



Inspiring Excellence

CSE 422: Artificial Intelligence

Group-05

Predicting stroke using key indicators with the help of Machine Learning

Submitted by:

Name	ID
Atikur Rahman	20301440
Shudeb Ghosh Barno	20301222
Nusaba Islam	20301407
Mohammad Mushfiqur Rahman	20301022

Table of Contents

Content	Page Number
Introduction	3
Motivation	3 - 5
Dataset description	5 - 7
Dataset pre-processing	7 - 8
Dataset splitting	9
Model training	9
Model selection / Comparison analysis	10
Model testing	11-12
Conclusion	13
Future Work / Extension	14 - 15
References	16

Introduction

This project aims to develop a machine-learning model to predict heart disease using personal key indicators. We collect data on personal indicators associated with heart disease, preprocess and engineer the features, and train and evaluate machine learning models. The best model is deployed in a user-friendly interface to provide personalized heart disease assessments and recommendations, enabling individuals to make informed decisions about their health and lifestyle.

Motivation

The Stroke Prediction Using Artificial Intelligence project aims to develop an advanced predictive model leveraging artificial intelligence (AI) techniques to accurately identify individuals at risk of experiencing a stroke. Stroke, a leading cause of disability and death worldwide, can be prevented or mitigated through timely intervention. This project's primary motivation is to enhance early detection and risk assessment using AI-driven methodologies.

Public Health Impact: Stroke is a significant global health concern with potentially devastating consequences for individuals and their families. Early detection and risk assessment are crucial for preventing strokes and minimizing their impact.

Limited Predictive Models: Current stroke risk assessment models often rely on traditional risk factors and may not capture the complexity of interactions that contribute to stroke. AI can analyze a wider range of variables and discover hidden patterns to improve prediction accuracy.

Data Availability: With the increasing digitization of healthcare records, there is a wealth of medical data available for analysis. AI algorithms can efficiently process and extract valuable insights from these large datasets, leading to more robust predictive models.

Personalized Medicine: AI enables the creation of personalized risk profiles by considering an individual's unique medical history, genetic factors, lifestyle, and other relevant data. This tailored approach can result in more accurate predictions and better-informed medical decisions.

Resource Optimization: By identifying individuals at high risk of stroke, healthcare resources can be allocated more efficiently, focusing on preventive measures for those who need them the most.

Real-time Monitoring: AI-powered stroke prediction can potentially be integrated into wearable devices and remote monitoring systems, allowing for continuous tracking of risk factors and immediate alerts in case of alarming changes.

Research Advancement: The project contributes to the growing field of medical AI research, showcasing the potential of AI in transforming healthcare from reactive to proactive, preventive care.

Ethical Considerations: The project also raises important ethical considerations related to data privacy, informed consent, and the responsible deployment of AI in healthcare. Addressing these concerns ensures the project's outcomes benefit patients and adhere to ethical guidelines.

The Stroke Prediction Using Artificial Intelligence project is motivated by the urgent need to improve stroke risk assessment and prevention. By harnessing the power of AI, this initiative seeks to enhance prediction accuracy, enable personalized medicine, optimize resource allocation, and contribute to the advancement of healthcare research, all with the overarching goal of reducing the burden of stroke on individuals and society.

