Stroke Prediction Using Machine Learning Models

Data science and advanced Python concepts workshop for Neuroscience

Course code: 8527502001 | **Lecturer:** Osnat Bar Shira

Submitted by:

Amit Davidpur, 207496878 | **Meshi Ben Oz,** 207287566

Introduction

The negative impact of stroke on society has led to concerted efforts to improve its management and diagnosis. With the increasing synergy between technology and medical diagnosis, healthcare providers are creating opportunities for better patient management by systematically mining and archiving patients' medical records. The paper we chose systematically analyzes various factors within electronic health records to enable more effective stroke prediction.

Our final project concentrates on replicating the findings of the original study. While we do not aim to identify the key features for stroke prediction, we implement the machine learning models applied in the original research. This replication seeks to validate the reported results and explore the effectiveness of the models used. We will focus on implementing and testing the machine learning models described in the paper. By doing so, we aim to validate the reported model performance and confirm the feasibility of the methodologies presented.

We selected this paper due to its innovative application of machine learning techniques in the field of stroke prediction, a topic with significant importance in healthcare. In addition The study provides valuable insights into leveraging electronic health records to improve predictive accuracy, making it both relevant and impactful for understanding and addressing this important medical challenge.

Methods

Data

Electronic Health Records (EHR), or Electronic Medical Records (EMR), are digital repositories of patient information maintained by qualified medical practitioners. These records include vital signs, diagnostic results, and medical examinations.

The dataset used in the original study was sourced from a healthcare hackathon challenge organized by McKinsey & Company and is publicly available on Kaggle. It includes various patient attributes and a binary target variable indicating whether a patient suffered a stroke.

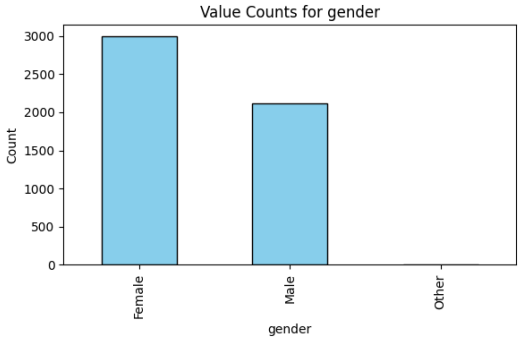
The dataset is highly imbalanced, with most records belonging to patients who did not suffer a stroke. This imbalance poses challenges for training machine learning models and requires careful consideration during analysis.

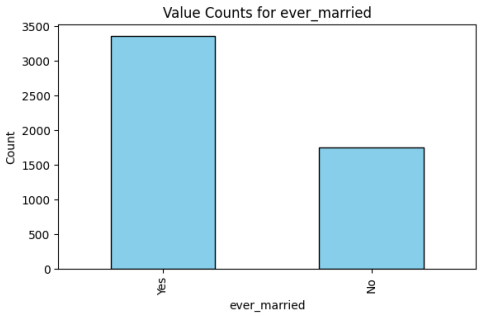
In our project, we used the same dataset; however, the version we accessed is significantly smaller than the one described in the original study. This reduced dataset size may impact the accuracy and generalizability of our findings, particularly in training and evaluating machine learning models.

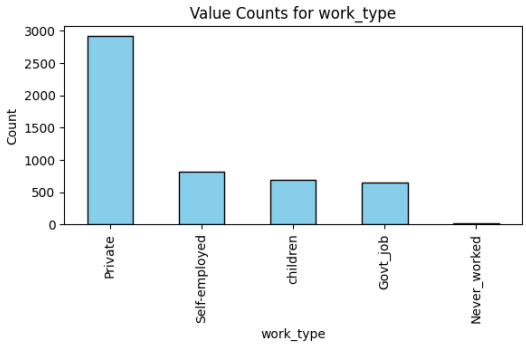
As done in the paper, dimensionality reduction techniques, such as principal component analysis (PCA), were applied to transform the high-dimensional feature space into a more interpretable lower-dimensional subspace. This approach enabled the researchers to better understand the relative importance of each feature and its contribution to stroke prediction. Additionally, the study benchmarked several popular machine learning classification algorithms on the dataset, demonstrating the effectiveness of these models in predicting strokes.

