PES UNIVERSITY
Data Analytics- EC campus
Section: F,G,I and J

Format for Literature Survey Report

1.Project Title	Brain Stroke Prediction	
2.Team Name	ATOM	
3.Team Members	SRN1:PES2UG20CS431	Name1: Emil Bluemax
	SRN2:PES2UG20CS433	Name2:J.P.DANIEL CHRISTOPHER
	SRN3:PES2UG20CS414	Name3:ADITYA R KHOT
4.Dataset used	stroke-prediction-dataset.csv	
5.Link for the Dataset	https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset	
6. Github link:	https://github.com/danielchristopher513/Data_analytics_team_atom	

7.Problem Statement:

Stroke is a disease that affects the arteries leading to and within the brain.

A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot or ruptures.

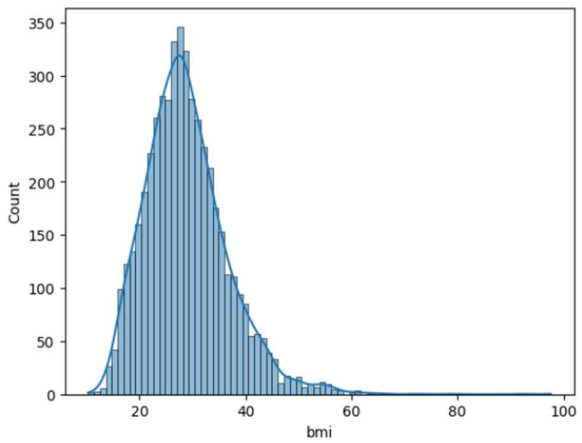
According to the WHO, stroke is the 2nd leading cause of death worldwide.

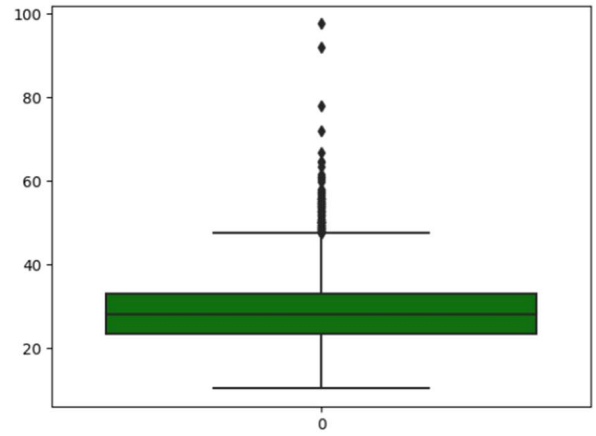
80% of the time these strokes can be prevented, so putting in place proper education on the signs of stroke is very important.

Early detection of stroke is a crucial step for efficient treatment and ML can be of great value in this process.

8.EDA and Visualization

There are a total of 5110 rows and 12 attributes in the dataset.
After performing the initial EDA on the dataset we observed that for the attribute bmi there are 201 N/A values.
On visualising the bmi attribute we observe that bmi values are positivly skewed and there exist outliers.

