Stroke Prediction Using Data Mining

Aims to build a predictive model using machine learning algorithms to classify whether a person is likely to suffer a stroke based on healthcare-related features.

CPS 844 - Data Mining

Prof. Cherie Ding

Group 100

Hetu Virajkumar Patel - 501215707

Nilay Thakorbhai Patel - 501215918

INTRODUCTION

In recent years, the increasing availability of healthcare data has provided significant opportunities for data mining and machine learning to contribute to improved medical decision-making and preventive care. Stroke, one of the leading causes of death and disability worldwide, is a life-threatening condition that requires prompt diagnosis and intervention. Early prediction and prevention play a pivotal role in reducing its fatality and long-term impacts.

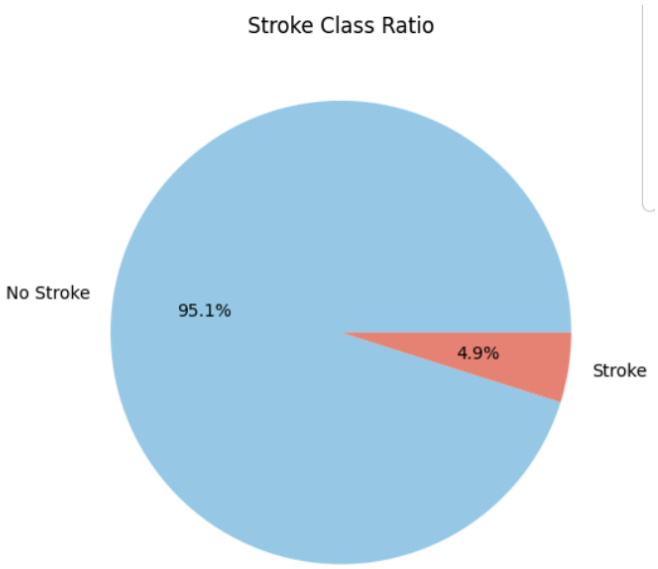
This report outlines a comprehensive data mining approach to the "**Stroke Prediction Dataset**" sourced from **Kaggle**. The dataset contains health-related records and a binary label indicating whether a person has had a stroke. The objective is to develop predictive models that can accurately identify individuals at risk of stroke based on their demographic and medical attributes, framing this as a binary classification problem. The methods include data preprocessing, exploratory data analysis (EDA), and the implementation of five machine learning models: Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Additionally, the use of Synthetic Minority Over-sampling Technique (SMOTE) and Recursive Feature Elimination (RFE) is explored. A stacked ensemble model is also built to assess improvements in performance by combining multiple base learners.

DATASET DESCRIPTION

The Stroke Prediction Dataset, sourced from Kaggle, contains 5,110 instances and 12 features, including 11 input variables and 1 target variable. The target variable, "stroke," is binary and indicates whether an individual has had a stroke (1) or not (0). The features include a mix of categorical and numerical attributes:

- **Categorical Features:** Gender, Ever Married, Work Type, Residence Type, Smoking Status
- **Numerical Features:** Age, Hypertension, Heart Disease, Avg Glucose Level, BMI

There are missing values in the BMI column, which contain 201 missing entries. This dataset is ideal for predictive modeling based on demographic and health-related factors and is instrumental for developing machine learning models aimed at stroke prediction.



Exploratory Data Analysis and Preprocessing

The initial data exploration involved identifying missing values, checking for imbalanced classes, and understanding the feature distributions. The missing values in the BMI column were imputed with the mean value to avoid data loss. Label encoding was used for categorical variables such as gender, marital status, and smoking status, converting them to numeric values suitable for machine learning algorithms.

To address the class imbalance, the **Synthetic Minority Over-sampling Technique (SMOTE)** was applied, generating synthetic samples for the **minority class (stroke)**. Additionally, the "id" column, which was not useful for prediction, was dropped, and feature scaling using StandardScaler was performed to normalize the data. This ensured that all features contributed equally to the model, improving performance.