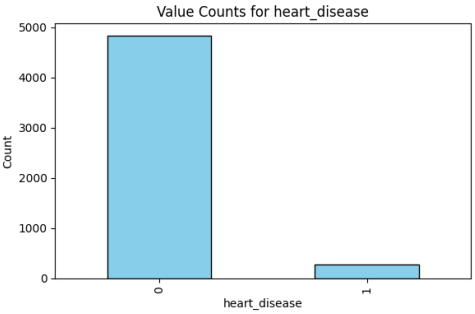
Preprocessing

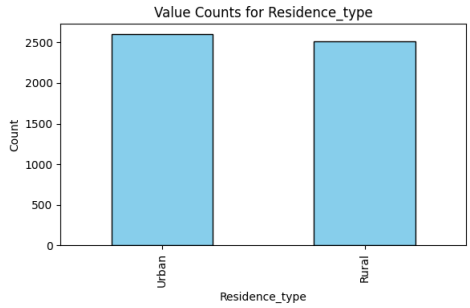
In this stage, we checked the data for duplicates, missing values, and performed descriptive statistical analysis. As the preprocessing steps were not detailed in the paper, we relied on the knowledge acquired during the course to carry out the necessary preprocessing. We considered that this might influence the accuracy of the models and subsequent analyses.
We found no duplicates, and the only missing values were in the BMI column. To visualize the data distribution, we created bar plots for the categorical columns and histograms for the numerical ones.

