**Module 7**

**Final Research Paper**

Devika Rajendran

Colorado State University Global

MIS581: Capstone: Business Intelligence and Data Analytics

Dr. Steve Chung

08/27/2023

**Abstract**

Brain stroke is a serious medical condition that can cause long-term disability or even death. Therefore, it is important to be aware of the signs and symptoms of a stroke and seek immediate medical attention if they occur. This project aims to develop various machine learning models, such as Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT) Classification, to predict the likelihood of stroke in patients based on a variety of demographic, lifestyle, and health factors. The case study also focuses on comparing all three predictive models to identify the better-performing models. We use a dataset of over 5,000 patient records to train and evaluate several different models, including logistic regression and random forest. The project goes through various steps: data preprocessing, descriptive analysis, explanatory analysis, and predictive modeling. The results suggest that our model has the potential to be a useful tool for healthcare providers in identifying patients at high risk of stroke and initiating early interventions to prevent stroke occurrence. All the predictive models showed an accuracy score above 90%, with the Decision Tree model providing the best AUC value of 0.6.

**Introduction**

Brain stroke is a severe and potentially life-threatening medical condition that arises when blood flow to the brain is interrupted, causing damage to brain tissue. It represents a significant global health concern, impacting millions of individuals annually. The World Health Organization (WHO) reports an alarming figure of approximately fifteen million stroke cases yearly, resulting in a devastating loss of life, with one person dying every four to five minutes in the affected population. In the United States, stroke ranks as the sixth leading cause of death, affecting around 795,000 individuals who experience severe effects of strokes, which can cause significant challenges in their daily lives (Tazin et al., 2021).

The consequences of brain stroke can be severe and long-lasting, with many individuals experiencing significant deficits in physical, cognitive, and emotional functioning. A brain stroke can also have emotional and psychological impacts. Many individuals who have suffered a stroke experience depression, anxiety, and other mood disorders. They may also struggle with social isolation and a decreased quality of life, as they can no longer engage in activities they once enjoyed.

As brain stroke significantly impacts individuals and their families, it is critical to develop effective strategies for preventing and managing the condition. This includes developing more accurate predictive models and improving access to timely and effective interventions such as rehabilitation and support services (Yu et al., 2020). Doing so can help minimize the incidence and severity of brain stroke and improve the quality of life for those affected by this debilitating condition.

**Objectives**

The primary aim of this research is to develop a machine learning-based prediction model for brain stroke. To achieve this aim, this study will leverage a dataset of brain stroke patients to train and test our model. Specifically, we will investigate the relationship between various risk factors and the likelihood of experiencing a brain stroke. This project aims to identify the most critical predictors of brain stroke and develop a model that can accurately predict the likelihood of experiencing a stroke.

This research project aims to analyze the impact of various factors on the likelihood of experiencing a brain stroke. Specifically, the project investigates the following factors:

* Gender
* Residence type
* Glucose level
* Heart disease
* Hypertension
* Marital status
* Work type
* Smoking Status
* BMI

By analyzing the relationship between these factors and the likelihood of experiencing a brain stroke, the study aims to identify the most important predictors of the condition. This research project aims to leverage machine learning algorithms, such as Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT) Classification, to train various models for accurate prediction of brain stroke. The case study will compare the models to identify the most effective and accurate prediction model for brain stroke.

**Overview of the Study**

The dataset contains both categorical and numerical variables that help forecast the chance of a stroke in patients. These variables cover details like age, gender, smoking habits, heart disease, hypertension, marital status, residence and work type, average glucose level, and BMI. The table also indicates the unit of measurement associated with numerical variables. For instance, age might be measured in years, and BMI might be measured in kg/m². Categorical variables are described with lists of all the possible categories or options that the variable can take.