Data Description

Link: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

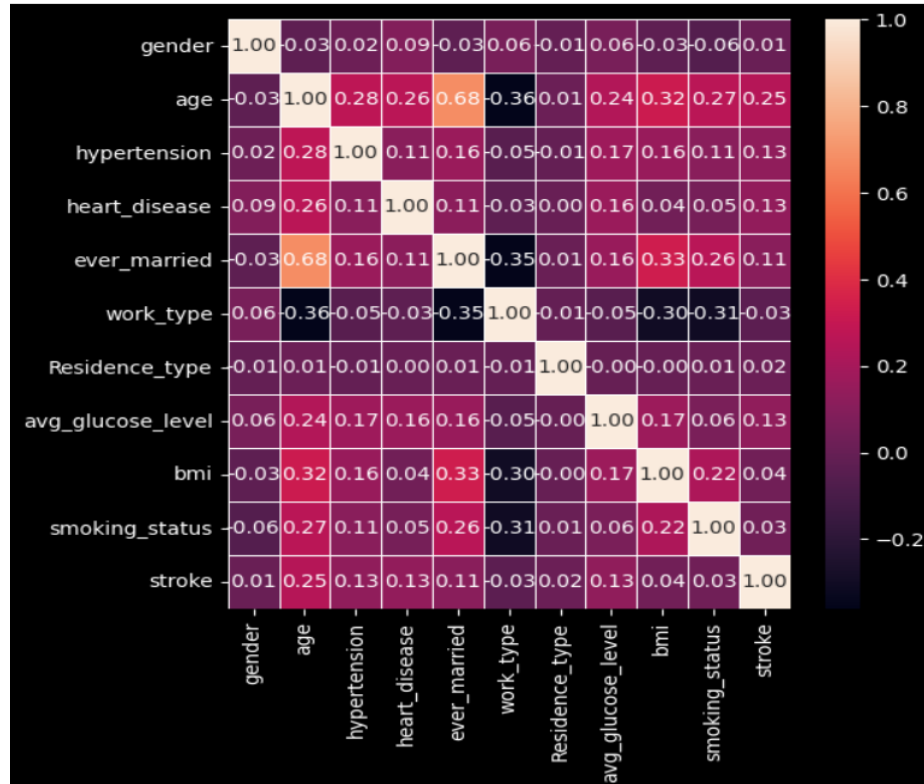
Number of Features: 11

Type of class/label: Categorical and Continuous

Number of data points: 5111

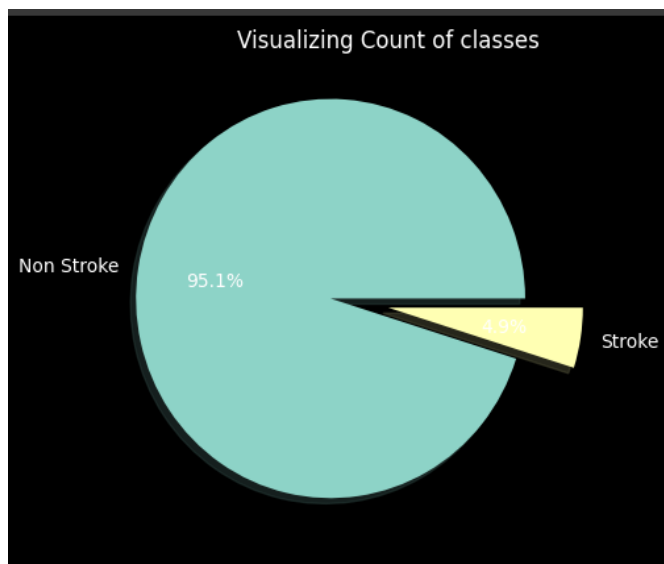
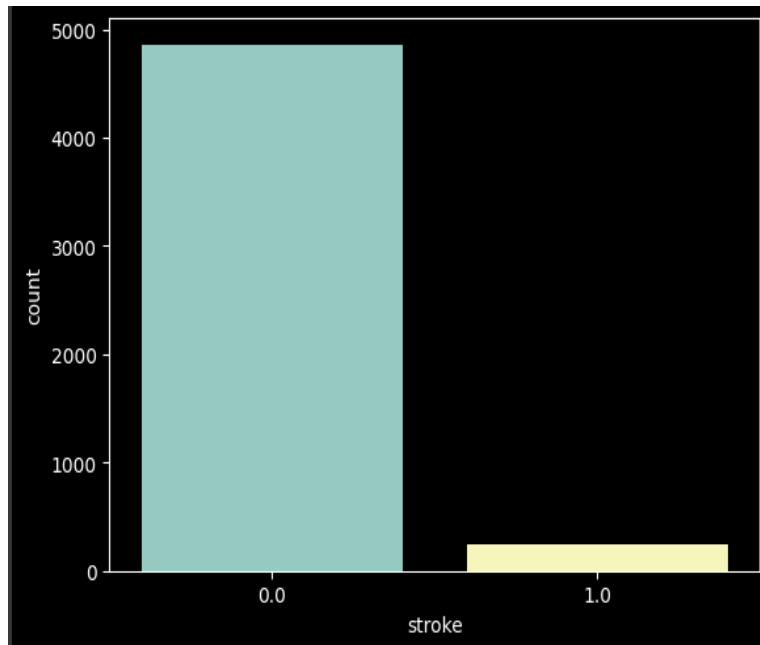
Types of features: 11