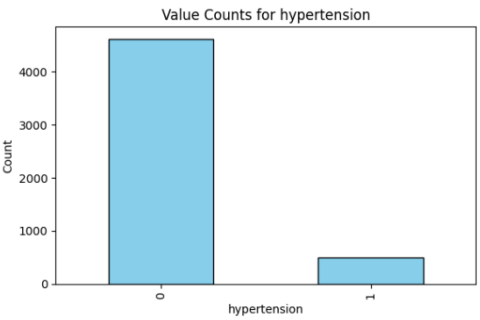


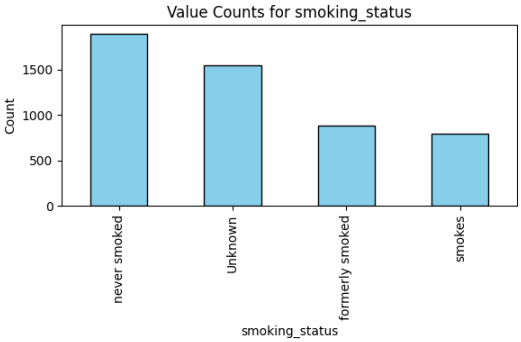


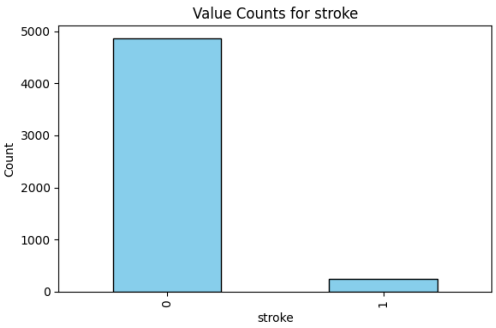




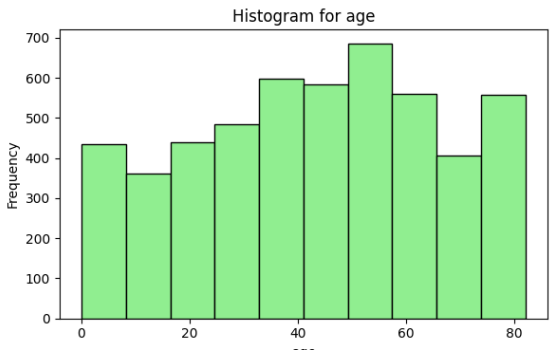


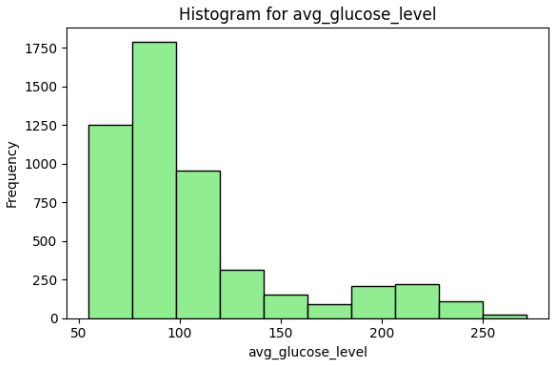


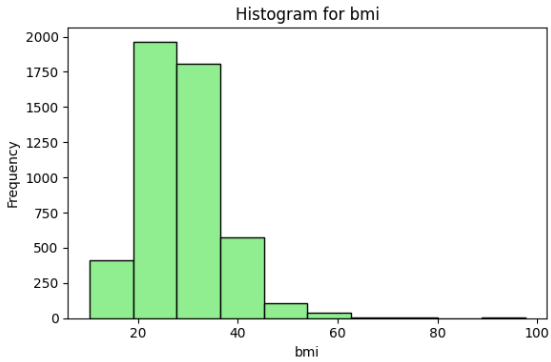




From the graphs, it is evident that most of the patients are healthy and have not experienced a stroke, suggesting an imbalanced dataset. Additionally, the majority of patients are women, and most have been or are currently married.







Looking at the histograms, we observed that the age distribution is close to normal, peaking around the age of 55. Most patients have an average glucose level between 70 and 100, which is considered normal according to medical literature. Regarding BMI, the majority of patients fall within the 20-30 range.

As mentioned, the BMI column had missing values. After careful consideration, we decided to group patients by gender and average glucose level. Based on these groups, we filled the missing BMI values using the median value for each group.

BMI before:

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1

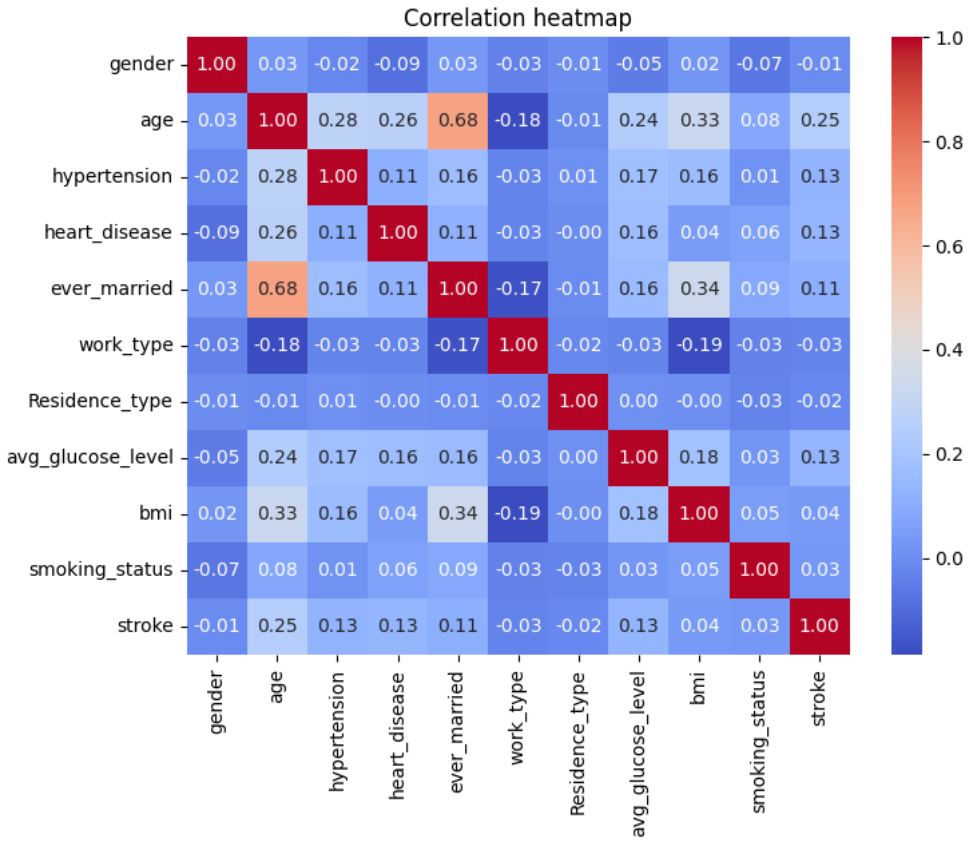
BMI after:

