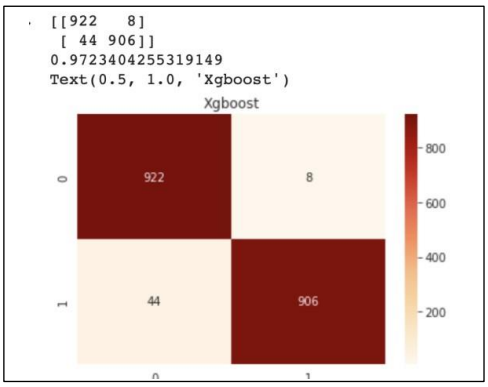
Naïve Bayes:



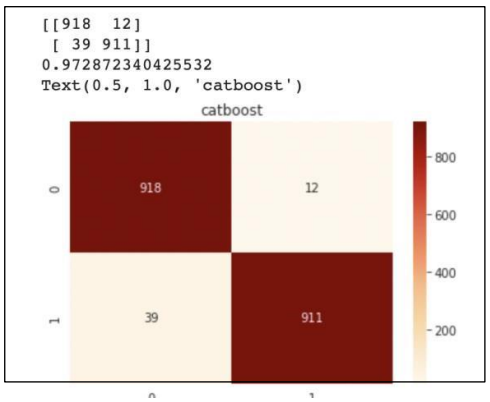
SVM:



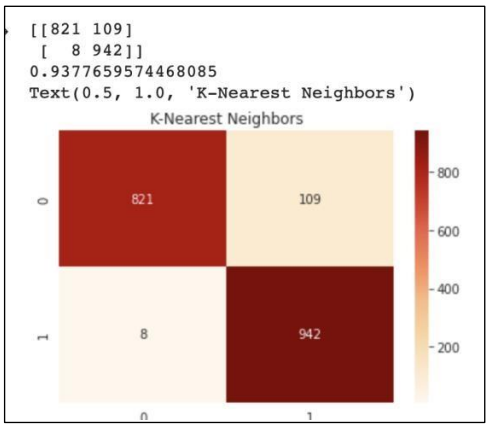
Xgboost:



Catboost:



K-Nearest Neighbors:



REFERENCES

[1] Stroke Association, "Stroke Statistics," Stroke Association, 2020. https://www.stroke.org.uk/what-is-stroke/stroke-statistics

[2] J. M. Shikany, M. M. Safford, O. Soroka, P. Newby, T. M. Brown, R. W. Durant, and S. E. Judd, ''Abstract P520: Associations of dietary patterns and risk of sudden cardiac death in the reasons for geographic and racial differences in stroke study differ by history of coronary heart disease,'' Circulation, vol. 141, no. 1, p. AP520, Mar. 2020.

[3] P. B. Gorelick, "New horizons for stroke prevention: PROGRESS and HOPE," The Lancet Neurology, vol. 1, no. 3, pp. 149–156, Jul. 2002, doi: 10.1016/s1474-4422(02)00070-4.

[4] "(PDF) Stroke prediction through Data Science and Machine Learning Algorithms," ResearchGate. https://www.researchgate.net/publication/352261064_Stroke_prediction_through_Data_Science_and_Machine_Learning_Algorithms

[5] Kaggle, "Kaggle: Your Home for Data Science," Kaggle.com, 2019. https://www.kaggle.com/

[6] A. Kumar, “Python - Replace Missing Values with Mean, Median & Mode,” Data Analytics, Jul. 23, 2020. https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/

[7] “Missing Data: Two Big Problems with Mean Imputation,” The Analysis Factor, Oct. 15, 2020. https://www.theanalysisfactor.com/mean-imputation/

[8] B. Angelov, “Working with Missing Data in Machine Learning,” Medium, Dec. 13, 2017. https://towardsdatascience.com/working-with-missing-data-in-machine-learning-9c0a430df4ce

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.