A PROJECT REPORT
on

"Heart Stroke Risk Assessment: A Comparative Study of Supervised and Unsupervised Learning Models"

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER SCIENCE AND ENGINEERING
BY

KANDALA ABHIGNA	20051694
RASAGNA T	20051696
KANKANALA L S V S NAGAMANI CHARAN	2005028
VINAYA KAMAL D N	2005907

UNDER THE GUIDANCE OF

Pradeep Kandula

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024

May 2023

A PROJECT REPORT
on

"Heart Stroke Risk Assessment: A Comparative Study of Supervised and Unsupervised Learning Models"

Submitted to
KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN COMPUTER SCIENCE AND ENGINEERING

BY

KANDALA ABHIGNA	20051694
RASAGNA T	20051696
KANKANALA L S V S NAGAMANI CHARAN	2005028
VINAYA KAMAL D N	2005907

UNDER THE GUIDANCE OF

Pradeep Kandula

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA -751024

May 2023

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is to certify that the project entitled

"Heart Stroke Risk Assessment: A Comparative Study of Supervised and Unsupervised Learning Models"

submitted by

Kandala Abhigna	20051694
Rasagna T	20051696
Kankanala L S V S Nagamani Charan	2005028
Vinaya Kamal D N	2005907

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT Deemed to be University, Bhubaneswar. This work is done during the year 2022-2023, under our guidance.

Date: 10/12/2023

PROJECT GUIDE:
Pradeep Kandula
Professor, School of Computer Engg,
KIIT Deemed to be University

Acknowledgments

We are profoundly grateful to Professor **Pradeep Kandula** of **KIIT** for his expert guidance and continuous encouragement throughout to see that this project meets its target from its commencement to its completion.

Rasagna T
Kandala Abhigna
Kankanala L S V S Nagamani Charan
Vinaya Kamal D N

ABSTRACT

Our project aims to address the ongoing global health issue of heart strokes by focusing on innovative methods to improve risk assessment and early intervention. We introduce novel techniques by conducting a thorough comparative analysis of supervised and unsupervised learning models that are specifically tailored for assessing the risk of heart stroke.

Within the domain of supervised learning, our project utilizes annotated data to forecast the likelihood of experiencing a stroke, by making use of extensive patient data. In contrast, unsupervised models investigate the identification of patterns in data without relying on pre-established labels. The project thoroughly assesses these models by utilizing actual patient datasets, not only to measure their performance but also to emphasize their individual capabilities and constraints.

The primary goal is to provide healthcare professionals and researchers with practical knowledge to fully comprehend the practical effectiveness of supervised and unsupervised learning models in predicting heart strokes. These insights are essential for the development of targeted and efficient strategies to prevent strokes, enhance patient care, and mitigate the substantial impact of strokes on individuals and healthcare systems globally.

By adopting a project-oriented perspective, we aim to bridge the gap between theoretical research and practical implementation. The project findings offer nuanced guidance, enabling healthcare professionals to tailor interventions based on the strengths of each approach. This, in turn, enhances the precision and efficiency of preventive measures.

Our project goes beyond merely comparing models; it serves as a practical guide for healthcare stakeholders, offering tangible knowledge to enhance decision-making processes. As we work towards the project's objectives, we anticipate contributing to a paradigm shift in healthcare interventions, ushering in an era of targeted and informed strategies to effectively address the prevalence and impact of heart strokes on a global scale.

Contents

List of Figures

I. Introduction

A stroke is a severe medical incident that usually happens when there is a disruption or decrease in the blood flow to a particular area of the brain, resulting in a blockage of essential oxygen and nutrients. Consequently, this leads to the progressive deterioration of neurons in the brain.

The early signs of a stroke include a range of symptoms such as paralysis, sensory numbness, impaired vision, headache, and nausea. The World Health Organization (WHO) emphasizes the seriousness of the problem, classifying stroke as the second most significant cause of death worldwide, responsible for a significant 11% of all global deaths.