Table 1:

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Description** |
| ID | Nominal | Unique identifier for each patient |
| Gender | Nominal | Gender of the patient (Male/Female/Other) |
| Age | Interval | Age of the patient (in years) |
| Hypertension | Binary | Whether the patient has hypertension or not (1/0) |
| Heart disease | Binary | Whether the patient has any heart disease or not (1/0) |
| Ever married | Nominal | Whether the patient is currently married or not (Yes/No) |
| Work type | Nominal | Type of occupation of the patient (Children/Government job/Never worked/Private/Self-employed) |
| Residence type | Nominal | Type of residence of the patient (Urban/Rural) |
| Avg glucose level | Interval | The average glucose level in the patient's blood (in mg/dL) |
| BMI | Interval | Body mass index of the patient (in kg/m^2) |
| Smoking status | Nominal | Whether the patient is a smoker or not (formerly smoked/never smoked/smokes) |
| Stroke | Binary | Whether the patient has had a stroke or not (1/0) |

The case study is intended to compare brain stroke predictive models using Decision Tree (DT) Classification, Random Forest (RF), and Logistic Regression (LR) algorithms. All three machine learning algorithms used in this study are widely used in healthcare and are effective in predicting various medical conditions. Random Forest (RF) is a type of decision tree-based model that can handle large datasets and is known for its accuracy. Logistic Regression (LR) is a linear model that is commonly used in medical research to predict the probability of an event occurring. Decision Tree (DT) Classification is a tree-based model useful for predicting outcomes based on a set of variables.

**Research Questions and Hypotheses**

**Research Questions**

1. Is the probability of having a stroke higher for individuals who have heart disease?
2. Is there a higher likelihood of strokes in patients with hypertension than in those without?
3. Are men more susceptible to experiencing a brain stroke?
4. What is the range of average glucose levels that correlates with a higher likelihood of stroke occurrence?

**Overview of Hypotheses**

Hypothesis testing is a way to check if the assumptions about population parameters in a group of data are correct. It helps you determine the relationship between two statistical variables in a dataset. For performing hypothesis testing on a business problem, the initial step is to state the null and alternate hypotheses for the test. The decision to accept or reject the null hypothesis is based on the result of the statistical test.

Here, the hypotheses are designed to investigate the impact of various demographic and lifestyle factors on stroke risk. The null hypothesis suggests that there is no impact or relationship between the variables, whereas the alternative hypothesis proposes that there is one. By examining these hypotheses and studying the outcomes, researchers and healthcare experts can gain a deeper insight into the elements that contribute to the risk of stroke. This understanding can aid in creating improved strategies for preventing and treating strokes in patients.

**Hypothesis for the Project**

1. Is the probability of having a stroke higher for individuals who have heart disease?

Null hypothesis (Ho): The occurrence of heart disease does not increase the likelihood of a stroke.

Alternate hypothesis (Ha): The occurrence of heart disease increases the likelihood of a stroke.

1. Is there a higher likelihood of strokes in patients with hypertension than in those without?

Null hypothesis (Ho): There is no significant difference in the likelihood of a stroke between patients with and without hypertension.

Alternate hypothesis (Ha): There is a significant difference in the likelihood of a stroke between patients with and without hypertension.

1. Are men more susceptible to experiencing a brain stroke?

Null hypothesis (Ho): Gender does not have a significant impact on the likelihood of a stroke. Alternate hypothesis (Ha): Gender has a significant impact on the likelihood of a stroke.

1. What is the range of average glucose levels that correlates with a higher likelihood of stroke occurrence?

Null hypothesis (Ho): There is no correlation between the range of average glucose levels and the likelihood of stroke occurrence.

Alternate hypothesis (Ha): There is a correlation between the range of average glucose levels and the likelihood of stroke occurrence.

**Literature Review**

The research study by Tazin et al. (2021) aimed to develop accurate and robust machine-learning models for detecting and predicting stroke disease. The study used machine learning algorithms like random forest, logistic regression, and k-nearest neighbor to build models capable of effectively predicting the likelihood of stroke in patients. The study was conducted on a dataset of 10,000 patients, sourced from the University of California, Irvine (UCI) Machine Learning Repository. The outcomes demonstrated that the newly developed machine learning models accurately anticipated the probability of stroke in patients. This study sheds light on the potential of robust machine learning techniques and algorithms in enhancing the precision of stroke prediction models, contributing valuable insights into their advancement.