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
Female	61.0	0	0	Yes	Self-employed	Rural	202.21	31.8	never smoked	1
Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1

Features analysis

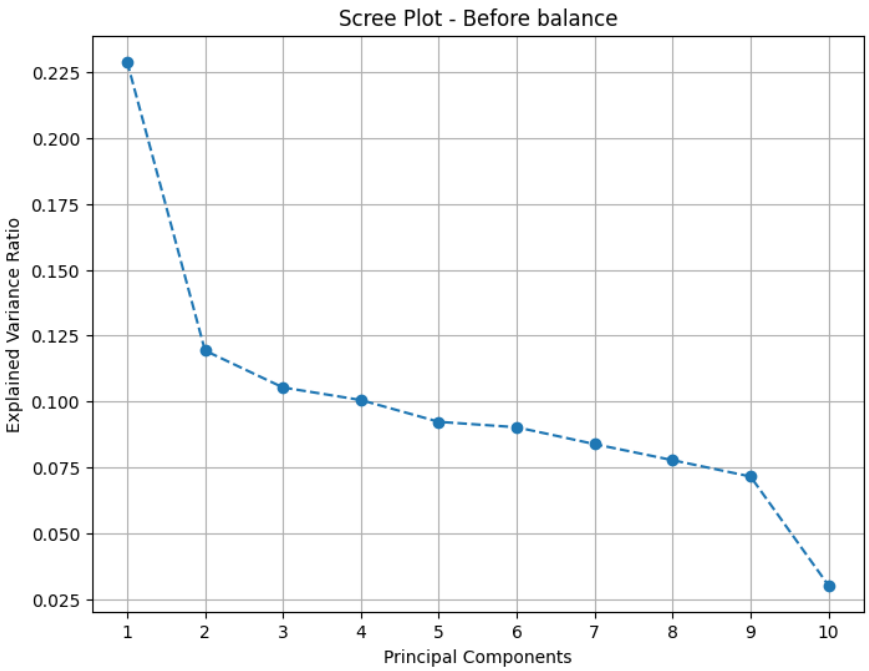
In the second stage, we performed an analysis of the relationships between the features and balanced the data.

First, we examined the correlations. As described in the article, we utilized Pearson's correlation coefficient to determine the relationships between various patient features. A correlation heatmap was created, where red indicates a positive correlation and blue indicates a negative correlation. The intensity of the color reflects the strength of the correlation, with deeper shades representing stronger relationships between the features.



Based on the heatmap, we observed a strong positive correlation between age and marital status (ever_married), with a correlation value of 0.68. However, for the other features, there is little to no correlation.

Following the study's approach, we applied PCA to transform the dataset into uncorrelated principal components, capturing maximum variance. Features were standardized by scaling to have a mean of zero and a standard deviation of one. A scree plot was used to identify components explaining most of the variance.



