Project Report

On

STROKE PREDICTION ANALYSIS

Course Number: CSP 571(Data Preparation & Analysis)

Instructor: **Oleksandr Narykov**

Team Members:

Ramchandra Reddy Sathu	**A20526126**
Venkata Madhu Vinay Nalamati	**A20526859**
Akash Chaturvedi Battula	**A20549559**
Vijay Goud Kodipyaka	**A20529999**

Department of Computer Science

ILLINOIS INSTITUTE OF TECHNOLOGY

STROKE DATASET

❖ *Explanation of Dataset:*

The Stroke Prediction Dataset comprises 5,110 records, each representing an individual patient. It includes 11 clinical features pertinent to stroke prediction, such as age, gender, hypertension status, heart disease history, and lifestyle factors like smoking status.

The Columns in the Stroke dataset are:

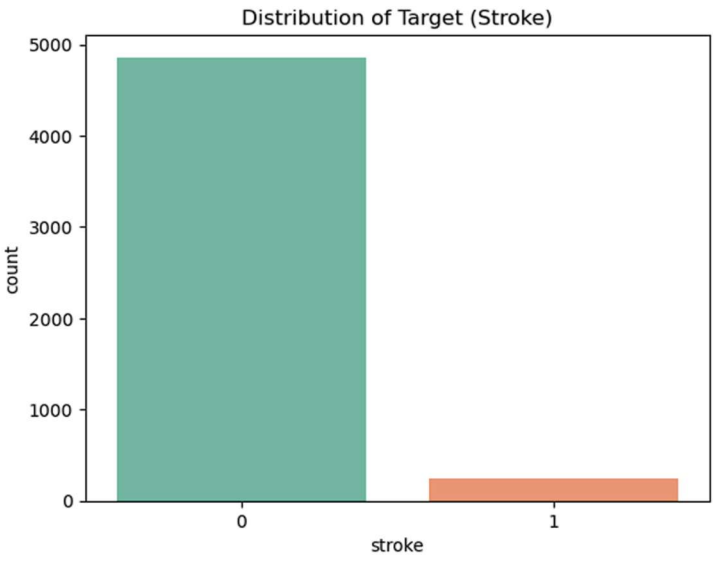
- Gender, Age, Hypertension, Heart disease, ever married, Work type, Residence type, Avg glucose level, bmi, smoking status, stroke:

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...
5105	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 11 columns

❖ *Distribution of Target:*

From the above dataset we find the distribution of target by using the visualization of bar graph. The bulk of the samples fall into the "No Stroke" group (denoted as '0'), with a count above 4,500, whereas the "Stroke" category (denoted as '1') is much smaller, with less than 500 occurrences



MACHINE LEARNING PROBLEM

The objective is to develop a predictive model capable of effectively predicting stroke occurrences. The proposed solution involves data exploration, preprocessing to address missing values and potential class imbalance, selection and training of a suitable classification model, and evaluation based on metrics like precision, recall, and F1-score. The goal is to create a robust and generalizable model that enhances stroke prediction accuracy based on clinical and demographic features.

LIBRARIES USED

- pandas: For data manipulation and analysis (e.g., handling missing values, filtering, and grouping data).
- numpy: For numerical operations and efficient computation on arrays.
- matplotlib and seaborn: For data visualization to explore patterns, trends, and distributions.
- scikit-learn: Provides tools for model selection, training, and evaluation. Includes algorithms like logistic regression, decision trees, random forests, and SVM.

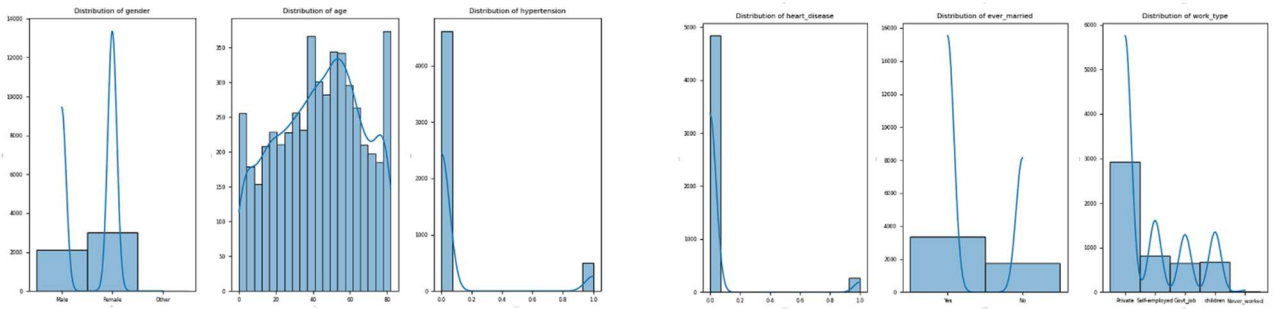
DATA PRE-PROCESSING

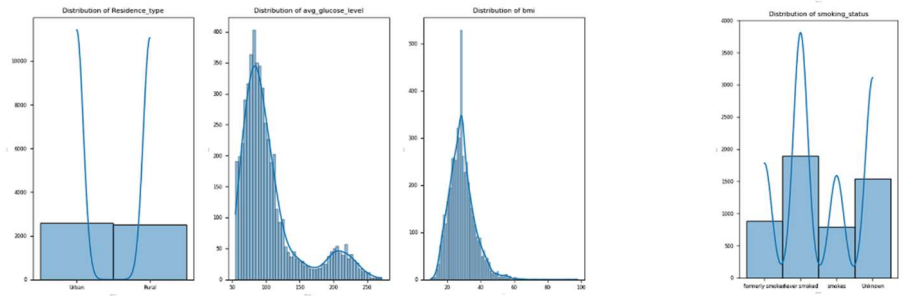
- Used SimpleImputer to fill missing bmi values with the mean.
- Dropped the id column as it is irrelevant for prediction.
- Converted categorical features into numerical formats and Standardized numerical features for uniformity.
- Addressed imbalance using techniques like SMOTE during training.