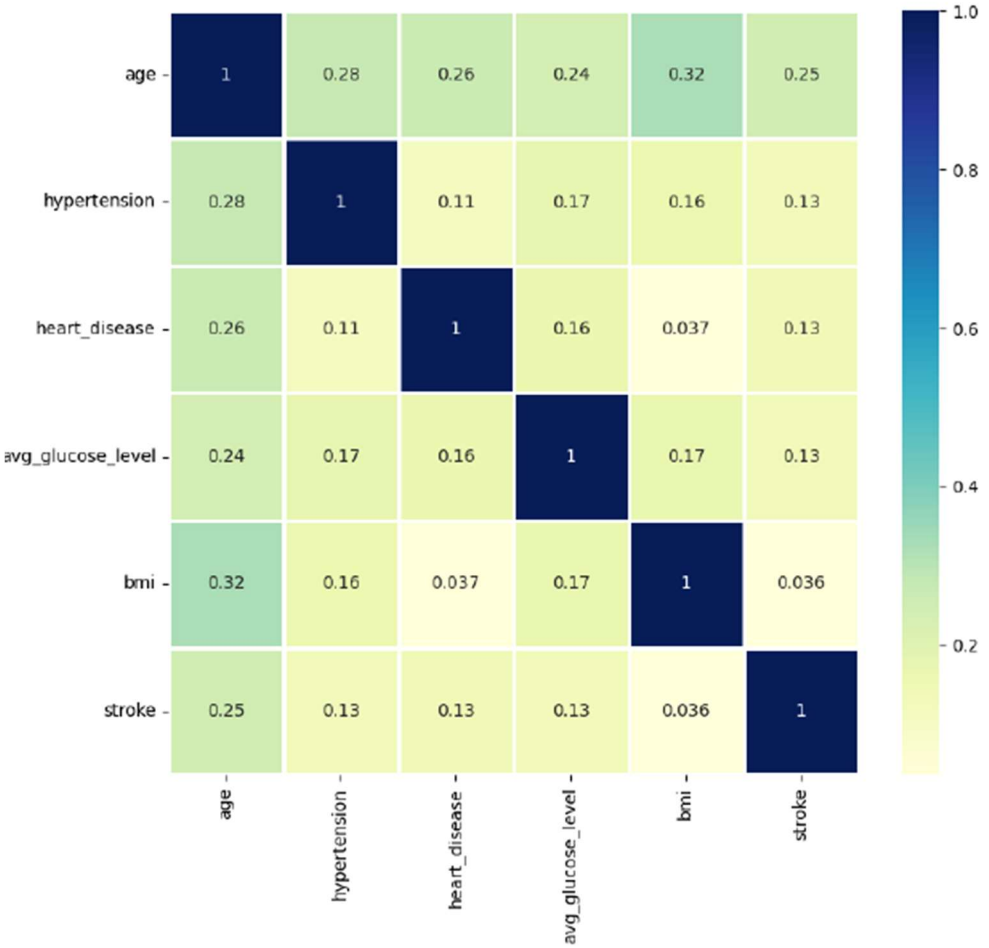





```
People who got stroke and their BMI is NA: 40
People who got stroke and their BMI is given: 249
percentage of people with stroke in Nan values to the overall dataset:
16.06425702811245
```

Since we have very few instances where stroke was positive, we can't remove those N/A values . Hence we fill the missing N/A values are imputed using the median value of bmi as it is least affected by the outliers.

After plotting the correlation matrix between the attributes of the dataset we obtain a heatmap

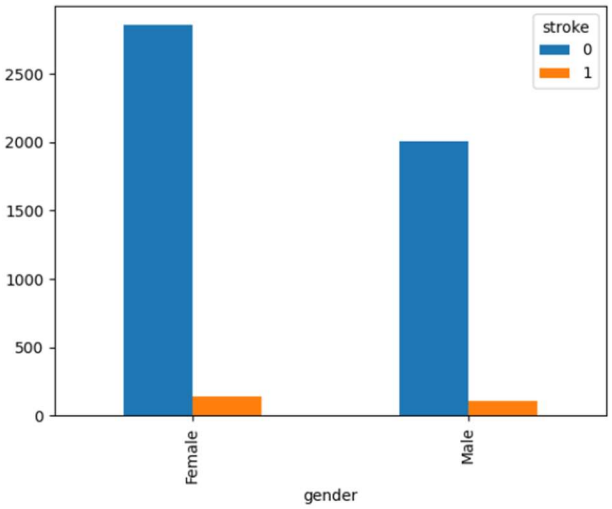


From this we observe that there is very low correlation among the attributes, the highest correlation observed was between age and bmi with a value of 0.32 all other correlation value's were less than 0.3