Explained variance by each principal component:

PC1: 22.89%

PC2: 11.93%

PC3: 10.53%

PC4: 10.06%

PC5: 9.23%

PC6: 9.02%

PC7: 8.38%

PC8: 7.77%

PC9: 7.16%

PC10: 3.03%

In the paper, 10 principal components were derived, with 8 components explaining 88.2% of the variance. In our analysis, as shown above, the variance explained by the top 8 principal components was 89.81%.

We checked the contributions of the different features in the first and second principal components.

```
Sorted Contributions of Features to PC1:
age                0.56
ever_married       0.51
bmi                0.38
avg_glucose_level  0.29
hypertension       0.28
heart_disease      0.25
work_type          0.22
smoking_status     0.12
gender             0.02
residence_type     0.02
```

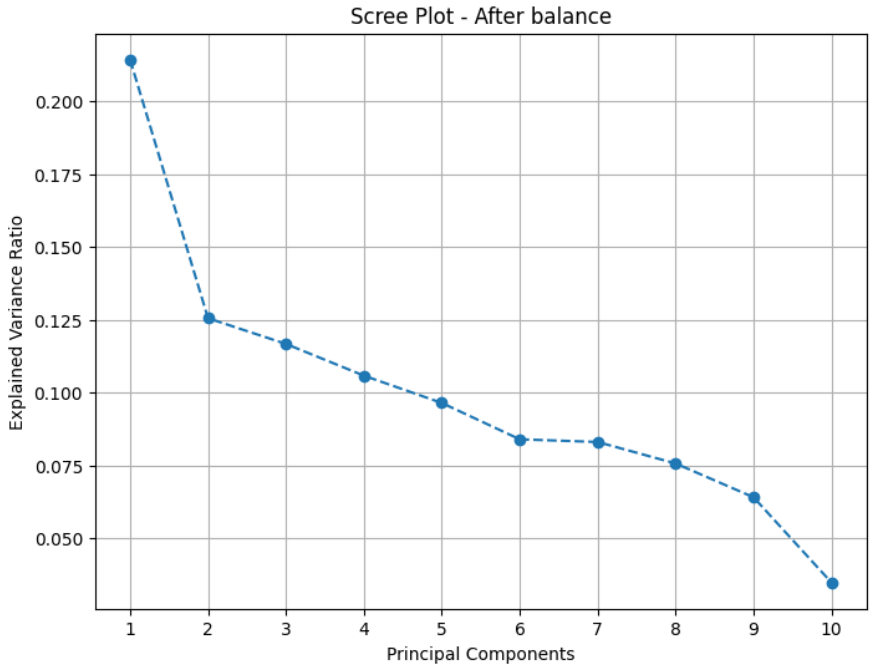
```
Sorted Contributions of Features to PC2:
gender             0.58
heart_disease      0.47
work_type          0.40
avg_glucose_level  0.35
bmi                0.23
hypertension       0.20
ever_married       0.20
smoking_status     0.17
age                0.05
residence type     0.01
```

In the paper, the features age and ever_married are among the most significant contributors to PC1, similar to our findings. For PC2, both analyses highlight gender and work_type as key contributors, with noticeable similarities in their ranking.

Balance the data

The original dataset is unbalanced because there are more samples possessing negative stroke labels (4,861), as compared to positive label stroke samples (249). We balanced the data by considering all the positive stroke samples, and then randomly picked equal number of negative stroke samples from the rest. This resulted in a balanced dataset with an equal number of positive and negative stroke samples, with 249 samples from each group.

After balancing the dataset, we performed PCA and included a scree plot to examine whether the variance in the data and the features contributing most to it remained the same.



Explained variance by each principal component:

PC1: 21.43%

PC2: 12.55%

PC3: 11.67%

PC4: 10.58%

PC5: 9.64%

PC6: 8.39%

PC7: 8.30%

PC8: 7.56%

PC9: 6.40%

PC10: 3.47%

We checked the contributions of the different features in the first and second principal components.