DATA DISTRIBUTION

❖ *Distribution of columns:*

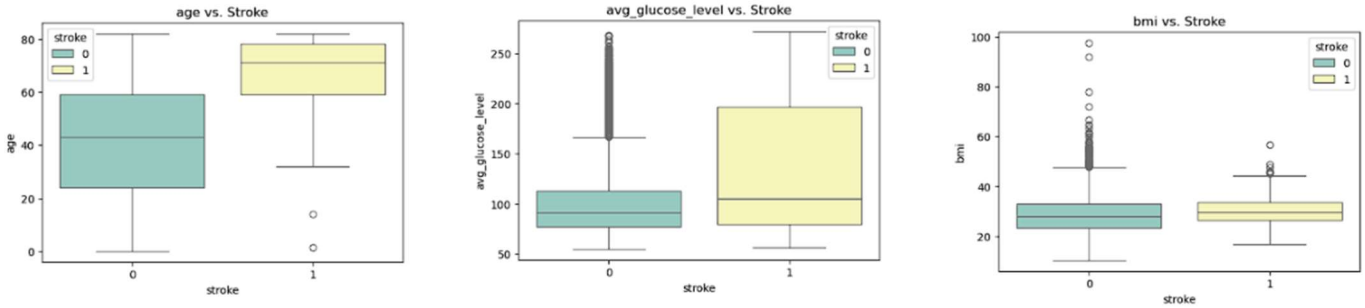
The visualizations depict the distributions of several variables in the dataset. The gender distribution is unequal, with "Male" and "Female" being far more prevalent than "Other." According to the age distribution, most people are in their middle years and do not have hypertension or heart disease. A greater number of people are or have been married, and the majority work in the private sector, with "Never worked" being the least common group. The residential categories are evenly split between "urban" and "rural." Average glucose levels and BMI exhibit center clustering, with occasional outliers at higher levels. Finally, smoking status shows that "never smoked" is the most common, followed by "unknown." These maps show an overview of the dataset's demographic, health, and lifestyle factors.





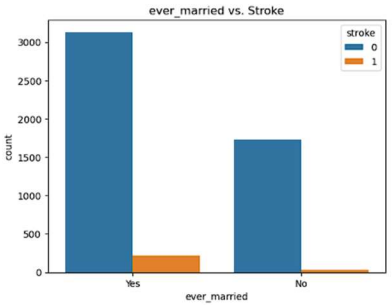
❖ *Numerical Features:*

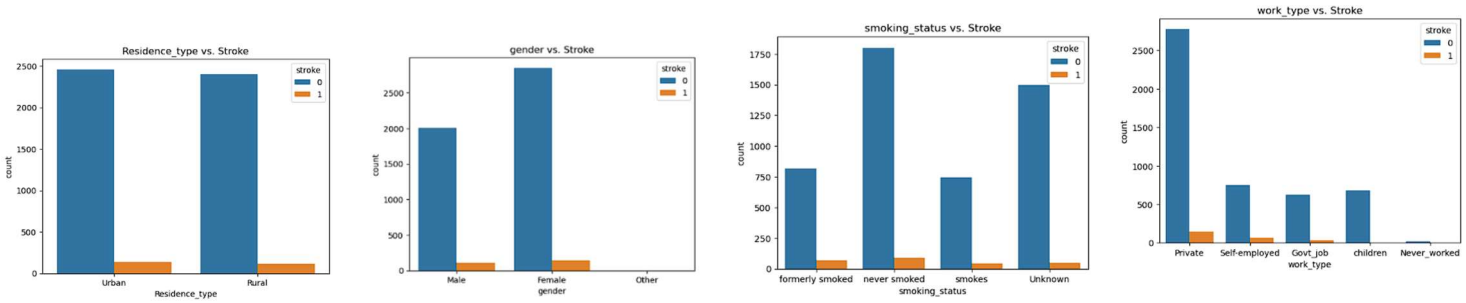
The boxplots show the link between age, average glucose level, and BMI with the risk of stroke. Individuals who have had a stroke are often older, having a greater and narrower median age range than those who have not had one. Average glucose levels are much higher in stroke patients, indicating a possible link between high glucose levels and stroke incidence. BMI distributions differ little between the two groups, with similar medians and a wide range of outliers in both categories. Overall, age and glucose levels appear to be more strongly related to stroke risk than BMI.



❖ *Categorical Features:*

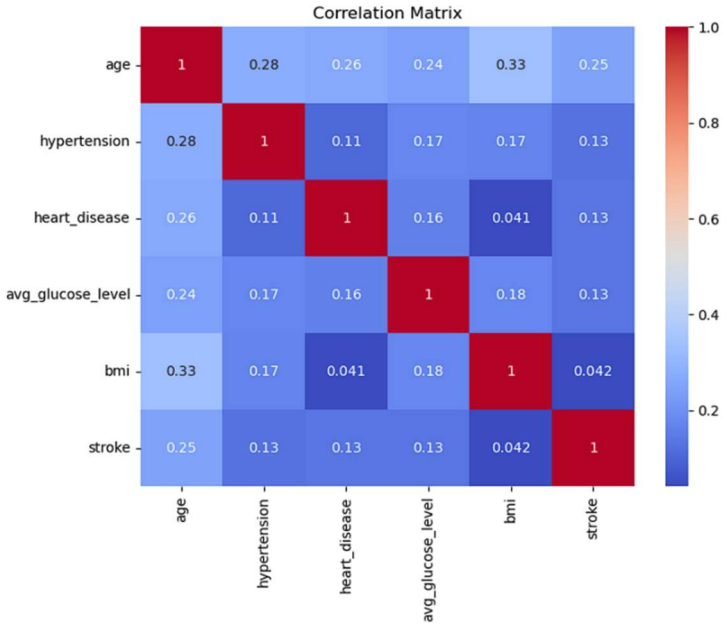
The bar charts show the link between different category characteristics and stroke incidences. Stroke cases are somewhat more common in females, whereas the "Other" gender category has a low proportion. Married people had a greater stroke rate, which may correspond with age. Most strokes occur among those who work in private or self-employment, whereas "children" and "never worked" groups have extremely few incidents. The distribution of strokes in rural and urban regions is equal, showing no substantial connection. Smoking status is associated with greater stroke rates in "never smoked" and "formerly smoked" groups, while the huge "unknown" category confuses interpretation. These trends indicate that marital status and employment type may be more strongly associated with stroke incidence than gender, residence, or smoking status.





❖ *Correlation:*

The correlation matrix shows the strength of correlations between different variables in the dataset. Age has a moderate positive connection with BMI (0.33) and a weak link with stroke (0.25), implying that older people are somewhat more likely to have strokes and have higher BMI levels. Hypertension has modest relationships with stroke (0.13), age (0.28), and average_glucose_level (0.17), indicating that it may be a slight factor to stroke risk. Heart disease has a slight correlation with stroke (0.13) and age (0.26), with little connection with other factors. The average glucose level shows a modest positive connection with stroke (0.13) and hypertension (0.17), indicating that those with these diseases had slightly higher glucose levels. BMI has a low connection with stroke (0.042), but a somewhat high link with age (0.33). Stroke had the greatest relationships with age (0.25), hypertension (0.13), and average glucose level (0.13), while these correlations are still minor. The low correlations between most variables and stroke indicate that stroke risk is controlled by several factors rather than a single element. Furthermore, BMI and heart disease have practically no significant direct relationship with stroke. Overall, the findings suggest a complicated interaction of factors influencing stroke risk, with age being the most significant of the identified variables.



❖ *Chi-Square Analysis:*

The results of the Chi-Square test show how categorical features and the target variable are related. With a p-value of 0.7895 and a relatively low Chi-Square statistic of 0.47, gender does not appear to have a statistically significant link with the objective variable. Conversely, a high Chi-Square score (58.92) and a statistically significant p-value (0.0000) show a substantial correlation between ever married and the objective variable. Work type also exhibits a strong correlation, with a p-value of 0.0000 and a Chi-Square statistic of 49.16. With a p-value of 0.2983 and a poor Chi-Square statistic of 1.08, residence type does not appear to be significantly associated. With a significant p-value (0.0000) and a Chi-Square statistic of 29.15, smoking status shows a moderate connection.

```
Feature: gender
Chi-Square Statistic: 0.47
P-Value: 0.7895
Degrees of Freedom: 2
```

```
Feature: ever_married
Chi-Square Statistic: 58.92
P-Value: 0.0000
Degrees of Freedom: 1
```

```
Feature: work_type
Chi-Square Statistic: 49.16
P-Value: 0.0000
Degrees of Freedom: 4
```

```
Feature: Residence_type
Chi-Square Statistic: 1.08
P-Value: 0.2983
Degrees of Freedom: 1
```

```
Feature: smoking_status
Chi-Square Statistic: 29.15
P-Value: 0.0000
Degrees of Freedom: 3
```

ONE-HOT ENCODING

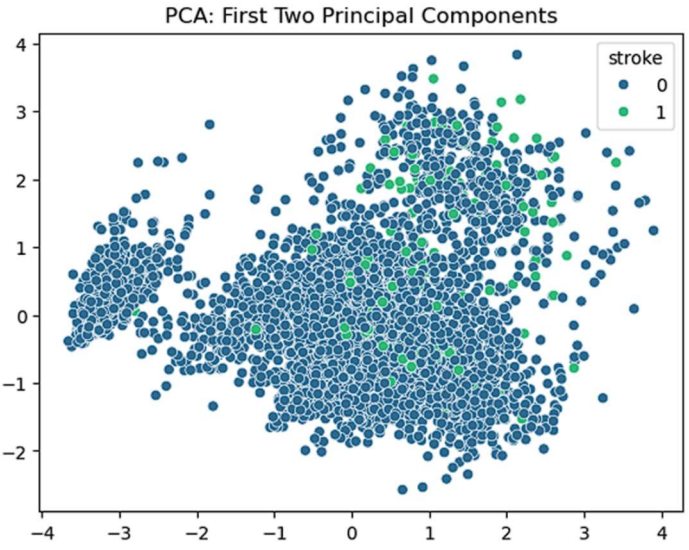
A dataset with a combination of encoded category and numerical attributes is represented by the table. The continuous numerical variables of age, BMI, and average blood sugar are probably scaled or normalized. Each value in the numerical encoding of categorical features—such as smoking status, ever married, work type, residence type, and gender—represents a distinct category. Heart disease and hypertension are binary variables that show whether they exist or not. Although numerical representation is compact in the current encoding, it may suggest an ordinal relationship between categories. There would be no ordinal assumption if one-hot encoding were employed, since distinct binary columns would be utilized to represent each category. This method treats each category separately, making it ideal for models that are sensitive to numerical values for compactness. Numerical encoding works well, but in most situations, one-hot encoding guarantees superior handling of categorical data. Below is the head of the dataset after applying one-hot encoding.

	age	avg_glucose_level	bmi	smoking_status	ever_married	work_type	Residence_type	gender	hypertension	heart_disease
0	1.051434	2.706375	1.001234e+00	1	1	2	1	1	0	1
1	0.786070	2.121559	4.615554e-16	2	1	3	0	0	0	0
2	1.626390	-0.005028	4.685773e-01	2	1	2	0	1	0	1
3	0.255342	1.437358	7.154182e-01	3	1	2	1	0	0	0
4	1.582163	1.501184	-6.357112e-01	2	1	3	0	0	1	0

DIMENSIONALITY REDUCTION APPROACHES

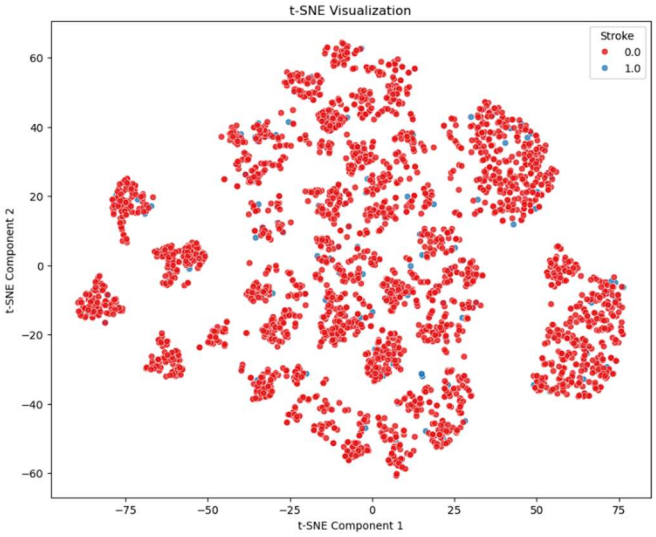
- ***PCA:***

The most significant variance in the data is captured by the PCA plot, which visualizes the dataset reduced to two principle components. While the "stroke" category is dispersed and overlaps with the main cluster, the bulk of data points fall into the "no stroke" category, which forms a dense cluster. The overlap between the stroke and no-stroke categories suggests that, in this limited dimensional space, the variables in the dataset might not be able to clearly differentiate between the two groups. This implies that there can be intricate relationships in the dataset that are unable to properly capture in two dimensions. The inability to distinguish clearly indicates that further feature engineering or more sophisticated modelling approaches are required to enhance classification performance. Although PCA does not specifically optimize for class separation, it is useful for visualizing high-dimensional data. Although this map offers a helpful summary of the data structure, it might not accurately depict its predictive capabilities.



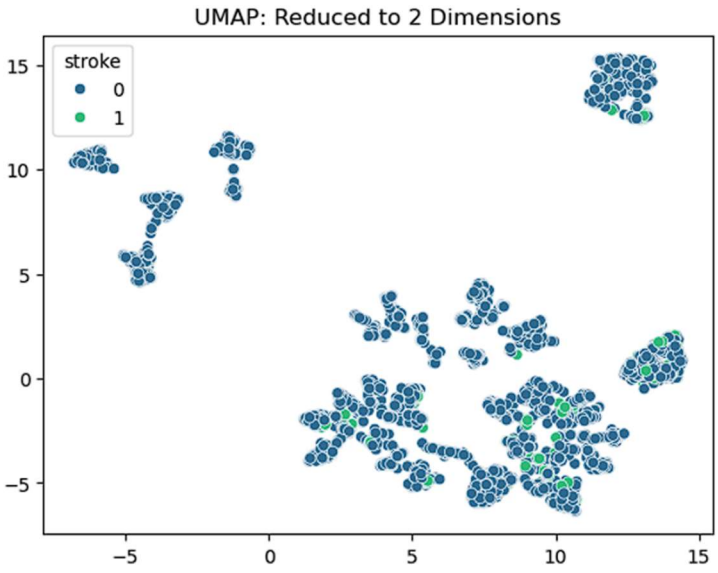
❖ ***UMAP:***

Clusters that indicate similarities in the high-dimensional data are displayed in the UMAP plot, which visualizes the dataset in two dimensions. Every point represents a distinct person; green indicates stroke instances, while blue indicates no stroke. Stroke cases are unevenly dispersed throughout clusters with little discernible separation, despite the clusters suggesting some underlying structure in the data. Potential areas of attention are indicated by certain clusters that have somewhat greater frequencies of stroke cases. It appears difficult to differentiate between stroke and no-stroke patients in this area due to the overlap between the two categories. While UMAP does not optimize for class separation, it performs a good job of grouping related data points for exploratory research. This graphic makes clear that to enhance classification performance, more features or sophisticated modelling approaches are required.