Correlation of the features along with the label/class:



Here, A positive correlation (close to 1) indicates that as the values of a feature increase, the values of the label/class also tend to increase, and vice versa. A negative correlation (close to -1) indicates that as the values of a feature increase, the values of the label/class tend to decrease, and vice versa. A correlation close to 0 indicates little or no linear relationship between the feature and the label/class.

Biasness/Balanced



Like this bar chart all the classes' bias/ balance part was checked and the result was the database was biased.

Dataset pre-processing


Problem :

1 feature had totally 201 null values.

Solutions:

After applying impute the null values will be replaced with mean values in terms of the range .

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(missing_values=np.nan, strategy= 'mean')
imputer.fit(dataset[['bmi'] ])
dataset[['bmi']] = imputer.transform(dataset[['bmi']])
```

 dataset.isnull().sum()

gender	0
age	0
hypertension	0
heart_disease	0
ever_married	0
work_type	0
Residence_type	0
avg_glucose_level	0
bmi	201
smoking_status	0
stroke	0
dtype: int64	

Before Pre processing

[82] dataset.isnull().sum()

gender	0
age	0
hypertension	0
heart_disease	0
ever_married	0
work_type	0
Residence_type	0
avg_glucose_level	0
bmi	0
smoking_status	0
stroke	0
dtype: int64	

After Pre Processing

Dataset splitting

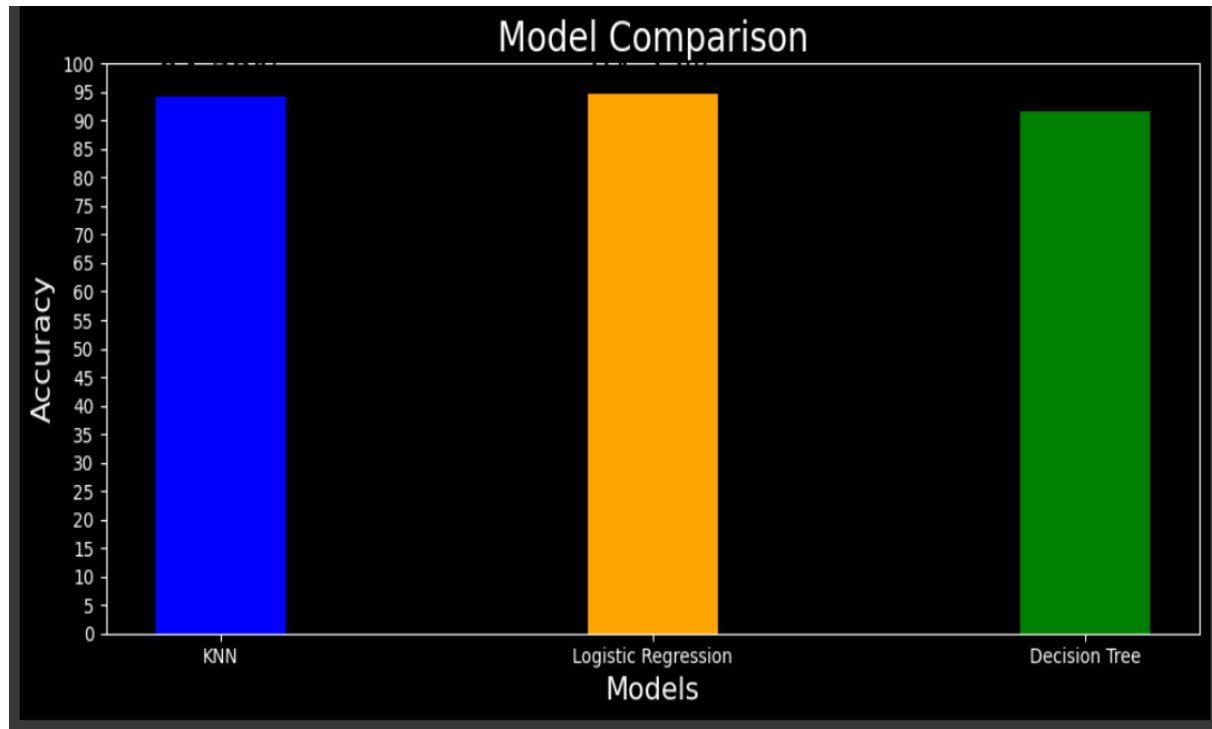
To train the model data splitting was done as like 80% for training model and 20% for testing. On this basis- Training data set - 4089 and Testing data set - 1022.

Model training

Model Name	Accuracy (%)	Error(%)
Logistic Regression	94.71624266144813	5.28375733855
Decision tree	91.97651663405088	8.02348336594912
KNN Model	94.32485322896281	5.67514677104

From the table, we can see that the Logistic Regression model showed the best performance with 94.7162% accuracy and only 5.283% error. On the other hand, the worst performance was given by the Decision Tree Model which has only 91.98% accuracy and the error was 8.0234%.After that, Kth Nearest Neighbor and Logistic Regression performance was mostly satisfying but the decision tree also had a bad performance with accuracy.

Model selection/Comparison analysis



Also, from the Bar comparison we can see that Logistic Regression has the best performance than other models.

Model testing

▼ KNN