Problem Definition

The task at hand is to predict the likelihood of an individual having a stroke based on demographic and medical features. The problem is a **binary classification**, where the target variable indicates **stroke** occurrence (1) or absence (0). Given the class imbalance, with more non-stroke cases, special attention was given to ensuring the model performed well for both classes. Evaluation metrics such as **accuracy, F1-Score, and ROC AUC** were used to assess model performance.

MODEL EVALUATION

Logistic Regression [LR]:

Logistic Regression is a classification algorithm that predicts the probability of a binary outcome, in this case, whether an individual will experience a stroke (1) or not (0) based on features like age, gender, hypertension, heart disease, BMI, and smoking habits. The model is trained by minimizing the **logistic loss** (also called **log-loss** or **binary cross-entropy**) over the dataset.

Logistic Regression Model Using all Features And Recursive Feature Elimination (RFE):

```
--- Logistic Regression ---
Accuracy: 0.7881748071979434
F1 Score: 0.7964426877470355
ROC AUC: 0.848913560666138
Confusion Matrix:
 [[727 248]
 [164 806]]
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.75      0.78       975
           1       0.76      0.83      0.80       970

    accuracy                           0.79      1945
   macro avg       0.79      0.79      0.79      1945
weighted avg       0.79      0.79      0.79      1945
```

```
--- Logistic Regression (RFE) ---
Accuracy: 0.7763496143958869
F1 Score: 0.782608695652174
ROC AUC: 0.8491821305841925
Confusion Matrix:
 [[727 248]
 [187 783]]
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.75      0.77       975
           1       0.76      0.81      0.78       970

    accuracy                           0.78      1945
   macro avg       0.78      0.78      0.78      1945
weighted avg       0.78      0.78      0.78      1945
```

The **Logistic Regression** model achieved an **accuracy of 0.79**, an **F1 score of 0.80**, and a **ROC AUC of 0.85**, with a **confusion matrix** showing that it correctly predicted **727 non-stroke** and **806 stroke** cases. The model's **precision**, **recall**, and **F1 score** for the **no-stroke class (0)** were **0.82**, **0.75**, and **0.78**, while for the **stroke class (1)**, they were **0.76**, **0.83**, and **0.80**. These results indicate that the model performs well in predicting both classes, but with class imbalance issues, further improvements like **SMOTE** could be considered to enhance stroke prediction.

In comparison, the **Logistic Regression (RFE)** model, which uses **Recursive Feature Elimination (RFE)** for feature selection, had a slightly lower **accuracy of 0.78** and **F1 score of 0.78**, though it maintained the same **ROC AUC of 0.85**. It correctly predicted **727 non-stroke** and **783 stroke** cases. The precision, recall, and F1 score for the no-stroke class (0) were **0.80**,

0.75, and **0.77**, and for the stroke class (1), **0.76**, **0.81**, and **0.78**, respectively. While the accuracy is slightly reduced, the RFE model's primary benefit lies in **feature selection**, which could potentially enhance model interpretability and reduce overfitting. The slight drop in predictive power suggests that while RFE helps streamline the model, it may also lead to the exclusion of important features that could slightly impact overall performance.

Decision Tree [DT]:

The **Decision Tree** is a classification algorithm that splits the dataset into subsets based on the feature that minimizes impurity (using either **Gini Impurity** or **Entropy**). It recursively partitions the data, aiming to create homogeneous subsets with respect to the target variable, which in this case is predicting whether a person has had a stroke or not.

Decision Tree Model Using all Features And Recursive Feature Elimination (RFE):

```
--- Decision Tree ---
Accuracy: 0.9095115681233933
F1 Score: 0.9101123595505618
ROC AUC: 0.9095347607718741
Confusion Matrix:
 [[878  97]
 [ 79 891]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.90      0.91       975
           1       0.90      0.92      0.91       970

    accuracy                           0.91      1945
   macro avg       0.91      0.91      0.91      1945
weighted avg       0.91      0.91      0.91      1945
```

```
--- Decision Tree (RFE) ---
Accuracy: 0.9285347043701799
F1 Score: 0.9284611425630468
ROC AUC: 0.928538197197991
Confusion Matrix:
 [[904  71]
 [ 68 902]]
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.93      0.93       975
           1       0.93      0.93      0.93       970

    accuracy                           0.93      1945
   macro avg       0.93      0.93      0.93      1945
weighted avg       0.93      0.93      0.93      1945
```

The **Decision Tree** model achieved an accuracy of **0.91**, an **F1 score of 0.91**, and a **ROC AUC of 0.91**, with balanced precision and recall across both stroke and non-stroke classes. The model correctly predicted **878 non-stroke** and **891 stroke** cases, indicating strong performance in distinguishing between the two classes. However, **Decision Trees** are prone to overfitting, particularly with deeper trees, which can result in models that perform well on training data but may struggle to generalize to new data.

The **Decision Tree (RFE)** model, incorporating **Recursive Feature Elimination (RFE)** for feature selection, slightly outperformed the base model with an accuracy of **0.93**, an **F1 score of 0.93**, and a **ROC AUC of 0.93**. The model correctly predicted **904 non-stroke** and **902 stroke** cases. RFE helped streamline the model by eliminating less important features, reducing overfitting, and improving generalization. This resulted in a more interpretable model that focused on the most relevant features, achieving superior performance compared to the base **Decision Tree** model.

Random Forest:

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and reduce overfitting. By aggregating predictions from many trees, it enhances the stability and accuracy of the model, making it less sensitive to individual tree variances and better at generalizing to new data.

Random Forest Model Using all Features And Recursive Feature Elimination (RFE):

```
--- Random Forest ---
Accuracy: 0.9598971722365038
F1 Score: 0.9602446483180428
ROC AUC: 0.992548770816812
Confusion Matrix:
 [[925  50]
 [ 28 942]]
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.95      0.96       975
           1       0.95      0.97      0.96       970

    accuracy                           0.96      1945
   macro avg       0.96      0.96      0.96      1945
weighted avg       0.96      0.96      0.96      1945
```

```
--- Random Forest (RFE) ---
Accuracy: 0.9156812339331619
F1 Score: 0.9175050301810865
ROC AUC: 0.9697145122918319
Confusion Matrix:
 [[869 106]
 [ 58 912]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.89      0.91       975
           1       0.90      0.94      0.92       970

    accuracy                           0.92      1945
   macro avg       0.92      0.92      0.92      1945
weighted avg       0.92      0.92      0.92      1945
```

The **Random Forest** model achieved an impressive accuracy of **0.96**, an **F1 score of 0.96**, and a **ROC AUC of 0.99**, reflecting its high performance in predicting both stroke and non-stroke cases. The model correctly predicted **925 non-stroke** and **942 stroke** cases, with precision,

recall, and F1 score for both classes above **0.95**, showcasing its strong ability to balance both classes. The high **ROC AUC** indicates excellent classification capabilities, making it a highly effective model for this task.

The **Random Forest (RFE)** model, which incorporates **Recursive Feature Elimination (RFE)**, slightly underperformed compared to the base model with an accuracy of **0.92**, an **F1 score of 0.92**, and a **ROC AUC of 0.97**. The model correctly predicted **869 non-stroke** and **912 stroke** cases. While RFE helped in selecting the most important features, reducing model complexity, it led to a slight reduction in overall performance compared to the full feature set. Despite this, the model still showed strong classification performance, with good precision and recall, and improved generalization by focusing on fewer but more significant features.

K-Nearest Neighbors [KNN]:

K-Nearest Neighbors (KNN) is a simple, non-parametric algorithm used for classification tasks. It works by finding the "K" nearest neighbors to a data point and making predictions based on the majority class among those neighbors. KNN is intuitive and effective when the data has a clear structure but can struggle with high-dimensional data or imbalanced classes.

KNN Model Using all Features And Recursive Feature Elimination (RFE):

```
--- KNN ---
Accuracy: 0.8966580976863753
F1 Score: 0.9049645390070922
ROC AUC: 0.9546576790906687
Confusion Matrix:
 [[787 188]
 [ 13 957]]
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.81      0.89       975
           1       0.84      0.99      0.90       970

    accuracy                           0.90      1945
   macro avg       0.91      0.90      0.90      1945
weighted avg       0.91      0.90      0.90      1945
```

```
--- KNN (RFE) ---
Accuracy: 0.877120822622108
F1 Score: 0.8843734881470731
ROC AUC: 0.9318234205656886
Confusion Matrix:
 [[792 183]
 [ 56 914]]
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.81      0.87       975
           1       0.83      0.94      0.88       970

    accuracy                           0.88      1945
   macro avg       0.88      0.88      0.88      1945
weighted avg       0.88      0.88      0.88      1945
```

The **KNN** model achieved an accuracy of **0.90**, an **F1 score of 0.90**, and a **ROC AUC of 0.95**, demonstrating solid performance in predicting stroke and non-stroke cases. The confusion matrix indicated **787 non-stroke** and **957 stroke** cases were correctly predicted, with high precision and recall for the stroke class (1) at **0.84** and **0.99**, respectively. However, the recall for the no-stroke class (0) was lower at **0.81**, indicating some difficulty in identifying non-stroke cases, which is common for KNN with class imbalance.

The **KNN (RFE)** model, which applies **Recursive Feature Elimination (RFE)** for feature selection, performed slightly worse, achieving an accuracy of **0.88**, an **F1 score of 0.88**, and a **ROC AUC of 0.93**. With **792 non-stroke** and **914 stroke** cases predicted correctly, the model demonstrated a similar pattern with higher recall for the stroke class but a lower recall for the non-stroke class. RFE, by eliminating less important features, resulted in a simplified model that traded off some predictive power for potentially better interpretability and reduced complexity. Despite the drop in performance, the model still showed reasonable classification results with a focus on fewer features.

Support Vector Machines:

Support Vector Machines (SVM) are supervised learning models used for classification tasks, particularly effective in high-dimensional spaces. SVM works by finding the hyperplane that best separates different classes, aiming to maximize the margin between the classes. It's particularly powerful for datasets with clear margins of separation and is robust to overfitting, especially in high-dimensional spaces.

SVM Model Using all Features And Recursive Feature Elimination (RFE):

```
--- SVM ---
Accuracy: 0.8406169665809768
F1 Score: 0.8515325670498084
ROC AUC: 0.911634152788792
Confusion Matrix:
 [[746 229]
 [ 81 889]]
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.77      0.83       975
           1       0.80      0.92      0.85       970

    accuracy                           0.84      1945
   macro avg       0.85      0.84      0.84      1945
weighted avg       0.85      0.84      0.84      1945
```

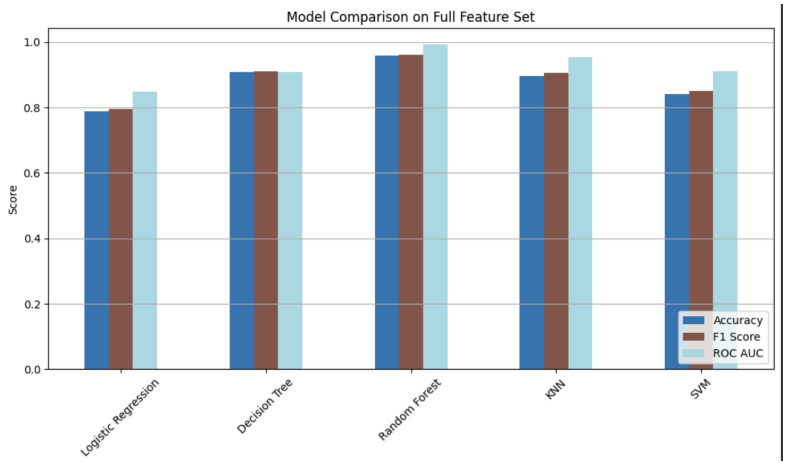
```
--- SVM (RFE) ---
Accuracy: 0.7840616966580977
F1 Score: 0.7922848664688428
ROC AUC: 0.8709807031456516
Confusion Matrix:
 [[724 251]
 [169 801]]
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.74      0.78       975
           1       0.76      0.83      0.79       970

    accuracy                           0.78      1945
   macro avg       0.79      0.78      0.78      1945
weighted avg       0.79      0.78      0.78      1945
```

The **SVM** model achieved an accuracy of **0.84**, an **F1 score of 0.85**, and a **ROC AUC of 0.91**, providing solid performance in classifying stroke and non-stroke cases. The confusion matrix revealed **746 non-stroke** and **889 stroke** cases correctly predicted. Precision and recall for the non-stroke class (0) were **0.90** and **0.77**, while for the stroke class (1), they were **0.80** and **0.92**. Despite the slightly lower precision and recall for the non-stroke class, SVM performed reasonably well with good balance, though some misclassification of non-stroke cases occurred.

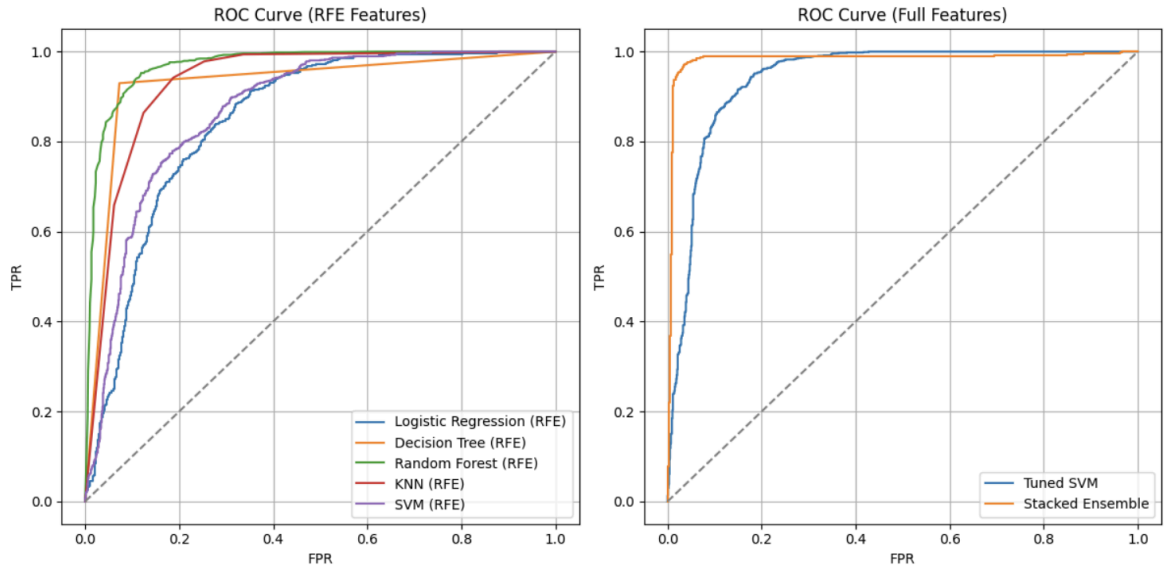
The **SVM (RFE)** model, after applying **Recursive Feature Elimination (RFE)**, showed a drop in performance, achieving an accuracy of **0.78**, an **F1 score of 0.79**, and a **ROC AUC of 0.87**. The confusion matrix showed **724 non-stroke** and **801 stroke** cases predicted correctly. The precision and recall for the non-stroke class (0) were **0.81** and **0.74**, while for the stroke class (1), they were **0.76** and **0.83**. The reduction in accuracy and other metrics suggests that RFE, while improving model interpretability by eliminating less important features, led to a loss of valuable information and a simpler, but less accurate, model.

MODEL COMPARISON

The following graph compares the **Accuracy**, **F1 Score**, and **ROC AUC** highlighting the **Random Forest** model as the top performer, achieving the highest scores across all metrics, followed by the **Decision Tree**. **KNN** and **SVM** show strong performance but with slightly lower scores, especially in accuracy and F1 score. **Logistic Regression** has the lowest performance, particularly in **Accuracy** and **F1 Score**, reflecting its struggle with class imbalance. Overall, ensemble methods outperform others in terms of predictive power and generalization.



The ROC curve comparison between **RFE features** and **full features** highlights the impact of feature selection on model performance. The curve for **full features** shows a higher **ROC AUC**, indicating better overall classification performance with the inclusion of all available features. In contrast, the **RFE features** curve, while still strong, demonstrates a slight reduction in **ROC AUC**, reflecting a trade-off between simplicity and predictive power. Despite the drop, **RFE** helps improve model interpretability and reduce overfitting, though it comes at the cost of slightly less discriminative ability.



Additional Observations: Tuned SVM and Stacked Ensemble

In addition to the baseline models, we evaluated the performance of the **Tuned SVM** and **Stacked Ensemble** models to assess the impact of hyperparameter tuning and ensemble learning techniques on predictive accuracy. The **Tuned SVM** was optimized to improve classification performance, while the **Stacked Ensemble** model combined multiple base algorithms to leverage their strengths and enhance model generalization.

```
--- Tuned SVM ---
Accuracy: 0.8827763496143959
F1 Score: 0.88996138996139
ROC AUC: 0.9418408670367433
Confusion Matrix:
 [[795 180]
 [ 48 922]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.82      0.87       975
           1       0.84      0.95      0.89       970

    accuracy                           0.88      1945
   macro avg       0.89      0.88      0.88      1945
weighted avg       0.89      0.88      0.88      1945
```

The **Tuned SVM** model achieved an accuracy of 88.28%, F1 score of 88.99%, and ROC AUC of 94.18%. These improvements over the base SVM model were driven by better precision for the non-stroke class (0) and improved recall for the stroke class (1). Despite slight trade-offs in precision and recall, the hyperparameter tuning allowed the model to better differentiate between the two classes, enhancing overall classification performance.

```
--- Stacked Ensemble ---
Accuracy: 0.9665809768637532
F1 Score: 0.9665809768637532
ROC AUC: 0.9836627015596088
Confusion Matrix:
 [[940  35]
 [ 30 940]]
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.96      0.97       975
           1       0.96      0.97      0.97       970

    accuracy                           0.97      1945
   macro avg       0.97      0.97      0.97      1945
weighted avg       0.97      0.97      0.97      1945
```

The **Stacked Ensemble** model outperformed all other models with an accuracy of 96.66%, F1 score of 96.66%, and ROC AUC of 98.37%. By combining multiple base classifiers, the ensemble model demonstrated superior precision and recall, especially in correctly predicting stroke and non-stroke cases. This confirms the value of ensemble learning in improving both accuracy and generalization, offering the best results in this analysis.

CONCLUSION

In conclusion, this study demonstrates the effectiveness of **data mining algorithms** in predicting the likelihood of stroke occurrence using healthcare-related features. Among the models evaluated, the **Random Forest** model emerged as the most accurate and robust, achieving the highest scores in terms of **accuracy**, **F1 score**, and **ROC AUC**. The use of techniques like **Recursive Feature Elimination (RFE)** and **Synthetic Minority Over-sampling Technique (SMOTE)** played a significant role in enhancing model performance by addressing class imbalance and improving feature selection. The comparison of individual models and ensemble methods highlights the potential of **stacking classifiers** to achieve superior predictive accuracy, with the **Stacked Ensemble** model outperforming all others.

Overall, this approach provides a valuable tool for **early stroke detection**, offering the potential to improve preventive care and timely intervention. The study also underscores the importance of selecting the right machine learning model and leveraging techniques like RFE and SMOTE to optimize predictive outcomes. Future research could explore the inclusion of additional features, **model tuning**, and **real-time implementation** in clinical settings to further enhance the accuracy and applicability of stroke prediction systems.

REFERENCES

FEDESORIANO (2021). *Stroke Prediction Dataset*. Kaggle.

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download

The Google Colab Notebook we created:

cps844_script_hetu.ipynb