❖ ***t-SNE:***

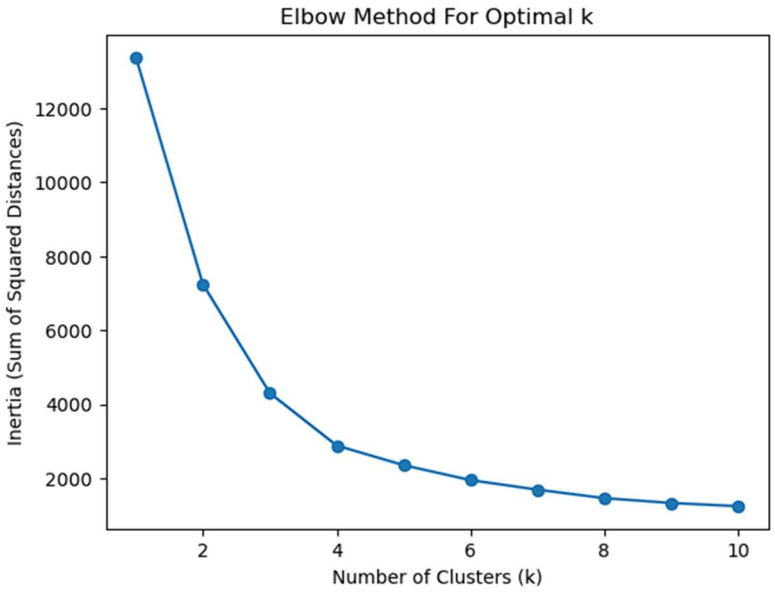
Individuals are shown as points in the t-SNE visualization, which projects high-dimensional data into two dimensions. Blue indicates stroke instances, whereas red indicates no stroke. Stroke cases are dispersed throughout these clusters without creating discrete groups, even though the data forms many clusters that reflect underlying classifications in the high-dimensional space. Due to the sparse distribution of blue dots among red clusters, it is difficult to distinguish between stroke and no-stroke cases. This implies that the two classes in the current representation might not be significantly distinguished by the dataset's properties. Global separability between classes is not given priority by t-SNE, despite its effectiveness in visualizing local structures and patterns. The dispersed stroke situations demonstrate the necessity of further feature engineering or sophisticated modelling methods.



All things considered, the plot provides information about the structure of the dataset, but it also highlights difficulties in differentiating stroke instances using the available attributes.

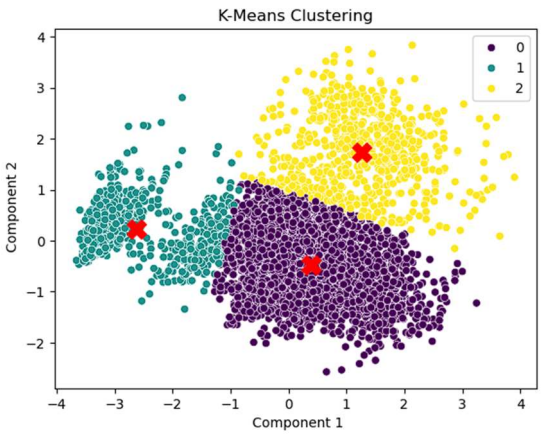
❖ ***Elbow Method:***

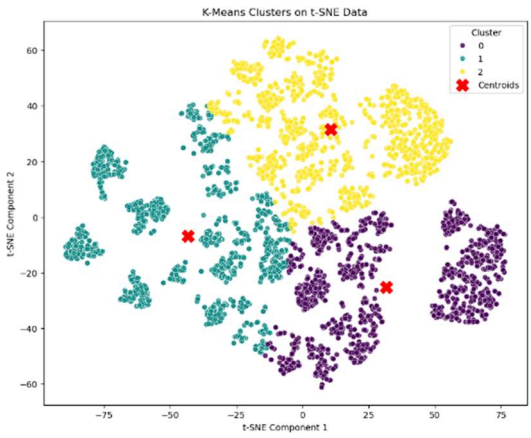
When employing clustering techniques such as k-means, the Elbow Method plot assists in identifying the ideal number of clusters (k) for a dataset. The y-axis displays inertia, which is the total of squared distances between data points and their cluster centroids, while the x-axis displays the number of clusters. As more clusters are added, the curve's initial rapid decrease in inertia slows down and forms a "elbow." The elbow, located at k = 3 or k = 4, shows the point at which clustering performance is no longer appreciably enhanced by the addition of clusters. This implies that three or four clusters strike a compromise between performance and simplicity, making them the best option for the dataset.



❖ ***k-means clustering:***

K-Means clustering applied to the data in two different representations is depicted in the two plots. With centroids (red stars) precisely positioned at the cluster centers, the left plot, which displays clustering on lower dimensions (such as PCA), displays compact and well-separated clusters. Because t-SNE focusses on maintaining local data structures rather than global separability, clusters in the right plot, which uses t-SNE components, appear less distinct. Centroids have significance in the PCA space, but in the nonlinear t-SNE space, they are harder to understand. Complementary perspectives of the clustering results are provided by the left plot, which emphasizes global structure and compactness, and the right plot, which highlights local patterns.





SMOTE ANALYSIS

The picture shows how to use SMOTE to handle unbalanced data and how to use cross-validation to assess three models (Logistic Regression, Random Forest, and XGBoost). By oversampling the minority class, SMOTE balances the dataset and enhances model performance on data that is under-represented. While Logistic Regression has the highest average ROC-AUC (83.49%) and the lowest accuracy (73.85%), Random Forest and XGBoost both attain high average accuracies (92.6%), suggesting superior discriminatory power. Overall classification performance is superior by Random Forest and XGBoost, however imbalanced predictions might be handled more effectively by Logistic Regression. Whether accuracy or class separation (ROC-AUC) is given priority determines which model is used.

```
Training data shape Before SMOTE
y_train: stroke
0    3889
1     199
Name: count, dtype: int64
Training data shape after SMOTE:
y_train_smote: stroke
0    3889
1    3889
Name: count, dtype: int64
```