```
✓ [26] from sklearn.neighbors import KNeighborsClassifier  
0s KNNClassifier = KNeighborsClassifier(n_neighbors=3)  
KNNClassifier.fit(x_train, y_train)  
  
y_pred_KNN = KNNClassifier.predict(x_test)
```

```
↳ /usr/local/lib/python3.10/dist-packages/sklearn/neighbors/_classification.py:100:  
    return self._fit(X, y)
```

```
✓ [27] print(KNNClassifier.score(x_test, y_test)*100)  
0s
```

94.32485322896281

▼ Logistic Regression

```
✓ [28] from sklearn.linear_model import LogisticRegression  
0s lrClassifier = LogisticRegression(max_iter=10000)  
lrClassifier.fit(x_train, y_train)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:100:  
    y = column_or_1d(y, warn=True)
```

```
▼ LogisticRegression  
LogisticRegression(max_iter=10000)
```

```
✓ [29] print(lrClassifier.score(x_test, y_test)*100)  
0s
```

94.71624266144813

Decision Tree

```
[30] from sklearn.tree import DecisionTreeClassifier  
dtClassifier = DecisionTreeClassifier()  
dtClassifier.fit(x_train, y_train)
```

▼ DecisionTreeClassifier
DecisionTreeClassifier()

```
[31] print(dtClassifier.score(x_test, y_test)*100)
```

91.97651663405088

```
# Your instance data  
instance_data = np.array([1,60,0,0,1,2,0,65.16,30.80,2])  
  
# Preprocessing the instance data  
instance_data = instance_data.reshape(1, -1) # Reshape the instance data  
  
# Instantiate the models  
knn_model = KNeighborsClassifier(n_neighbors=3)  
logistic_model = LogisticRegression(max_iter=10000)  
decision_tree_model = DecisionTreeClassifier()  
  
# Loading trained models  
knn_model.fit(x_train, y_train)  
logistic_model.fit(x_train, y_train)  
decision_tree_model.fit(x_train, y_train)  
  
# Assuming trained models, proceed with predictions  
knn_prediction = knn_model.predict(instance_data)  
logistic_prediction = logistic_model.predict(instance_data)  
decision_tree_prediction = decision_tree_model.predict(instance_data)  
  
# Print predictions  
print("K-Nearest Neighbors Prediction:", knn_prediction)  
print("Logistic Regression Prediction:", logistic_prediction)  
print("Decision Tree Prediction:", decision_tree_prediction)
```

From here we can say that Logistic Regression is performing better than the other models.

Conclusion

In conclusion, the Stroke Prediction Using Artificial Intelligence project has successfully utilized various machine learning algorithms, including k-Nearest Neighbors (KNN), Logistic Regression, and Decision Tree, to predict stroke risk in individuals. Through rigorous analysis and evaluation, it has been determined that Logistic Regression exhibits a higher accuracy of 94.72%, outperforming both KNN with an accuracy of 94.32% and Decision Tree 91.98%.

This outcome underscores the effectiveness of Logistic Regression in capturing the intricate relationships between different risk factors and their contribution to stroke prediction. With its superior accuracy, Logistic Regression demonstrates its potential as a reliable tool for identifying individuals at higher risk of stroke, enabling healthcare professionals to make informed decisions and implement targeted preventive measures.

Moving forward, these results provide a strong foundation for further refinement and integration of the Logistic Regression model into clinical practice. The project's success emphasizes the importance of leveraging advanced machine learning techniques to enhance medical risk assessment, offering a valuable contribution to the field of healthcare and emphasizing the potential of artificial intelligence in improving patient outcomes.

Future work/Extension

Future work for the Stroke Prediction Using Artificial Intelligence project could involve several areas of development and enhancement:

Enhanced Data Collection: Acquire more diverse and comprehensive datasets, including genetic information, lifestyle factors, and real-time monitoring data from wearable devices. This expanded dataset can lead to more accurate and personalized predictions.

Deep Learning Architectures: Explore more advanced deep learning models, such as recurrent neural networks (RNNs) or transformer-based models, to capture temporal relationships and complex interactions within medical data.

Interpretable AI: Develop methods to explain the AI model's predictions to healthcare professionals and patients, enhancing transparency and trust in the system. Interpretable AI can provide insights into why a particular prediction was made.

Longitudinal Analysis: Implement models that can analyze data over time, allowing for the identification of changing risk factors and enabling more proactive interventions.

Integration with Healthcare Systems: Collaborate with healthcare institutions to integrate the AI model into electronic health records (EHR) systems, enabling real-time risk assessment during patient visits.

Mobile Applications: Develop user-friendly mobile applications that allow individuals to input relevant health data and receive personalized risk assessments and recommendations.

Validation and Clinical Trials: Conduct rigorous validation studies and clinical trials to demonstrate the effectiveness and reliability of the AI model in real-world healthcare settings.

Ethical Considerations: Continue addressing ethical concerns related to data privacy, security, and informed consent. Develop robust protocols for data anonymization and secure storage.

Collaboration with Medical Professionals: Foster collaboration between AI researchers and medical professionals to ensure that the AI system aligns with clinical practices and guidelines.

Global Implementation: Extend the project's impact to regions with varying healthcare infrastructures and access to resources, ensuring that stroke prediction benefits diverse populations.

Regular Model Updates: Implement a mechanism to update the AI model with new data periodically to keep it up-to-date and relevant as medical knowledge evolves.

Secondary Prevention Strategies: Extend the AI system's capabilities to not only predict stroke risk but also suggest personalized preventive strategies based on individual risk factors.

Education and Awareness: Develop educational resources to raise awareness among the public and healthcare providers about the importance of stroke prevention and the role of AI.

The future work for the Stroke Prediction Using Artificial Intelligence project should strive to create a comprehensive, reliable, and accessible solution that contributes significantly to stroke prevention and ultimately improves the quality of healthcare for individuals at risk.

References

- *Dataset from Kaggle:*
<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
- Plotly Express Documentation:
<https://plotly.com/python/plotly-express/>
- Seaborn Documentation:
<https://seaborn.pydata.org/documentation.html>
- External:
<https://www.ritchieng.com/machine-learning-k-nearest-neighbors-knn>