Sorted Contributions of Features to PC1:

age	0.49
ever_married	0.46
bmi	0.40
avg_glucose_level	0.37
hypertension	0.28
work_type	0.28
heart_disease	0.24
residence_type	0.13
smoking_status	0.12
gender	0.05

Sorted Contributions of Features to PC2:

heart_disease	0.51
gender	0.50
smoking_status	0.33
avg_glucose_level	0.30
bmi	0.28
work_type	0.27
residence_type	0.26
ever_married	0.24
hypertension	0.08
age	0.02

We observed that, consistent with the findings reported in the paper, the contributions of the features remained largely unaffected after balancing the dataset. Age and ever_married had the highest contributions to PC1, while gender and heart disease contributed the most to PC2.

Models' implementation

In the final stage, we applied machine learning models to predict stroke outcomes in the data. Our goal was to identify the best hyperparameters for each model. To achieve this, as taught in the course, we created a function that performs grid search to find the hyperparameters that yield the optimal model.

We began by testing three models, as done in the paper: Decision Tree, Random Forest, and Neural Network. For each model, we selected relevant hyperparameters to search for the best ones.
As described in the paper, we first evaluated the models on the balanced dataset using all features, then tested them on the balanced dataset reduced to 2 and 8 principal components using PCA. This approach allowed us to assess the models' performance at different levels of feature reduction.
For each model, we used the grid search function to find the best hyperparameters and displayed them. Additionally, we presented the accuracy and other evaluation metrics for each model using the optimal hyperparameters.

Furthermore, we tested two additional models, SVM and Gradient Boosting, with the goal of identifying the best model for stroke prediction on our dataset. As before, we performed a hyperparameter search and reported the evaluation metrics for each model, both using all features and after dimensionality reduction with PCA, with 2 and 8 components.

Results

Results of the models over all features:

```
Decision Tree's best parameters:{'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}
Random Forest's best parameters:{'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 200}
Neural Network's best parameters:{'activation': 'relu', 'hidden_layer_sizes': (50,), 'solver': 'sgd'}
SVM's best parameters:{'C': 0.1, 'kernel': 'linear'}
Gradient Boosting's best parameters:{'learning_rate': 0.01, 'max_depth': 3, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 50}
```

Model	Precision	Recall	F-Score	Accuracy	Miss rate	Fall-out rate
Decision Tree	0.70	0.66	0.68	0.66	0.34	0.34
Random Forest	0.75	0.77	0.76	0.73	0.23	0.31
Neural Network	0.75	0.77	0.76	0.73	0.23	0.31
SVM	0.75	0.77	0.76	0.73	0.23	0.31
Gradient Boosting	0.75	0.72	0.73	0.71	0.28	0.29

Results of the models with the first 2 principal components:

```
Decision Tree's best parameters:{'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}
Random Forest's best parameters:{'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 100}
Neural Network's best parameters:{'activation': 'relu', 'hidden_layer_sizes': (50,), 'solver': 'sgd'}
SVM's best parameters:{'C': 10, 'kernel': 'poly'}
Gradient Boosting's best parameters:{'learning_rate': 0.01, 'max_depth': 3, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 50}
```

Model	Precision	Recall	F-Score	Accuracy	Miss rate	Fall-out rate
Decision Tree	0.63	0.49	0.55	0.57	0.51	0.34
Random Forest	0.71	0.65	0.68	0.66	0.35	0.32
Neural Network	0.74	0.66	0.70	0.69	0.34	0.28
SVM	0.64	0.71	0.67	0.62	0.29	0.49
Gradient Boosting	0.71	0.60	0.65	0.65	0.40	0.29