The research findings indicate that the random forest approach is more effective than other methods in predicting brain strokes, especially when evaluating cross-validation measurements. This study could be expanded by using a more extensive dataset and trying out different machine-learning models.

The study by Kaur et al. (2022) provides valuable insights into the development of early stroke prediction methods for the prevention of strokes. The use of machine learning algorithms and techniques in this study highlights the potential of these approaches in improving the accuracy of stroke prediction models. The researchers utilized machine learning algorithms, like support vector machines and artificial neural networks, to develop models that could accurately predict the risk of stroke in patients. The study utilized a dataset of 500 patients who had experienced a stroke within the last year. This research also collected data on patients' genetic and lifestyle factors, such as family history of stroke and physical activity level, which Tazin et al. did not consider. The results of the study showed that the developed machine learning models were able to accurately predict the risk of stroke in patients up to one year in advance.

The research by Sirsat et al. (2020) aims to assess the role of machine learning (ML) in addressing stroke-related issues, including but not limited to prevention and identification of risk factors, diagnosis, treatment, and prognostication. The study reviewed a total of 39 relevant studies and analyzed each one individually. Although numerous ML applications exist for stroke-related problems, the article only presents the most advanced studies in each category. This experimental analysis involved 77 regression techniques from 19 ML families over 84 datasets and analyzed optimal results for each ML approach. The research article also discusses the clinical implications of their findings.

Emon et al. (2020) conducted a study that examined how well different machine-learning methods can predict strokes. They used various sets of data to test each approach's performance and then thoroughly analyzed the outcomes. The research suggests that certain machine learning approaches, such as the k-nearest neighbor algorithm, effectively predict strokes. One of the strengths of this study is the use of multiple datasets to evaluate the performance of each machine-learning approach. This provides a more comprehensive analysis of the results and allows researchers to understand the most effective methods better. The authors also provide a detailed discussion of the implications of their findings and suggest potential areas for future research.

**Research design**

**Methodology**

This research is confirmatory in nature. The dataset used in this study is obtained from kaagle.com. The brain stroke dataset is a comprehensive collection of health-related information gathered from individuals. It encompasses a range of demographic and lifestyle attributes. The dataset chosen for the analysis comprises 5110 observations and 12 attributes. Each observation represents a patient and encompasses data utilized for forecasting the probability of a stroke. Variables like heart disease, BMI, hypertension, average glucose level, and stroke history provide relevant information about the patient's health status. The dataset also includes information about the patient's marital status, residence type, and work type, which could help to identify environmental factors that could be contributing to stroke risk. The binary variable stroke indicates whether a patient has experienced a stroke or not, serving as the primary focus of analysis.

The study will use various machine learning algorithms, including Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT) Classification, to develop and validate a prediction model for brain stroke. The performance of the models will be evaluated using various metrics, including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC).

The study will also investigate the relationship between each risk factor and the likelihood of experiencing a brain stroke using logistic regression analysis. The analysis results will be used to identify the most significant predictors of the condition and inform the development of the prediction model.

**Methods**

This capstone project will use Python libraries and tools to analyze the brain stroke dataset. Python libraries provide a powerful and versatile toolkit for developing and implementing the machine learning-based prediction model for brain stroke. By leveraging the capabilities of these libraries, researchers can efficiently preprocess, analyze, and visualize the data and develop accurate and reliable predictive models for brain stroke.

Before the development of the prediction model, the dataset will be preprocessed to prepare it for analysis. The brain stroke database has many "Unknown" data in the smoking status feature, and the BMI feature contains some null values, which can affect the predictions. This Data needs to be cleaned, and the research will use Panda's built-in functions for data cleaning and manipulation. The data will be transformed to make it suitable for analysis. This may include converting categorical variables to numerical variables. Then, the dataset will be split into training and testing sets. The training set will be used to train the prediction model, and the testing set will be used to validate the model.

Matplotlib and Seaborn libraries can be used to create visualizations of the Brain Stroke Prediction dataset, which can help identify patterns and relationships between different variables. NumPy can be used to perform various mathematical operations on the Brain Stroke Prediction dataset, such as calculating the mean, median, and standard deviation of the quantitative variables.

To test this hypothesis, logistic regression analysis can be used. Logistic regression can help to model the relationship between each predictor variable and the likelihood of stroke occurrence, controlling for other relevant variables. Also, the chi-square test can be utilized to examine the association between two types of categories or groups. For instance, A chi-squared test could be used to assess whether there is a significant association between gender and the occurrence of strokes.

**Limitations**

The study will focus only on the identified risk factors and may not consider other potential risk factors that could impact the likelihood of experiencing a brain stroke. If the data included more details on the possible warning signs, such as sudden eyesight compromise, often loss of balance or coordination, and Sudden headache with no cause, the predictions would be more effective, and the analysis would be able to identify the relationship between the warning signs and other lifestyle and health factors. Another limitation is incomplete or missing data points in BMI and smoking status. Missing values in BMI and smoking status may affect the accuracy of the predictive models as the model may not accurately capture the impact of obesity and smoking risk on the likelihood of experiencing a brain stroke. To address this issue, effective imputation techniques should be used to estimate missing values for these variables. It is important to carefully consider the best imputation technique and evaluate the impact of missing values on the accuracy of the model.

**Ethical Considerations**

When dealing with data related to brain strokes, it is essential to carefully consider the ethical implications associated with utilizing sensitive health-related data. This encompasses concerns related to data security, confidentiality maintenance, and the need for informed consent. Safeguarding data privacy is highly important when engaging with health data, as it contains personal information that can be used to identify individuals. Researchers and healthcare professionals must ensure patients' privacy by including techniques like data anonymization, encryption, and limiting access to authorized personnel only.

In the field of healthcare, Electronic Health Records (EHRs) are created to improve the quality and safety of patient care. They give healthcare providers quick access to accurate and current patient information. EHRs also make it easier for different healthcare professionals to share patient data, which helps them work together better and lowers the chances of mistakes in medical treatment.

**Data Preprocessing**

The brain stroke database contains 201 null values for BMI. As part of preprocessing, we replace null BMI values with the mean of the non-null BMI values (Figure 2). After replacing the null BMI values, the code checks for null values in the data frame using the *isnull()* and *sum()* functions from the panda's library. This is a common step in data preprocessing to ensure that the data is clean and ready for analysis.

**Findings**

**Descriptive Analysis**

The describe() function (Figure 3) provides descriptive statistics for each column of the dataset. For the stroke\_data dataset, the age column has a mean of 43.23 years and a standard deviation of 22.61 years. The hypertension column has a mean of 0.097 and a standard deviation of 0.297, indicating that hypertension is a relatively uncommon condition in the dataset. The heart disease column has a mean of 0.054 and a standard deviation of 0.226, also indicating that heart disease is relatively uncommon. The avg\_glucose\_level column has a mean of 106.15 mg/dL and a standard deviation of 45.28 mg/dL, indicating that there is significant variability in glucose levels across the dataset. The bBMIcolumn has a mean of 28.89 kg/m^2 and a standard deviation of 7.70 kg/m^2 (Table 1).

**Table 1**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | age | hypertension | heart\_disease | avg\_glucose\_level | BMI | stroke |
| count | 5110 | 5110 | 5110 | 5110 | 5110 | 5110 |
| mean | 43.226614 | 0.097456 | 0.054012 | 106.147677 | 28.893237 | 0.048728 |
| std | 22.612647 | 0.296607 | 0.226063 | 45.28356 | 7.698018 | 0.21532 |
| min | 0.08 | 0 | 0 | 55.12 | 10.3 | 0 |
| 25% | 25 | 0 | 0 | 77.245 | 23.8 | 0 |
| 50% | 45 | 0 | 0 | 91.885 | 28.4 | 0 |
| 75% | 61 | 0 | 0 | 114.09 | 32.8 | 0 |
| max | 82 | 1 | 1 | 271.74 | 97.6 | 1 |

**Explanatory Analysis**

Figure 4 depicts a histogram that illustrates the distribution of age for patients who have had a stroke. The resulting histogram shows that the highest frequency of strokes occurs in patients who are in their 70s and 80s. The frequency of strokes is also relatively high for patients in their 60s and 50s, with a lower frequency for patients in their 40s and below. Overall, the histogram suggests that age is an important factor in the likelihood of having a stroke, with older patients being at higher risk.

Figure 4 also depicts a boxplot to test the hypothesis that patient age and stroke are related and differ based on gender. In this boxplot, patient ages are grouped by gender and compared between those who have had a stroke and those who have not. The results indicate that for both men and women, patients who have had a stroke tend to be older than those who have not. This is evident from the higher median age observed in the group of patients who have experienced a stroke compared to the group of patients who have not experienced a stroke. The visuals don't provide proof of a potential relationship between gender and stroke occurrence. The pie chart shows that most patients with stroke in the dataset are female (56.6%), while the remaining patients are male (43.4%) (Figure 5).

Figure 6 displays a boxplot that helps to compare the average glucose levels of two groups: patients who have had a stroke and those who have not. From the boxplot, it is evident that there is a notable difference in how glucose levels are distributed between these two groups. The median of glucose levels for the patients who have experienced stroke is comparatively greater than the median glucose level for those who have not. This finding strongly supports the alternative testing, stating there is a connection between the range of average glucose levels and the probability of experiencing a stroke.

The given boxplot in Figure 6 analyzes whether the patient's age and stroke are related and differ based on hypertension status. The output of the code shows that there is a difference in the distribution of patient ages between those who have had a stroke and those who have not had a stroke and that this difference varies based on hypertension status. This indicates that hypertension status may be an important factor in the relationship between age and stroke occurrence. However, further analysis is necessary to determine the significance of this difference and to establish whether hypertension status is a significant factor in the likelihood of stroke occurrence.

**Hypothesis Testin**g

The Python code in Figure 9 creates a contingency table of heart disease and stroke occurrence using the crosstab() function from Pandas and then calculates the chi-squared test statistic and p-value using the chi2\_contingency() function from SciPy. Here, the chi-squared test statistic is 90.25956, which indicates that there is a significant difference between the observed and expected frequencies in the contingency table. The p-value is 2.0887845685229236e-21, which is less than the commonly used significance level of 0.05. This means that there is strong evidence to reject the null hypothesis. So, it is evident that there is a correlation between heart disease and stroke occurrence, which is supported by the data.

In Figure 10, The t-test statistic is 5.998714431283543, which indicates that there is a significant difference between the average stroke occurrence in the two groups being compared. The p-value is 3.673335722473369e-09, which is less than the common significance level of 0.5. So, we can reject the null hypothesis and conclude that there is a higher likelihood of strokes in patients with hypertension than in those without.

The hypothesis being tested in Figure 11 is for the business question, "Are men more susceptible to experiencing a brain stroke?" The result of the chi-squared test is a test statistic of 0.47258662884530234 and a p-value of 0.7895490538408245. The p-value is greater than the commonly used significance level of 0.05, which means there is insufficient evidence to reject the null hypothesis. Therefore, we cannot conclude that gender has a significant impact on experiencing a brain stroke based on this test alone.

The code in Figure 12 creates two groups of patients based on their average glucose levels (high glucose group with patients whose average glucose level is above 140, and low glucose group with patients whose average glucose level is 140 or below) and then uses the t-test to compare the average stroke occurrence in each group. The result shows a p-value of 3.004392030233343e-1, which is less than 0.05. So, here, we can reject the null hypothesis and conclude that there is a correlation between the range of average glucose levels and the likelihood of stroke occurrence.

Table 2

|  |  |  |
| --- | --- | --- |
| Business Question | P-value | Hypothesis Test Result (Null Hypothesis Accept/Rejected) |
| Is the probability of having a stroke higher for individuals who have heart disease? | 2.09E-21 | Reject |
| Is there a higher likelihood of strokes in patients with hypertension than in those without? | 3.67E-09 | Reject |
| Are men more susceptible to experiencing a brain stroke? | 0.789549054 | Accept |
| What is the range of average glucose levels that correlates with a higher likelihood of stroke occurrence? | 3.00E-01 | Reject |

**Predictive Model Comparison**

The model comparison results (Figure 14) show that both the Random Forest and Logistic Regression models performed similarly, with the same accuracy score of 0.9393346379647749. This score was higher than the accuracy score of the Decision Tree model, which had a score of 0.9168297455968689. This suggests that either the Random Forest or Logistic Regression models may be a better choice for brain stroke prediction than the Decision Tree model based on the selected dataset.

* Random Forest accuracy: 0.9393346379647749
* Logistic Regression accuracy: 0.9393346379647749
* Decision Tree accuracy: 0.9168297455968689

The ROC curve plotted the true positive rate (y-axis) against the false positive rate (x-axis) for different threshold values (Figure 15). The area under the ROC curve (AUC) is also a commonly used metric for evaluating classifier performance. Generally, an AUC of 0.7 or higher is considered to be a good classifier, while an AUC of less than 0.5 indicates that the classifier is worse than random guessing. Here, the AUC of the Decision tree is near 0.6, which is greater than that of the Random Forest with an AUC of 0.51 and Logistic Regression with an AUC of 0.50.

There can be many reasons why the accuracy score is lower, and the area under the ROC curve is higher for the Decision Tree model compared to the other models. One possible explanation is that the Decision Tree model is better at distinguishing between positive and negative cases but is not as good at correctly classifying all instances. This means that the model may have a higher true positive rate (sensitivity) and a lower true negative rate (specificity), resulting in a higher AUC but a lower accuracy score.

**Conclusion**

The model comparison results suggest that the Random Forest and Logistic Regression models may be better choices for the particular problem of brain stroke prediction than the Decision Tree model. Both models achieved high accuracy scores, with the same score of 0.9393346379647749. Based on the model comparison results, it appears that both the Random Forest and Logistic Regression models achieved high accuracy scores, which were higher than the accuracy score of the Decision Tree model. However, it is important to consider multiple evaluation metrics and to interpret them in the context of the specific problem. Further analysis and experimentation may be needed to optimize the models and improve their performance. Overall, by implementing the recommendations outlined in this project, we can work towards improving the predictive models for brain stroke prediction and ultimately contribute to better patient outcomes and a more effective healthcare system.

**Recommendations**

Based on the findings of this project, it is recommended to further analyze the data, experiment with different predictive models and evaluation metrics, consider incorporating additional features or data sources, evaluate the models in a clinical setting, and continuously monitor and evaluate the models to ensure their accuracy and effectiveness over time. By implementing these recommendations, the predictive models for brain stroke prediction can be further optimized, leading to better patient outcomes and a more effective healthcare system.

**Screenshots**

**Figure (Screenshot) 1**

*Brain Stroke Dataset*

**A screenshot of a computer

Description automatically generated**

**Figure (Screenshot) 2**

*Data Preprocessing (Part 1)*

A screenshot of a computer

Description automatically generated

**Figure (Screenshot) 3**

*Data Preprocessing (part 2)*

A screenshot of a computer

Description automatically generated

**Figure (Screenshot) 4**

*Stroke Distribution by Age and Patient Age and Stroke Grouped by Gender*

A screenshot of a computer

Description automatically generated

**Figure (Screenshot) 5**

*Pie Chart to Show Gender Distribution of Patients with Stroke*

A screen shot of a computer

Description automatically generated

**Figure (Screenshot) 6**

*Boxplots Against Average Glucose Levels and Stroke and Patient age grouped by hypertension*

A screenshot of a computer

Description automatically generated

**Figure (Screenshot) 7**

*The pie chart shows the age group of patients with stroke.*

A screenshot of a computer

Description automatically generated

**Figure (Screenshot) 8**

*Data Preprocessing (part 2)*

A screenshot of a computer

Description automatically generated

**Figure (Screenshot) 9**

*Data Preprocessing (part 2)*

*A screenshot of a computer

Description automatically generated*

**Figure (Screenshot) 10**

*Data Preprocessing (part 2)*

*A screenshot of a computer

Description automatically generated*

**Figure (Screenshot) 11**

*Data Preprocessing (part 2)*

*A screenshot of a computer

Description automatically generated*

**Figure (Screenshot) 12**

*Data Preprocessing (part 2)*

*A screenshot of a computer

Description automatically generated*

**Figure (Screenshot) 13**

*Transforming the work\_type feature for building the predictive models*

*A screenshot of a computer

Description automatically generated*

**Figure (Screenshot) 14**

*Building and comparing predictive models using DT Classification, RF, and LR algorithms.*

*A screenshot of a computer

Description automatically generated*

**Figure (Screenshot) 15**

*Plotting the ROC curves for the model comparison.*

A screenshot of a computer

Description automatically generated

**References**

Colorado State University Global. (2022) MIS581-1. Module 6: Analytics Approach for Business Problems, *Critical Thinking, Canvas*. https://csuglobal.instructure.com/courses/80519/assignments/1615344

Emon, M. U., Keya, M. S., Meghla, T. I., Rahman, M. M., Al Mamun, M. S., & Kaiser, M. S.

(2020, November). Performance analysis of machine learning approaches in stroke

prediction. In 2020 4th International Conference on Electronics, Communication, and

Aerospace Technology (ICECA) (pp. 1464-1469).

<https://ieeexplore.ieee.org/abstract/document/9297525>

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic

(ROC) curves. Canadian Journal of Emergency Medicine, 8(1), 19-20.

<https://www.cambridge.org/core/journals/canadian-journal-of-emergency>

medicine/article/understanding-receiver-operating-characteristic-roc-curves/

Jansen, D., & Warren, K. (2020, June). Grad Coach. What is Research Methodology? A Plain

Language Explanation & Definition (With Examples).

<https://gradcoach.com/what-is-research-methodology/>

Kaur, M., Sakhare, S. R., Wanjale, K., & Akter, F. (2022). Early stroke prediction methods for

Prevention of strokes. Behavioural Neurology, 2022.

<https://www.hindawi.com/journals/bn/2022/7725597/>

Nagpal, A., & Gabrani, G. (2019, February). Python for data analytics, scientific and technical

applications. In 2019 Amity international conference on artificial intelligence

(AICAI) (pp. 140-145). IEEE.

https://ieeexplore.ieee.org/abstract/document/8701341

Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Monirujjaman Khan, M.

(2021). Stroke disease detection and prediction using robust learning approaches.

Journal of healthcare engineering, 2021.

<https://downloads.hindawi.com/journals/jhe/2021/7633381.pdf>

Singh, M. S., & Choudhary, P. (2017, August). Stroke prediction using artificial intelligence.

In 2017 8th Annual Industrial Automation and Electromechanical Engineering

Conference (IEMECON) (pp. 158-161). IEEE.

https://pdfs.semanticscholar.org/df5c/7d1bd7a59009dc51b9db903aa7f144241879.pdf

Sirsat, M. S., Fermé, E., & Camara, J. (2020). Machine learning for brain stroke: a

Review. Journal of Stroke and Cerebrovascular Diseases, 29(10), 105162.

<https://www.sciencedirect.com/science/article/abs/pii/S1052305720305802>

Swamynathan, M. (2017). Mastering machine learning with python in six steps: A practical

implementation guide to predictive data analytics using python. Manohar

Swamynathan.

<https://link.springer.com/book/10.1007/978-1-4842-2866-1>