STRATIFIED K-FOLD CROSS VALIDATION

Stratified K-Fold Cross-Validation was employed to evaluate the model's performance, ensuring fair and robust assessment. This technique splits the dataset into 5 folds, where each fold preserves the class distribution of the target variable, addressing the challenge of imbalanced datasets. This approach provides a balanced representation of all classes in both training and validation sets, reducing bias and variance in performance metrics. By offering a reliable and fair evaluation, Stratified K-Fold enhances the model's generalizability and ensures its robustness across different subsets of the data.

```
Logistic Regression - Cross-Validation Scores:
  Fold 1: Accuracy = 0.7225, ROC-AUC = 0.8189
  Fold 2: Accuracy = 0.7543, ROC-AUC = 0.8509
  Fold 3: Accuracy = 0.7286, ROC-AUC = 0.8268
  Fold 4: Accuracy = 0.7356, ROC-AUC = 0.8296
  Fold 5: Accuracy = 0.7515, ROC-AUC = 0.8483
Logistic Regression - Average Accuracy: 0.7385
Logistic Regression - Average ROC-AUC: 0.8349
```

```
Random Forest - Cross-Validation Scores:
  Fold 1: Accuracy = 0.9156, ROC-AUC = 0.7972
  Fold 2: Accuracy = 0.9340, ROC-AUC = 0.8230
  Fold 3: Accuracy = 0.9181, ROC-AUC = 0.8065
  Fold 4: Accuracy = 0.9241, ROC-AUC = 0.7844
  Fold 5: Accuracy = 0.9412, ROC-AUC = 0.7947
Random Forest - Average Accuracy: 0.9266
Random Forest - Average ROC-AUC: 0.8012
```

```
XGBoost - Cross-Validation Scores:
  Fold 1: Accuracy = 0.9193, ROC-AUC = 0.7739
  Fold 2: Accuracy = 0.9364, ROC-AUC = 0.8184
  Fold 3: Accuracy = 0.9193, ROC-AUC = 0.7891
  Fold 4: Accuracy = 0.9302, ROC-AUC = 0.8056
  Fold 5: Accuracy = 0.9290, ROC-AUC = 0.7749
XGBoost - Average Accuracy: 0.9269
XGBoost - Average ROC-AUC: 0.7924
```

UNSUPERVISED LEARNING TECHNIQUES

- ***Logistic Regression:***

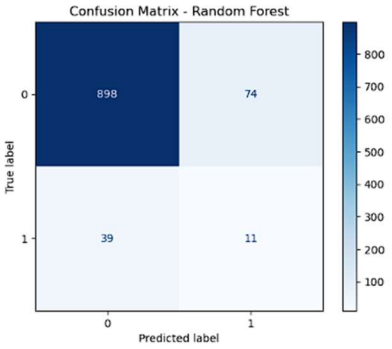
The majority class (no stroke) performs well according to the evaluation of the logistic regression model; however the minority class (stroke) is difficult to forecast. The model struggles with class imbalance, achieving high precision (0.99) and F1-score (0.86) for Stroke but low precision (0.14) and F1-score (0.24) for No Stroke. The imbalance problem is shown by the confusion matrix, which displays many false positives for stroke and a few false negatives. The ROC-AUC value of 0.82, however, suggests strong overall discriminatory ability. Enhancements like threshold tweaking or resampling could improve the minority class's performance.

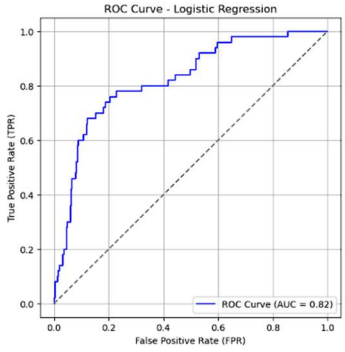
```
Logistic Regression Report:
              precision    recall  f1-score   support

           0       0.99      0.76      0.86       972
           1       0.14      0.78      0.24        50

    accuracy                           0.76      1022
   macro avg       0.56      0.77      0.55      1022
weighted avg       0.94      0.76      0.83      1022

ROC-AUC Score: 0.8234567901234568
```





- ***Random Forest:***

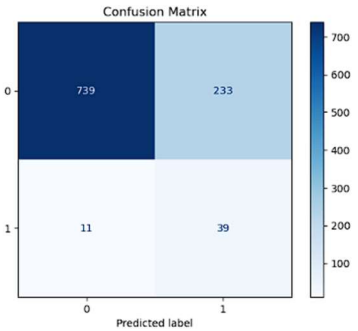
With an overall accuracy of 89%, the Random Forest model performs well for the majority class (No Stroke), exhibiting good precision (0.96), recall (0.92), and an F1-score (0.94). Its poor precision (0.13), recall (0.22), and F1-score (0.16) for the minority class (Stroke) result in a high number of false negatives and false positives. The model's bias towards the majority class and inability to accurately forecast strokes are highlighted by the confusion matrix. The model must be improved, such as by resampling or modifying class weights, in order to better handle the class imbalance, even though the ROC-AUC score of 0.79 indicates modest discriminatory capacity. All things considered; the model works well for the majority class but falls short when it comes to detecting strokes.