Results of the models with the first 8 principal components:

```
Decision Tree's best parameters:{'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2}
Random Forest's best parameters:{'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 50}
Neural Network's best parameters:{'activation': 'tanh', 'hidden_layer_sizes': (50, 50), 'solver': 'sgd'}
SVM's best parameters:{'C': 1, 'kernel': 'linear'}
Gradient Boosting's best parameters:{'learning_rate': 0.01, 'max_depth': 3, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200}
```

Model	Precision	Recall	F-Score	Accuracy	Miss rate	Fall-out rate
Decision Tree	0.61	0.51	0.56	0.55	0.49	0.40
Random Forest	0.76	0.74	0.75	0.73	0.26	0.28
Neural Network	0.77	0.77	0.77	0.75	0.23	0.28
SVM	0.75	0.76	0.75	0.73	0.24	0.31
Gradient Boosting	0.72	0.74	0.73	0.70	0.26	0.35

Discussion and Conclusion

The project aimed to replicate and validate the findings of a study that applied machine learning techniques to electronic health records (EHR) for stroke prediction. The primary objective was to assess the performance of various models, including Decision Tree, Random Forest, Neural Networks, SVM, and Gradient Boosting, using both original features and dimensionality-reduced data through PCA.

To assess the methodologies outlined in the study, we began by implementing three models: Decision Tree, Random Forest, and Neural Network. Following hyperparameter optimization using grid search, we achieved accuracy results comparable to the original study. For the Decision Tree model, our accuracy was 0.66, slightly below the article's reported 0.74. Both the Random Forest and Neural Network models achieved an accuracy of 0.73, nearly identical to the article's reported 0.74. These findings reflect the robustness of these models, even with differences in dataset size and preprocessing.

We further evaluated these models using PCA-reduced datasets, as done in the original study. Using the first two PCA components, we observed a decline in accuracy: the Decision Tree achieved 0.57 (compared to 0.73 in the article), the Random Forest 0.66 (compared to 0.69), and the Neural Network 0.69 (compared to 0.74). With the first eight PCA components, which captured 88% of the data variance, the Decision Tree's accuracy dropped to 0.55 (compared to 0.73), while the Random Forest and Neural Network reached 0.73 and 0.75, closely aligning with the article's results of 0.72 and 0.75, respectively. Notably, the Neural Network consistently outperformed the other models across all conditions, particularly with eight PCA components.

In the second phase, we expanded our analysis by incorporating SVM and Gradient Boosting models. Using the full set of features, the SVM model achieved an accuracy of 0.73, surpassing the 0.68 reported in the article, while Gradient Boosting yielded an accuracy of 0.71. On the dataset reduced to two PCA components, accuracy for SVM and Gradient Boosting dropped to 0.62 and 0.65, respectively. However, with eight PCA components, SVM reached 0.73 and Gradient Boosting 0.70, confirming the effectiveness of dimensionality reduction while maintaining competitive performance.

Our results reinforce the Neural Network model's superiority, particularly with eight PCA components, where it consistently achieved the highest accuracy of 0.75, matching the original study. These findings validate the reliability of machine learning models for stroke prediction and emphasize the critical role of hyperparameter tuning and dimensionality reduction in optimizing predictive performance. The slight variations in accuracy between our results and the article's are attributed to differences in dataset size and preprocessing, underscoring the importance of transparent data handling in replicable research.

Accuracy scores comparison:

Model	**All features** (Original Paper)	**All features** (Our Findings)	**2 PCA** (Original Paper)	**2 PCA** (Our Findings)	**8 PCA** (Original Paper)	**8 PCA** (Our Findings)
Decision Tree	0.74	0.66	0.73	0.57	0.73	0.55
Random Forest	0.74	0.73	0.69	0.66	0.72	0.73
Neural Network	0.74	0.73	0.74	0.69	**0.75**	**0.75**
SVM	0.68	0.73	-	0.62	-	0.73
Gradient Boosting	-	0.71	-	0.65	-	0.70

References:

Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., & John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics*, *2*, 100032.

Link to the data:
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset