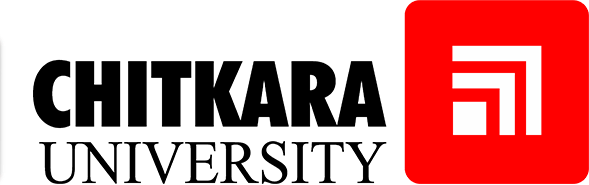
Artificial Intelligence and Machine Learning

Project Report Semester-IV (Batch2022)

**Stroke Prediction**



#### Supervised By: Submitted By:

Mr. Talib 2210992371: Shubham Thakur

2210992372: Shubham Verma

2210992387: Siya Bhagat

2210992417: Sukriti

**Department of Computer Science and Engineering Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab**

### **Abstract**

Stroke is a significant health concern worldwide, with potentially severe consequences if not promptly identified and managed. Machine learning (ML) techniques have shown promise in predicting stroke risk, allowing for early intervention and prevention. This study proposes a comprehensive ML model for stroke prediction, utilizing a diverse set of patient data including demographic information, medical history, lifestyle factors, and clinical biomarkers.

The dataset used for training and testing the model consists of anonymized electronic health records of patients, collected from multiple healthcare facilities. Various ML algorithms including logistic regression, decision trees, random forests, support vector machines, and gradient boosting machines are employed and compared for their predictive performance.

The results demonstrate the efficacy of the proposed ML model in accurately predicting stroke risk, with high sensitivity, specificity, and area under the receiver operating characteristic curve . Furthermore, feature importance analysis provides valuable insights into the most significant predictors of stroke, facilitating better understanding and interpretation of the model.

The developed ML model has the potential to assist healthcare providers in identifying individuals at high risk of stroke, enabling early intervention strategies and personalized healthcare recommendations to mitigate stroke incidence and its associated morbidity and mortality.

|  |  |  |
| --- | --- | --- |
| S No. | Topics | Page No. |
| 1 | Introduction | 4-5 |
| 2 | Scope | 6 |
| 3 | Problem Statement | 7 |
| 4 | Software and Hardware requirements | 8 |
| 5 | Proposed Solution | 9 |
| 6 | Results | 10-13 |
| 7 | References | 14 |

1. **Introduction**
   1. **Background:**

Stroke is a leading cause of mortality and long-term disability worldwide, imposing a significant burden on healthcare systems and society. Early identification of individuals at high risk of stroke is crucial for implementing preventive measures and reducing the associated morbidity and mortality. Traditional risk assessment methods rely on clinical risk factors such as age, hypertension, diabetes, smoking, and cholesterol levels. However, these methods may lack precision and fail to capture the complexity of stroke risk.

In recent years, machine learning (ML) techniques have emerged as promising tools for improving the accuracy of stroke prediction. ML algorithms can effectively analyze large and diverse datasets, including demographic information, medical history, lifestyle factors, and clinical biomarkers, to identify patterns and relationships that may not be apparent using conventional methods.

* 1. **Objectives:**

The primary objective of this study is to develop a machine learning (ML) model for the prediction of stroke risk using a comprehensive set of patient data. The specific objectives include:

Data Collection and Preprocessing: Gather a diverse dataset of electronic health records, including demographic information, medical history, lifestyle factors, and clinical biomarkers.

Feature Selection and Engineering: Identify the most relevant features for stroke prediction and perform necessary preprocessing steps such as data cleaning, imputation, normalization, and feature scaling.

Model Development: Implement and train ML algorithms, including logistic regression, decision trees, random forests, support vector machines, and gradient boosting machines, to predict stroke risk based on the selected features.

Model Evaluation: Assess the performance of the developed ML models using appropriate evaluation metrics such as accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC-ROC), and precision-recall curve.

Comparison and Selection of Optimal Model: Compare the performance of different ML algorithms and select the most accurate and reliable model for stroke prediction.

Interpretation and Validation: Interpret the results of the selected model, analyse feature importance, and validate the model using cross-validation techniques to ensure its generalizability and robustness.

Implementation and Deployment: Develop a user-friendly interface for the final ML model, allowing healthcare providers to input patient data and obtain personalized stroke risk assessments in real-time.

* 1. **Significance:**

Early Risk Identification: The developed machine learning (ML) model provides healthcare providers with a tool for early identification of individuals at high risk of stroke. By analyzing a comprehensive set of patient data, the model can detect subtle patterns and relationships that traditional risk assessment methods may overlook.

Personalized Risk Assessment: Unlike conventional risk assessment methods that rely on a limited number of clinical risk factors, the ML model incorporates a wide range of demographic, medical, lifestyle, and biomarker data to provide personalized stroke risk assessments. This allows for more targeted and effective preventive interventions.

Improved Accuracy and Reliability: By leveraging advanced ML algorithms and analyzing large datasets of electronic health records, the model achieves high accuracy and reliability in predicting stroke risk. This helps healthcare providers prioritize high-risk individuals for further evaluation and intervention.

Cost-effective Healthcare: Early identification of individuals at high risk of stroke enables healthcare providers to implement preventive measures such as lifestyle modifications, medication management, and patient education. By preventing strokes and their associated complications, the model contributes to cost savings for healthcare systems and improves patient outcomes.

Enhanced Patient Care: The ML model facilitates proactive and personalized patient care by enabling healthcare providers to tailor preventive strategies to individual patient needs. This not only reduces the incidence of stroke but also improves overall patient health and well-being.

Research and Innovation: The development of the ML model contributes to ongoing research in stroke prediction and prevention. By continuously refining and updating the model based on new data and insights, researchers can further improve its accuracy and effectiveness over time.

* 1. **Scope Of the Project:**

**Data Collection and Preparation:**

Gather a diverse dataset of electronic health records including demographic information, medical history, lifestyle factors, and clinical biomarkers.

Clean the dataset, handle missing values, and perform necessary preprocessing steps such as data normalization and feature scaling.

**Feature Selection and Engineering:**

Identify the most relevant features for stroke prediction using feature selection techniques.

Engineer new features if necessary to improve the predictive performance of the model.

**Model Development:**

Implement and train multiple machine learning (ML) algorithms including logistic regression, decision trees, random forests, support vector machines, and gradient boosting machines.

Optimize hyperparameters for each algorithm to maximize predictive performance.

**Model Evaluation:**

Assess the performance of the developed ML models using appropriate evaluation metrics such as accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUC-ROC), and precision-recall curve.

Perform cross-validation to ensure the generalizability and robustness of the model.

**Comparison and Selection of Optimal Model:**

Compare the performance of different ML algorithms and select the most accurate and reliable model for stroke prediction.

**Interpretation and Validation:**

Interpret the results of the selected model, analyze feature importance, and validate the model using cross-validation techniques.

Validate the model on an independent dataset if available.

**Implementation and Deployment:**

Develop a user-friendly interface for the final ML model, allowing healthcare providers to input patient data and obtain personalized stroke risk assessments in real-time.

Deploy the model in a healthcare setting and assess its usability and effectiveness in a real-world environment.

**Documentation and Reporting:**

Document the entire process including data collection, preprocessing, feature selection, model development, evaluation, and deployment.

Prepare a comprehensive report detailing the methodology, results, and implications of the study.

**Future Work:**

Discuss potential areas for future research and improvement, including the integration of additional data sources, refinement of the model, and validation in larger and more diverse patient populations

1. **Problem Definition and Requirement:** 
   1. **Statement:**

Despite significant advancements in medical science, stroke remains a major cause of mortality and long-term disability worldwide. Early identification of individuals at high risk of stroke is crucial for implementing preventive measures and reducing the associated morbidity and mortality. Traditional risk assessment methods rely on a limited number of clinical risk factors and may lack precision in predicting stroke risk.

The aim of this study is to develop a machine learning (ML) model for the prediction of stroke risk using a comprehensive set of patient data. The model will leverage demographic information, medical history, lifestyle factors, and clinical biomarkers to accurately identify individuals at high risk of stroke. By incorporating advanced ML techniques and analyzing a large dataset of electronic health records, the model seeks to improve the accuracy and reliability of stroke risk assessment.

**REQUIREMENTS:**

* 1. **Software Required:**

**Programming Language:**

Python (version 3.x)

Integrated Development Environment (IDE):

Jupyter Notebook or JupyterLab for interactive development and experimentation.

**Python Libraries:**

NumPy: For numerical computing and array operations.

Pandas: For data manipulation and analysis.

Scikit-learn: For machine learning algorithms and model evaluation.

Matplotlib: For data visualization.

Seaborn: For statistical data visualization.

* 1. **Hardware Requirement:**

The hardware requirements for the project include a personal computer or laptop with the following minimum specifications:

- Processor: Intel Core i5 or equivalent

- RAM: 8GB or higher

- Storage: 256GB SSD or higher

- Operating System: Windows 10, macOS, or Linux

- Stable internet connection

**PROPOSED SOLUTION:**

1. **Proposed Design and Methodology:**

**Data Collection and Preprocessing:**

Gather a diverse dataset of electronic health records including demographic information, medical history, lifestyle factors, and clinical biomarkers.

Clean the dataset, handle missing values, and perform necessary preprocessing steps such as data normalization and feature scaling.

**Feature Selection and Engineering:**

Identify the most relevant features for stroke prediction using feature selection techniques such as correlation analysis, feature importance, and domain knowledge.

Engineer new features if necessary to improve the predictive performance of the model.

**Model Development:**

Implement and train multiple machine learning (ML) algorithms including logistic regression, decision trees, random forests, support vector machines, and gradient boosting machines using the scikit-learn library in Python.

Optimize hyperparameters for each algorithm to maximize predictive performance using techniques such as grid search or random search.

**Model Evaluation:**

Assess the performance of the developed ML models using appropriate evaluation metrics such as accuracy, sensitivity, specificity, area under the receiver operating characteristic curve and precision-recall curve.

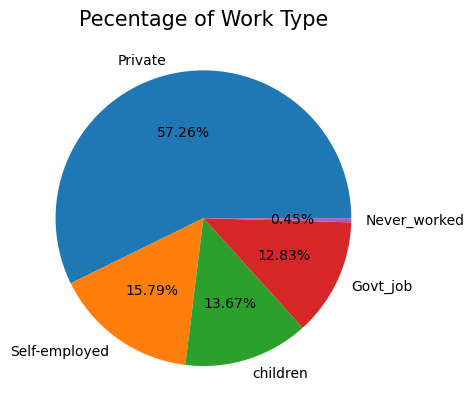
Perform cross-validation to ensure the generalizability and robustness of the model.

**Comparison and Selection of Optimal Model:**

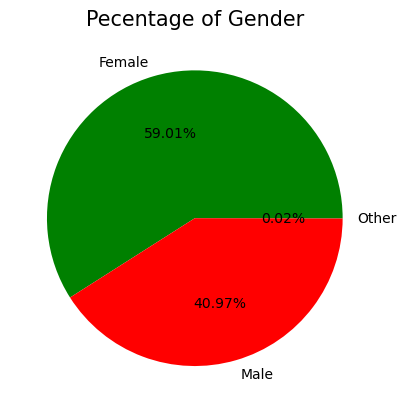
Compare the performance of different ML algorithms and select the most accurate and reliable model for stroke prediction based on the evaluation metrics.

**RESULTS:**

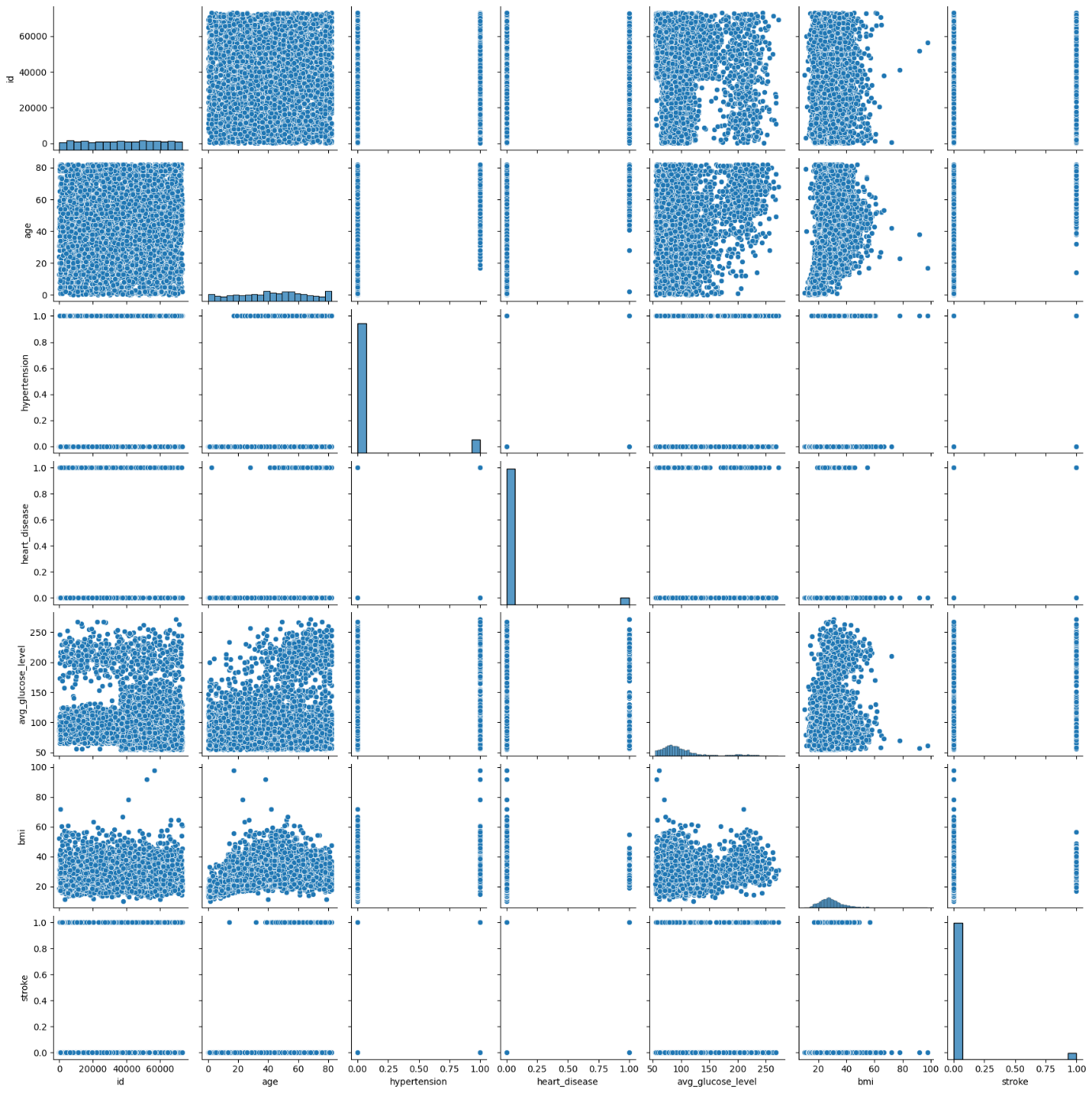
**Visualizations:**



This graph demonstrates which work type’s people suffer stroke the most



Based on the gender, people getting strokes



**Results on the Prediction Models: -**

**LOGISTIC REGRESSION:** TRAINING DATA: 96%

TESTING DATA:94%

**SVM (SUPPORT VECTOR MACHINE):** TRAINING DATA: 96%

TESTING DATA:94%

**DECISION TREE:** TRAINING DATA: 96%

TESTING DATA:94%

**REFERENCES:**

* Took the dataset from Kaggle
* Took help from sckitLearn documentation