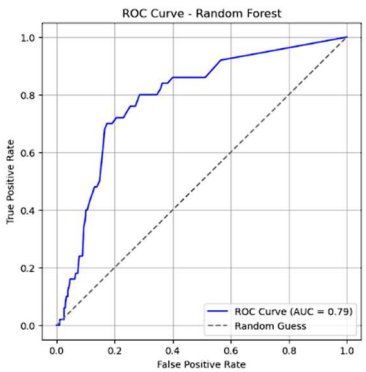
```
Random Forest Report:
              precision    recall  f1-score   support

           0       0.96      0.92      0.94       972
           1       0.13      0.22      0.16        50

    accuracy                           0.89      1022
   macro avg       0.54      0.57      0.55      1022
weighted avg       0.92      0.89      0.90      1022

Random Forest ROC-AUC Score: 0.7881893004115226
```





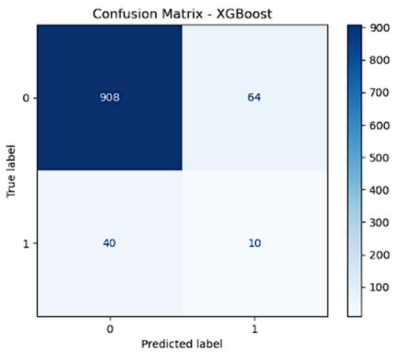
❖ ***XGBOOST:***

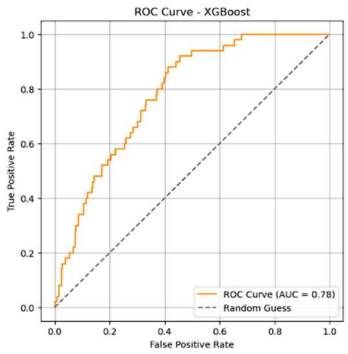
For the majority class (No Stroke), the XGBoost model performs well, achieving high accuracy (90%) and good precision (0.96), recall (0.93), and an F1-score (0.95). It has trouble detecting stroke patients, though, as evidenced by its poor precision (0.14), recall (0.20), and F1-score (0.16) for the minority class (Stroke). This imbalance is reflected in the confusion matrix, which has many false positives and false negatives for the minority class. The moderate overall discriminatory capacity but inadequate performance for stroke detection is indicated by the ROC-AUC score of 0.78. To improve minority class predictions, class imbalance must be addressed through tuning, class weighting, or resampling.

XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.96	0.93	0.95	972
1	0.14	0.20	0.16	50
accuracy			0.90	1022
macro avg	0.55	0.57	0.55	1022
weighted avg	0.92	0.90	0.91	1022

ROC-AUC Score: 0.78





HYPER PARAMETER TUNING

❖ ***Hyperparameter Tuning and early stopping:***

With an accuracy of 93% and enhanced minority class metrics (precision, recall, and F1-score), XGBoost had the highest overall performance following hyperparameter adjustment, with a balanced ROC-AUC of 0.82. Despite having a higher recall, logistic regression suffered with low accuracy for the minority class, despite having a high ROC-AUC (0.82). Random Forest scored poorly in the minority class, with the lowest ROC-AUC (0.79), while it had high metrics and good accuracy (89%) for the majority class. While Logistic Regression was still a solid candidate for balanced discrimination, XGBoost was much enhanced by hyperparameter adjustment, making it the strongest model for managing the minority class. However, Random Forest has to be further refined in order to forecast minority classes more accurately.

Best Logistic Regression Report:

	precision	recall	f1-score	support
0	0.99	0.76	0.86	972
1	0.14	0.78	0.24	50
accuracy			0.76	1022
macro avg	0.56	0.77	0.55	1022
weighted avg	0.94	0.76	0.83	1022

Best ROC-AUC Score: 0.82

Best Random Forest Report:

	precision	recall	f1-score	support
0	0.96	0.92	0.94	972
1	0.13	0.22	0.16	50
accuracy			0.89	1022
macro avg	0.54	0.57	0.55	1022
weighted avg	0.92	0.89	0.90	1022

Best Random Forest ROC-AUC Score: 0.79

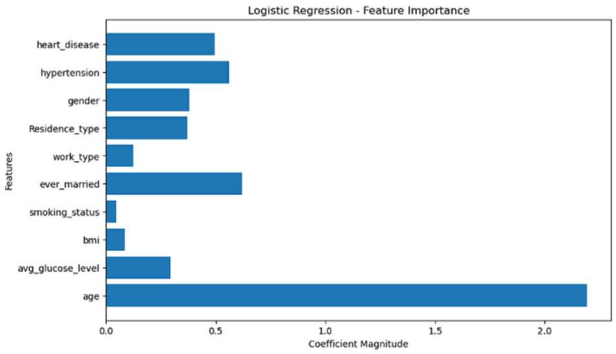
XGBoost Classification Report:

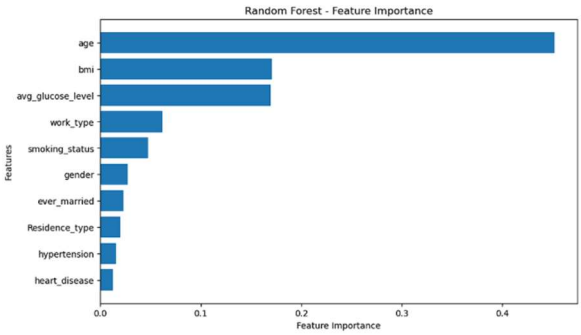
	precision	recall	f1-score	support
0	0.96	0.96	0.96	972
1	0.31	0.32	0.31	50
accuracy			0.93	1022
macro avg	0.64	0.64	0.64	1022
weighted avg	0.93	0.93	0.93	1022

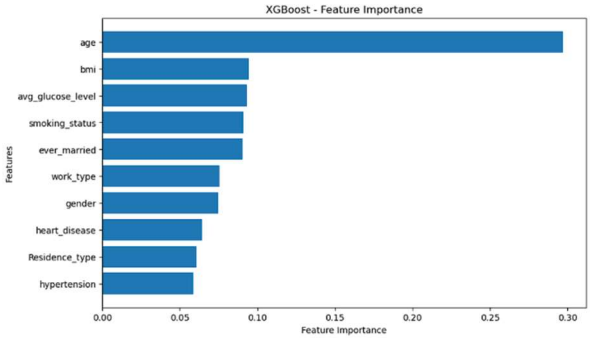
ROC-AUC Score: 0.82

❖ ***Feature Importance:***

The most significant predictor is consistent age, according to the feature importance rankings for Random Forest, XGBoost, and Logistic Regression. Age, avg_glucose_level, and ever_married are important characteristics in logistic regression, which prioritizes linear connections; residence_type and heart_disease have less bearing. Age is ranked as the most important feature by Random Forest and XGBoost, followed by BMI and average glucose level. Other features such as smoking status and work type are ranked as moderately important. Features like residence_type, heart_disease, and hypertension are consistently less significant across all models. While tree-based models offer more detail for secondary predictors like BMI and avg_glucose_level, these findings support the idea that age plays a dominant role in predictions.







RECURSIVE FEATURE ELIMINATION AND REGULARIZATION

The performance of XGBoost with Regularization and Logistic Regression (RFE with Regularization) on unbalanced data is compared. With a robust ROC-AUC of 0.82, both Logistic Regression techniques produce comparable results. The majority class performs well, whereas the minority class has poor precision. While XGBoost with Regularization has a slightly lower overall accuracy (0.89), it improves minority class recall (0.28) when compared to Random Forest and Logistic Regression. Random Forest has the lowest ROC-AUC (0.79) and suffers with minority class predictions, although achieving good accuracy (0.90). While XGBoost is better for minority class sensitivity, regularization guarantees better generalization for both Logistic Regression and XGBoost without appreciable metric changes.

RFE Logistic Regression Report:

	precision	recall	f1-score	support
0	0.99	0.76	0.86	972
1	0.14	0.78	0.24	50
accuracy			0.76	1022
macro avg	0.56	0.77	0.55	1022
weighted avg	0.94	0.76	0.83	1022

ROC-AUC Score: 0.82

Logistic Regression with Regularization Report:

	precision	recall	f1-score	support
0	0.99	0.76	0.86	972
1	0.14	0.78	0.24	50
accuracy			0.76	1022
macro avg	0.56	0.77	0.55	1022
weighted avg	0.94	0.76	0.83	1022

Best ROC-AUC Score: 0.82

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.94	0.95	972
1	0.11	0.14	0.12	50
accuracy			0.90	1022
macro avg	0.53	0.54	0.54	1022
weighted avg	0.91	0.90	0.91	1022

ROC-AUC Score: 0.79

XGBoost with Regularization Report:

	precision	recall	f1-score	support
0	0.96	0.92	0.94	972
1	0.15	0.28	0.19	50
accuracy			0.89	1022
macro avg	0.55	0.60	0.57	1022
weighted avg	0.92	0.89	0.90	1022

Best ROC-AUC Score: 0.78

NEURAL NETWORKS

With excellent precision (0.97), recall (0.84), and an F1-score (0.90), the neural network model performs well for the majority class (no stroke), resulting in an overall accuracy of 83%. Although the model's recall of 0.54 indicates a moderate sensitivity in identifying stroke cases, it struggles with poor precision (0.15) and an F1-score (0.23) for the minority class (Stroke). Despite the disparity in performance, the ROC-AUC score of 0.80 suggests decent overall discriminating between classes. The model works well for the majority class, but because of its high false positive rate, it incorrectly classifies many minority cases. Its sensitivity and accuracy for minority class predictions could be increased with improvements like resampling or sophisticated structures.

Neural Network Report:

	precision	recall	f1-score	support
0	0.97	0.84	0.90	972
1	0.15	0.54	0.23	50
accuracy			0.83	1022
macro avg	0.56	0.69	0.57	1022
weighted avg	0.93	0.83	0.87	1022

ROC-AUC Score: 0.80

CONCLUSION

To forecast the occurrence of strokes in an unbalanced dataset, this study assessed many machine learning models, such as Neural Networks, XGBoost, Random Forest, and Logistic Regression. Age, BMI, and avg_glucose_level were found to be the most significant predictors across models by feature importance analysis. Models like Random Forest and XGBoost performed well on the majority class but poorly on the minority class (Stroke). XGBoost demonstrated the best balance between accuracy and recall. Although they improved outcomes, strategies like SMOTE and hyperparameter tuning did not completely address the issues of class imbalance. To improve minority class prediction without compromising overall performance, future work should concentrate on sophisticated resampling strategies, threshold modifications, and model architecture optimization.

GITHUB AND DATASET LINKS

Github link: https://github.com/RamchandraReddy07/CSP_571_Project

Dataset link: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data

REFERENCES

[1] H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5128907&isnumber=5173046

[2] D. De Roure, Y. Gil and J. A. Hendler, "Guest editors' introduction: E-Science," in IEEE Intelligent Systems, vol. 19, no. 1, pp. 24-25, Jan.-Feb. 2004, doi: 10.1109/MIS.2004.1265881.

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1265881&isnumber=28315

[3] A.Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," in IEEE Access, vol. 9, pp. 39707-39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9370099&isnumber=9312710

[4] M. Khushi et al., "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," in IEEE Access, vol. 9, pp. 109960-109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9505667&isnumber=9312710