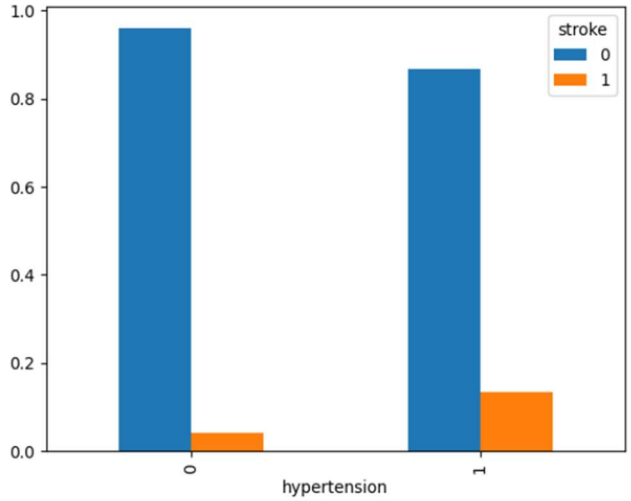
```
In [366]:
df['stroke'].sum()/len(df)*100
Out[366]:
4.87279843444227
```

Our main target function is 'stroke' and the instances who got stroke are very less and are in the minority (249) this is only 4.9 % of the instances, thus the distribution is very imbalanced. Hence we need to oversample the minority class for a more accurate and better prediction.

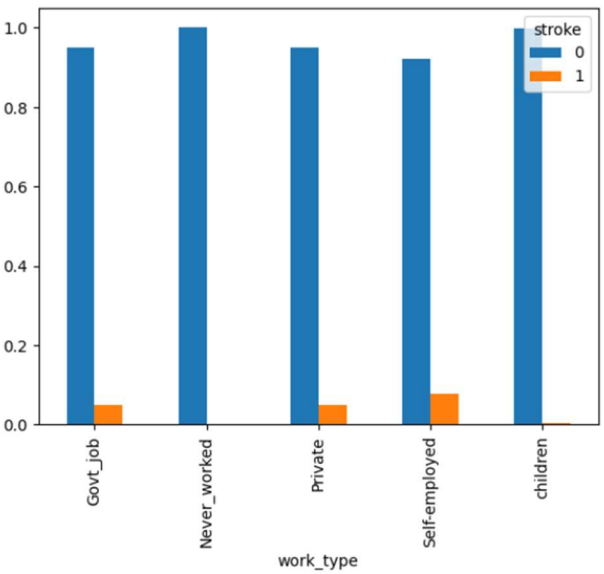
All the categorical values of attributes are converted into nominal numeric values for future processing



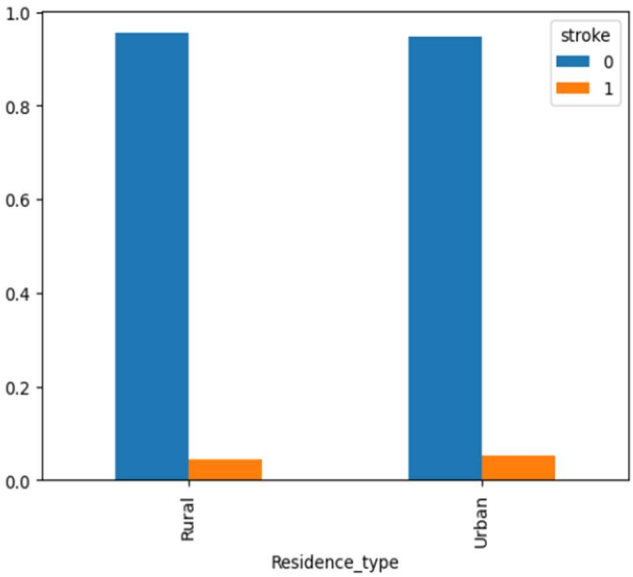
Relation between gender and stroke



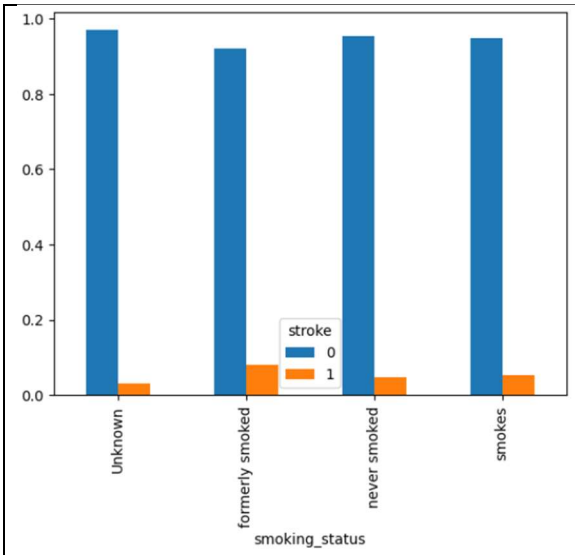
relation between hypertension and stroke



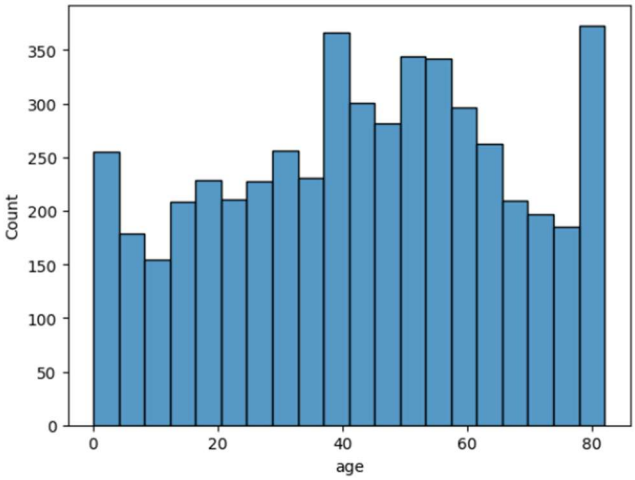
relation between work type and stroke

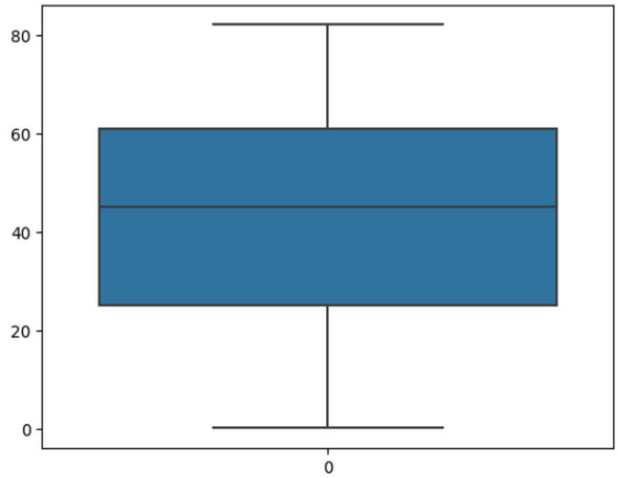


relationship between residence type and stoke

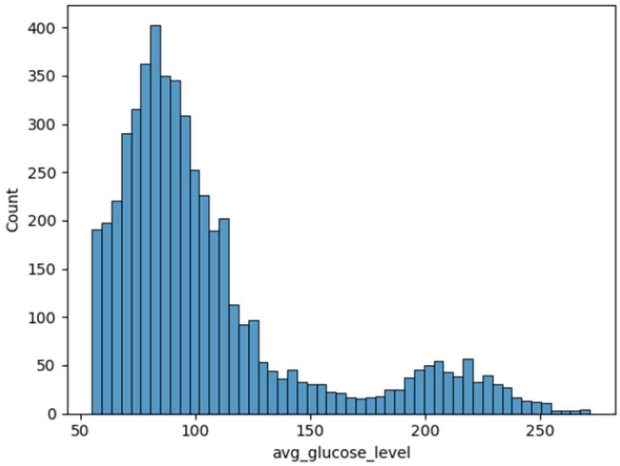


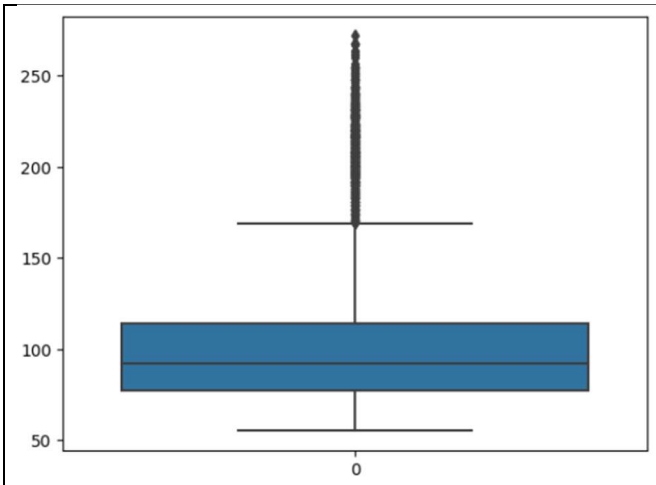
relation between smoking status and stoke





For the age parameter the data seems to be normaly distributed with no outliers





For avg glucose level the data is positivly skewed and there exist a lot of outliers

9. Summarize the Literature survey

1) This paper discussed the need of prediction of stroke. Seven different machine learning algorithms are used for the prediction of stroke. The prediction is performed on the basis of various features like age, gender, hypertension, heart disease, smoking status, BMI, average glucose levels, work type, residence type, and marital status. Among seven machine learning algorithms, AdaBoost Classifier, XGBoost Classifier, and the Random Forest Classifier gives the highest accuracies of95%, 96%, and 97% respectively.These were obtained by confusion matrices The accuracy results proved that these three algorithms can be used to predict stroke in real-life.Further, deep learning models can also be utilised in order to predict the risk of stroke more effectively without performing clinical tests. In future, similar approaches and methods can also be used to determine possible risk factors for various other diseases.

2)As the second-highest death worldwide, detecting stroke in an early stage will help the patient to have better and less harmful medication, reduce the cost of medication, help aid people's health, predict accurate outcomes. The highest accuracy value was obtained by Random Forest with a result of 94%. Random Forest has the advantage in classifying data because it works for data that has incomplete attributes, and is good for handling large sample data.For further research, an application or smart system suggested to be made in predicting the diagnosis of stroke, by adding other algorithms in machine learning to create better and more accurate models. It is recommended also to add attributes to the dataset such as recent strenuous activity, and occupations to strengthen the prediction.

3)This paper has considered gender, age, hypertension, heart disease, averag_ glucose level, BMI, smoking status feature attributes to predict stroke. The performance evaluation reveals that weighted voting provided the highest accuracy of about 97% compared to the commonly used other machine learning algorithms.These were obtained by confusion matrices . As a result, the weighted voting can be considered for the prediction of stroke. The relationship between these diseases and possibility of occurring stroke in a human individual has been evaluated.

4)This paper predicts the risk of strokes based on the classification approach. Eight different learning machines have been used to this data set, which reveals a 98% accuracy rate in voting classifier than other well-known approaches for prediction of stroke risks. As a consequence, voting classifier might be used to determine the risk of stroke. If a stroke can be predicted at an early stage, it is beneficial to the patient to prevent the stroke. Thus, this approach can be used in healthcare to identify the risk of stroke. In the future, brain imaging such as computed tomography and MRI may be combined as a feature to improve model performance. IoT system can be added in this system where the sensor will send the reading data to the cloud server. The machine learning model is also incorporated so that the webserver automatically receives the predicted accuracy.

5)In this paper, we proposed an ensemble framework (BSPE) based on supervised ML techniques for improving brain stroke prediction performance.An algorithm was proposed known as Hybrid Ensemble Learning for Brain Stroke Prediction (HEL-BSP) also feature engineering algorithm are reused - Composite Metric based Feature Selection (CMFS). The ensemble is made up of ML models such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), KNeighbours classifier, Gradient Boosting and Stochastic Gradient Descent (SGD). A prototype application is built using Python data science platform to evaluate the proposed framework and the underlying algorithm. The experimental results revealed that the ensemble of the prediction models

with majority voting approach could outperform individual prediction models. this analysis is focused on data-driven approach for brain

6) In this paper multiple models (SVM,RF,LR, DT,KNN) are explained in detail and all the pre analysis steps are explained and how to analyse the predictions are explained in detail it concluds that by using the proposed work we can predict whether a person is suffering from brain stroke or not very easily in the earlier stage using the requested parameters. Predicting the disease might decrease the causes of death. KNN model is used as the underlying prediction model and UI webpage is used to take the input parameters

7) The result was analyzed using confusion matrix and the accuracy is calculated using the accuracy formula, which resulted in showing higher accuracy rate for the combination od DT, PCA and NN model (Proposed model) which was 95%, 95.2% and 97.7% for STCL, STTIA and STAN datasets when compared to other model like NN model, PCA & NN model combinations which resulted in lower accuracy rates [82.5%, 78.6%, 90.9% and 92.5%, 92.9%, 95.5%] . This gives the conclusion that the combination model of Decision Tree, Principal Component Analysis, and Back Propagation NEural Network (DT, PCA and NN) is Better than the other teo NN, and PCA & NN models.

8) In this paper, the confusion matrix was used for analysis of the different algorithms used for accuracy like Naive Bayes theorem, J48, K-NN algorithm and Random Forest Algorithm, where from the performance analysis we came to know that Naive Bayes Performs better than all the other methods providing at accuracy of around 0.872, 0.801, 0.803, 0.955 for different classes like Ischemic stroke, Hemorrhagic stroke, Mini stroke, Brain Steam Stroke. and the other methods like K-NN, Random forest Classification and J48 also peformed well in the case while finding the accuracy greater than Naive bayes Algorithm.

9) The data alanysis techniques like Adaboostm1, J48, claasifiers are equally did well in getting accuracy better than that of the algorithms like Naive bayes classifier and Bayes Network classifiers by getting an accuracy of 0.957 and 0.957 for both TP rate and FP Rate respectively which is a greater accuracy when compared to the Bayes netwotk classifier and Naive Bayes Classifier methods

10. What is the specific problem your team is going to solve?

Predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status using the designed predictive model.

References:

1.S. Gupta and S. Raheja, "Stroke Prediction using Machine Learning Methods," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2022, pp. 553-558, doi: 10.1109/Confluence52989.2022.9734197.

2.N. S. Adi, R. Farhany, R. Ghina and H. Napitupulu, "Stroke Risk Prediction Model Using Machine Learning," 2021 International Conference on Artificial Intelligence and Big Data Analytics, 2021, pp. 56-60, doi: 10.1109/ICAIBDA53487.2021.9689740.

3.M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun and M. S. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020, pp. 1464-1469, doi: 10.1109/ICECA49313.2020.9297525.

4.R. Islam, S. Debnath and T. I. Palash, "Predictive Analysis for Risk of Stroke Using Machine Learning Techniques," 2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 2021, pp. 1-4, doi: 10.1109/IC4ME253898.2021.9768524.

5.A. Devaki and C. V. G. Rao, "An Ensemble Framework for Improving Brain Stroke Prediction Performance," 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2022, pp. 1-7, doi: 10.1109/ICEEICT53079.2022.9768579.

6.V. Krishna, J. Sasi Kiran, P. Prasada Rao, G. Charles Babu and G. John Babu, "Early Detection of Brain Stroke using Machine Learning Techniques," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021, pp. 1489-1495, doi: 10.1109/ICOSEC51865.2021.9591840.

7.M. Sheetal singh, Prakash choudhary, " Stroke Prediction using Artificial Intelligence ", 8th Annual

Industrial Automation and Electromechanical Engineering conference(IEMECON) 2017 DOI: 10.1109/IEMECON.2017.8079581

8.Tasfia Ismail Shoily, Tajul Islam, , Sumaiya Jannat, Sharmin Akter Tanna,Taslima Mostafa Alif, Romana Rahman Ema. " Detection of Stroke disease using Machine Learning Algorithms " 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) DOI: 10.1109/ICCCNT45670.2019.8944689

9.V. J. Jayalaxmi, V geetha, M. Ijaz, " Analysis and Prediction of Stroke using Machine Learning Algorithms " 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) | 978-1-6654-2829-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAECA52838.2021.9675545