In order to tackle the intricacy of stroke prediction, a classification approach is utilized, considering multiple factors including gender, age, presence of hypertension, glucose level, and smoking status. This approach seeks to ascertain the probability of an individual encountering a stroke by analyzing a provided dataset. Through the utilization of machine learning algorithms, the dataset undergoes meticulous analysis, facilitating the production of forecasts for a novel entry.

When trying to predict the probability of someone having a stroke, a wide range of factors is taken into account. The factors encompass age, gender, mean glucose level, smoking habit, body mass index, occupation, and place of residence. The objective is to utilize appropriate, efficient, and trustworthy machine learning algorithms to guarantee the accuracy of predictions and the dependability of results.

It is crucial to not only forecast the probability of a stroke but also to pinpoint particular areas of heightened risk within the given parameters. This focused approach allows for the delivery of healthcare advice to individuals who have an increased probability of suffering from a stroke.

Our initiative aims to make a substantial contribution to proactive healthcare interventions by implementing a nuanced and comprehensive strategy. The ultimate goal is to decrease the impact and prevalence of strokes worldwide.

II. Basic Concepts/ Literature Review

2.1 BASIC CONCEPTS

2.1.1 Pandas

The Pandas library in Python is highly praised for its robust features in data manipulation and analysis. With diverse data structures and functions, it efficiently facilitates data manipulation, making it an indispensable tool for data science professionals, built upon the foundation of the NumPy library.

2.1.2 NumPy

NumPy, known as Numerical Python, is widely used for scientific calculations in Python. It offers an efficient and user-friendly approach to array and matrix manipulation, coupled with an extensive set of mathematical functions.

2.1.3 Matplotlib

Matplotlib, a frequently used Python data visualization library, presents an array of tools for crafting visually appealing and precise plots, charts, and figures. With extensive customization options, users can create sophisticated visual representations for diverse scientific and engineering purposes, utilizing various plotting functions.

2.1.4 Sklearn

Scikit-learn, or Sklearn, a renowned machine learning library integrated with Python's scientific computing stack, offers a broad spectrum of algorithms for supervised and unsupervised learning. This includes classification, regression, clustering, and dimensionality reduction techniques, with a design focused on user-friendly functionality and enhanced readability.

2.1.5 Tensorflow

TensorFlow, an open-source machine learning library developed by the Google Brain team, offers a comprehensive set of tools and frameworks to streamline the development and deployment of machine learning models. Suitable for applications

like image recognition, natural language processing, and speech recognition, TensorFlow is intentionally designed for high adaptability and expandability, making it well-suited for research and development.

2.1.6 Scikit:

The library provide a wide range of tools for data preprocessing, model choice, and evaluation, making it a complete platform for creating machine learning pipelines. The flexibility of scikit-learn is one of its main advantages.

2.2 Models Used:

2.2.1 K-Nearest Neighbors

The utilization of the K-Nearest Neighbors algorithm continues to be a suitable method for predicting the likelihood of a stroke in an individual. This approach relies on the concept of proximity or closeness to make precise classification decisions. When specific parameters, such as blood sugar levels or BMI in relation to age, are considered, a significant number of data points show close proximity to each other when plotted on a graph. This proximity corresponds to the presence or absence of strokes.

By analyzing a new data entry that includes variables such as age, body mass index (BMI), and blood sugar levels, it is possible to determine how closely it relates to a specific class. This allows for the accurate classification of the data entry.

2.2.2 SVM (Support Vector Machine)

The Support Vector Classifier technique is highly suitable for binary classification tasks, where there are two distinct classes in the target variable. In this specific situation, where the result is categorized as either "high probability of a stroke" or "low probability of a stroke," the Support Vector Machine (SVM) utilizes a method called feature mapping to convert the data into a space with more dimensions.

This transformation allows the Support Vector Machine (SVM) to accurately classify data points, even when they do not show linear separability. In cases where data points cannot be separated by a straight line, a hyperplane can be used as a separator to effectively distinguish between them.

2.2.3 Random Forest Classifier

While the decision tree classifier demonstrates high accuracy, the random forest algorithm, despite its computational complexity, may provide superior effectiveness. The approach utilizes an ensemble of multiple decision trees to achieve its objective. Each tree in the ensemble is built using a data sample that is randomly selected from the training set with replacement, a technique known as the bootstrap sample.

This technique is beneficial in both classification and regression scenarios. In this specific situation, a classification problem showcases the ability to efficiently handle large amounts of data, demonstrating greater robustness and accuracy in comparison to the decision tree algorithm.

2.2.4 CatBoost classifier

The CatBoost algorithm is a machine learning technique that employs gradient boosting on decision trees. This can be categorized as an ensemble learning algorithm. Throughout the training process, decision trees are constructed in a sequential manner. Successive trees are built with reduced loss, increased accuracy, and enhanced learning compared to the previous trees.

The CatBoost algorithm utilizes the boosting technique to progressively build decision trees, where each subsequent tree leverages the knowledge acquired by its predecessor. It offers an advantage over random forest as it excels in handling categorical features.

2.2.5 XGBoost classifier

XGBoost, like CatBoost, is an ensemble learning method known for its exceptional speed and performance, consistently outperforming other algorithms designed for supervised learning tasks.

The approach entails the consolidation of numerous decision trees, known as base learners, that demonstrate low bias and high variance traits. CART trees, or Classification and Regression Trees, display minor deviations from traditional decision trees.

2.2.6 LGBM (Light Gradient Boosting Machine) Classifier

LightGBM is a fast, distributed, and efficient gradient boosting framework that utilizes decision tree algorithms. It is frequently used for various machine learning tasks, such as ranking and classification.

The model's training speed is increased, and its performance closely mirrors that of XG Boost. The approach employs two methodologies: Gradient-Based One-Side Sampling (GOSS) to maintain information accuracy and Exclusive Feature Bundling (EFB) to reduce the number of effective features.

2.2.7 K Means Clustering

The K-Means algorithm is a widely used clustering technique in the fields of machine learning and data mining. Clustering is categorized as an unsupervised machine-learning methodology. The data is divided into separate clusters based on the categories of the dependent variable in the dataset.

This technique guarantees convergence and demonstrates the ability to adapt to new instances in the test dataset. However, in this specific case, it has been proven through empirical evidence that the mentioned algorithm is less accurate than other supervised learning algorithms.

2.2.8 Agglomerative Clustering

Agglomerative Clustering, categorized as a type of hierarchical clustering technique, remains a prominent unsupervised machine learning method in dividing the population into multiple clusters. Data points within the same cluster exhibit greater similarity, while those across different clusters show dissimilarity.

This study employed a bottom-up approach, initially treating each individual data point as an independent cluster. As one progresses up the hierarchical structure, clusters are gradually combined. The present implementation of K-means in the project reveals inferior accuracy compared to this alternative approach, yet it still falls short of matching the accuracy achieved by other supervised learning methods.

III. Problem Statement & Requirement Specifications

3.1 Problem Statement

The healthcare sector faces a critical challenge in accurately predicting and preventing heart strokes, which rank as the second leading cause of global mortality. Current predictive models often rely on conventional approaches, lacking the integration of advanced machine learning techniques. To address this, our project aims to pioneer innovative approaches by conducting a comprehensive comparative analysis of supervised and unsupervised learning models for heart stroke risk assessment.

The primary problem revolves around the need for more effective and precise predictive models to assess the likelihood of an individual experiencing a heart stroke. Conventional methods often fall short in providing nuanced insights, and there is a pressing requirement for advanced models that can leverage diverse patient datasets. The complexity of the human body and the multifaceted nature of stroke risk factors demand a sophisticated approach that goes beyond traditional methodologies.

This project seeks to bridge this gap by implementing a systematic methodology that involves data collection from real-world patient datasets, data preprocessing for optimal model training, and the deployment of both supervised and unsupervised learning models. The challenge lies in accurately evaluating the performance of these models, understanding their respective strengths and limitations, and deriving actionable insights for healthcare practitioners and researchers.

Furthermore, the lack of tailored healthcare recommendations based on individual risk factors emphasizes the need for a recommendation system. The intricacies of age, gender, average glucose level, smoking status, body mass index, work type, and residence type must be considered to provide targeted interventions for individuals at heightened risk.

The successful completion of this project will not only contribute valuable insights to healthcare stakeholders but also pave the way for a paradigm shift in stroke prevention strategies. By adopting a project-oriented perspective, we aspire to offer tangible knowledge that enhances decision-making processes and ultimately

mitigates the devastating impact of heart strokes on individuals and healthcare systems worldwide. This project sets the stage for further research and optimization of forecasting models, marking a significant step toward proactive and informed healthcare interventions.

3.2 Project Planning

The implementation of the heart stroke risk assessment project will follow a structured approach to ensure accuracy, reliability, and ethical considerations. The project will adhere to industry-standard coding practices, employing rigorous testing procedures to identify and rectify any bugs or errors. In addition to the primary focus on supervised learning models, the project will integrate unsupervised learning models to comprehensively assess their performance in heart stroke risk prediction.

To ensure project integrity, potential risks and limitations such as data availability and quality, model complexity, and unforeseen events impacting health records will be identified. Ethical considerations will be paramount, addressing the responsible use of predictive models in healthcare fundraising activities and ensuring compliance with regulations and laws.

The overall objective is to create accurate and reliable heart stroke prediction models using both supervised and unsupervised learning algorithms, following industry standards and subjecting the models to rigorous evaluation.

3.3 Project Analysis

In this novel healthcare project, the primary goal is to develop a robust prediction model for heart strokes using both supervised and unsupervised learning models. The project phases include data collection, planning, design, and performance evaluation. Thorough data analysis is crucial to identify relevant factors affecting heart stroke risk, encompassing parameters like age, gender, hypertension, glucose level, smoking status, body mass index, work type, and residence type.

For model analysis, parameters, hyperparameters, and performance metrics will be thoroughly measured, ensuring that the models don't overfit the training data. Visualization techniques will aid in understanding predicted events and comparing them with historical trends. The analysis will extend to project limitations, ethical considerations, and the impact of unforeseen events on heart-stroke predictions.

The testing phase will involve running test data samples to validate accuracy and reliability. Comparative analysis between supervised and unsupervised learning models will determine the model that performs optimally.

In essence, the project aims to develop an accurate heart stroke prediction model using advanced machine learning algorithms, presenting data-driven insights to healthcare practitioners and researchers.

3.4 Design Constraints

Design constraints play a crucial role in developing a robust heart-stroke prediction model. The following design factors are pertinent to this task:

Data Availability:

The accuracy of the prediction model relies on the availability of high-quality historical patient data. Addressing limitations in data quality, such as missing or incorrect entries, is imperative.

Model Complexity:

Limiting the complexity of the supervised and unsupervised learning models ensures efficient training within reasonable time frames and hardware capabilities. Overly complex models may hinder practical applicability.

Regulatory Restrictions:

Adherence to ethical and legal requirements, including data privacy and healthcare regulations, is paramount. The project design should prevent the misuse of the model for illegal activities.

Interpretability:

Results should be presented in a manner understandable to stakeholders, including healthcare practitioners, regulators, and individuals. Visualization techniques and clear decision-making patterns should facilitate interpretation.

Scalability:

The predictive model must be scalable to handle large datasets and real-time environments. Design considerations, such as optimization models, ensure efficiency and robustness in scalability.

By incorporating these design constraints, the project aims to develop an accurate, ethical, interpretable, and scalable heart stroke prediction model, contributing valuable insights to healthcare interventions.

IV. Implementation

4.1 Methodology

The methodology for our project, aimed at enhancing heart stroke risk assessment and early intervention through a comprehensive comparative analysis of supervised and unsupervised learning models, involves a systematic approach. The following steps outline the process:

Data Collection:

Gather real-world patient datasets containing comprehensive information, including age, gender, hypertension, glucose level, smoking status, body mass index, work type, and residence type. Ensure the datasets represent diverse demographics and cover a broad range of health profiles.

Data Preprocessing:

Clean the data by handling missing values, outliers, and inconsistencies. Format and convert raw data into a suitable structure for model training and evaluation. Apply techniques such as data normalization, scaling, and feature selection to enhance model performance.

Supervised Learning Model:

Utilize labeled data to train supervised learning models, employing algorithms such as logistic regression, decision trees, and support vector machines. Optimize hyperparameters based on performance metrics like accuracy, precision, recall, and F1 score.

Unsupervised Learning Model:

Explore unsupervised learning models, including clustering algorithms such as k-means or hierarchical clustering, for data-driven pattern discovery without predefined labels. Evaluate the effectiveness of unsupervised models in identifying patterns and potential risk factors.

Performance Evaluation:

Rigorously evaluate both supervised and unsupervised models using metrics such as accuracy, sensitivity, specificity, and area under the ROC curve. Compare the performance of the two approaches to elucidate their strengths and limitations.

Insights Generation:

Derive actionable insights from the evaluation results, providing a deeper understanding of the effectiveness of supervised and unsupervised learning models in predicting heart strokes.

Recommendation System:

Develop a recommendation system based on the identified risk factors and patterns to offer targeted healthcare recommendations for individuals at heightened risk.

Visualization:

Employ visualization techniques such as charts, bar charts, and histograms to present patterns and trends in heart-stroke risk factors.

Deployment:

Deploy the developed models and recommendation system to facilitate real-time predictions and interventions. By following this comprehensive methodology, our project aims to contribute valuable insights to healthcare practitioners and researchers, fostering a proactive approach to heart stroke prevention and care.

4.2 Data Selection:

The dataset provided is extensive, covering a broad spectrum of parameters and factors that may influence the probability of an individual having a stroke.

The key factors considered in determining the similarity are gender, average glucose levels, BMI, age, hypertension, history of heart disease, and smoking status. Nevertheless, it is crucial to recognize that both work and lifestyle exert a substantial influence.

This dataset is comprehensive and encompasses a wide range of parameters and factors that contribute to the likelihood of an individual experiencing a stroke.

These are the key attributes that contribute to determining the same: gender, average glucose levels, BMI, age, hypertension, history of heart disease, and smoking status. However, work and lifestyle factors also have an influence.

- **Data Description:**

id	Integer
gender	Categorical: male, female or other
age	Float
hypertension	Categorical: 1 or 0
heart_disease	Categorical: 1 or 0
ever_married	Categorical: 1 or 0

work_type	Categorical: Private, Self employed, Govt job, Never_worked, children
Residence_type	Categorical: Urban, Rural
avg_glucose_level	Float
bmi	Float
smoking_status	Categorical: never smoked, formerly smoked, smoking, unknown
stroke	Categorical: 1 or 0

Fig. 1: Data description

- **Model:**

  Given that the output variable is stroke, which can only take on the values of either yes or no, this problem can be classified as a binary classification problem. We employ logistic regression. Logistic regression is a regression technique that is applicable when the response variable is binary.

  In order to assess the efficacy of a logistic regression model, it is necessary to construct a confusion matrix. This matrix, presented as a 2 × 2 table, displays the predicted values generated by the model in comparison to the actual values derived from the test dataset.

V. Exploratory Analysis:

- **The number of samples with hypertension Based on stroke**

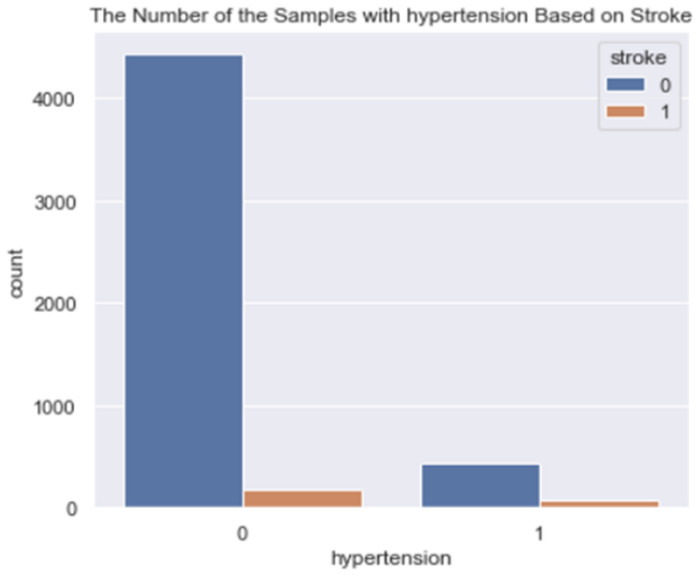


Fig. 2: The number of samples with hypertension Based on stroke

- **The number of samples with heart disease Based on stroke**

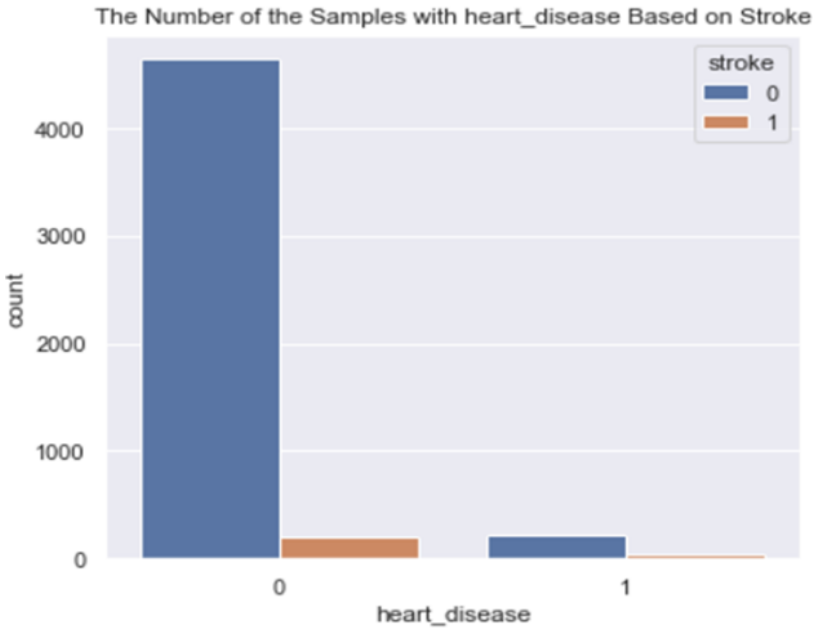


Fig. 3: The number of samples with heart disease Based on stroke

- **Distribution Plot - Avg Glucose Level**

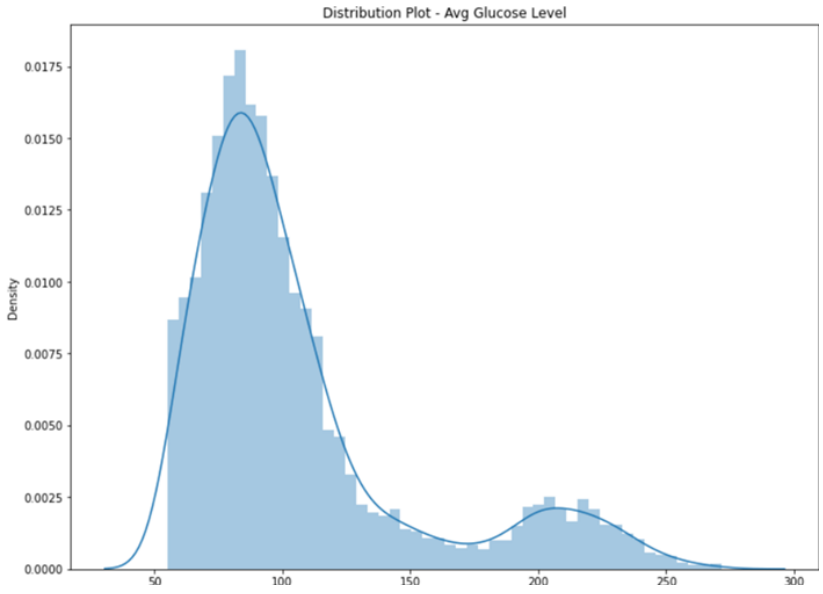


Fig. 4: Distribution Plot - Avg Glucose Level

- **linear model plot - Avg Glucose Level**

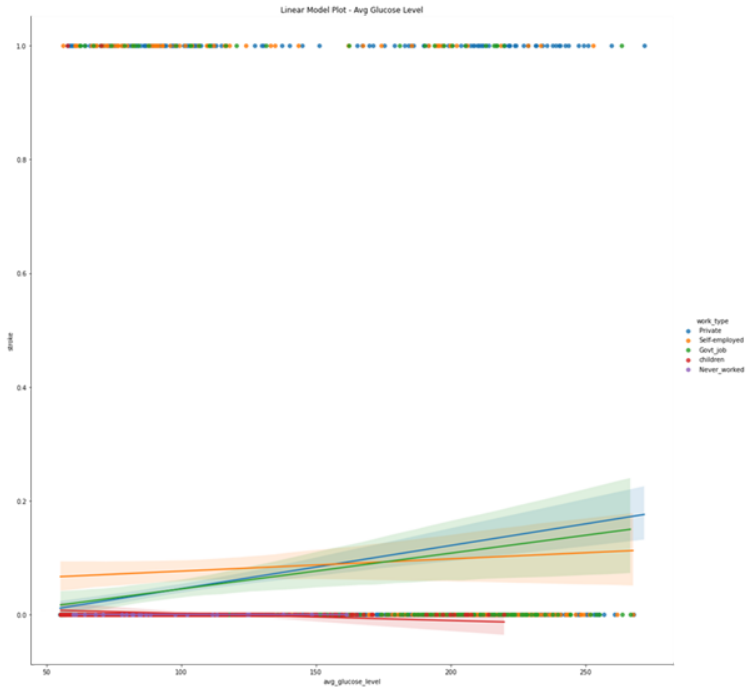


Fig. 5: linear model plot - Avg Glucose Level

- **Scatter plot**

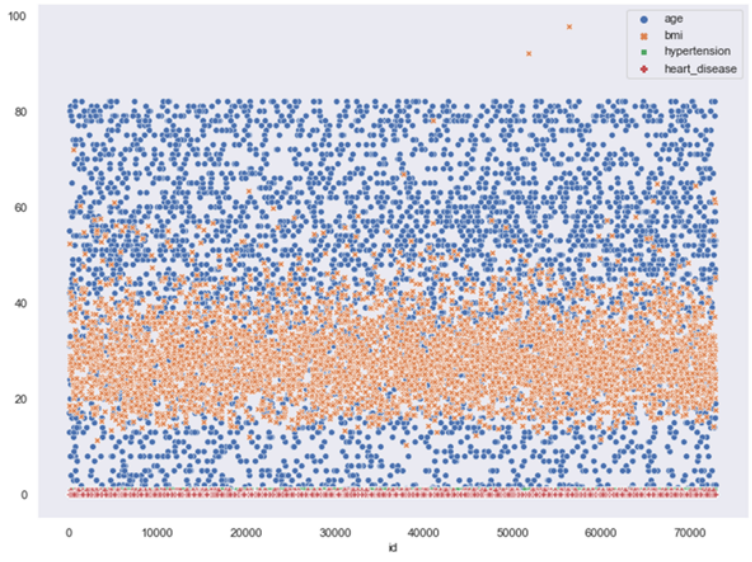
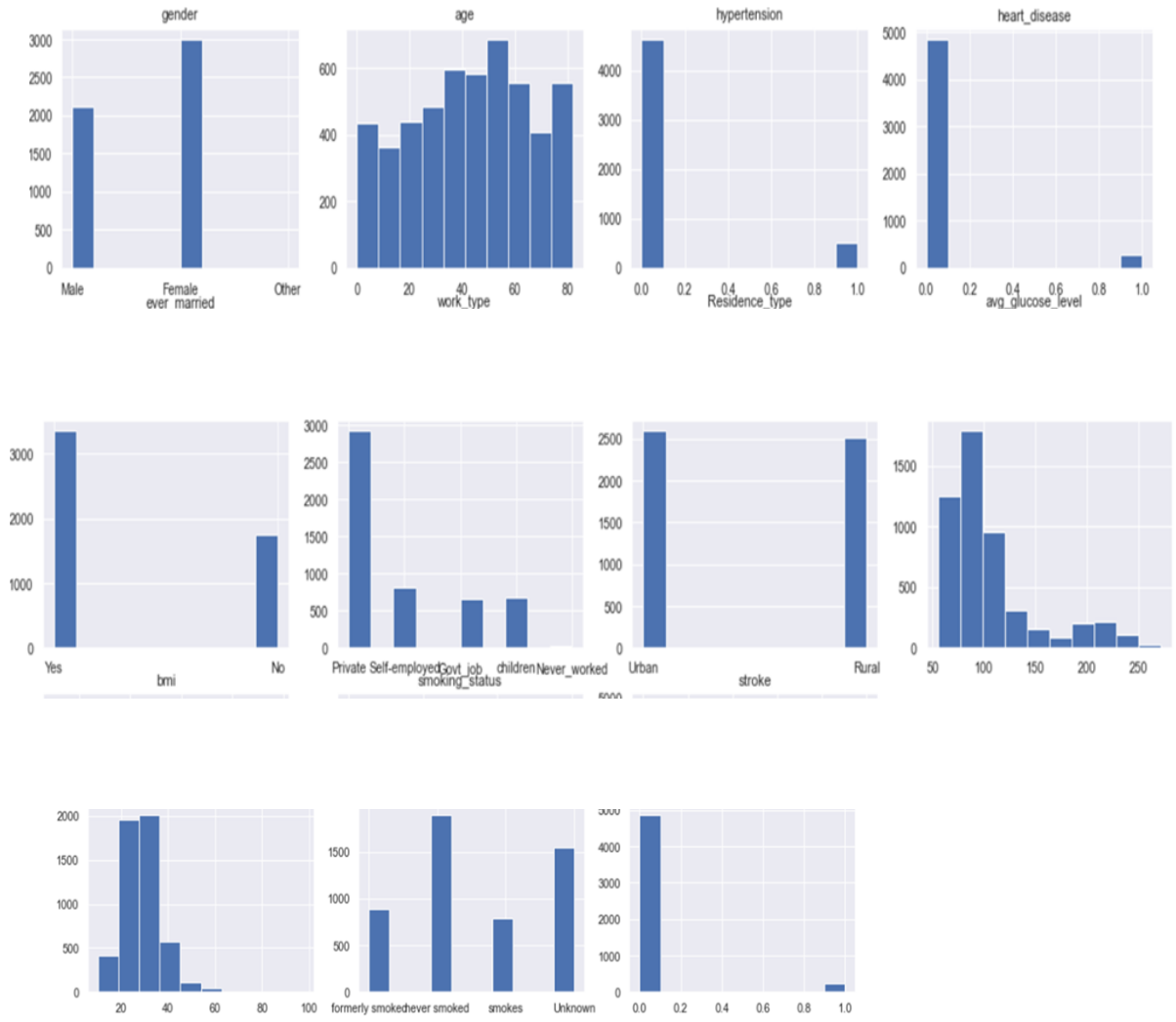


Fig. 6: Scatter plot

- **Pair Plots**



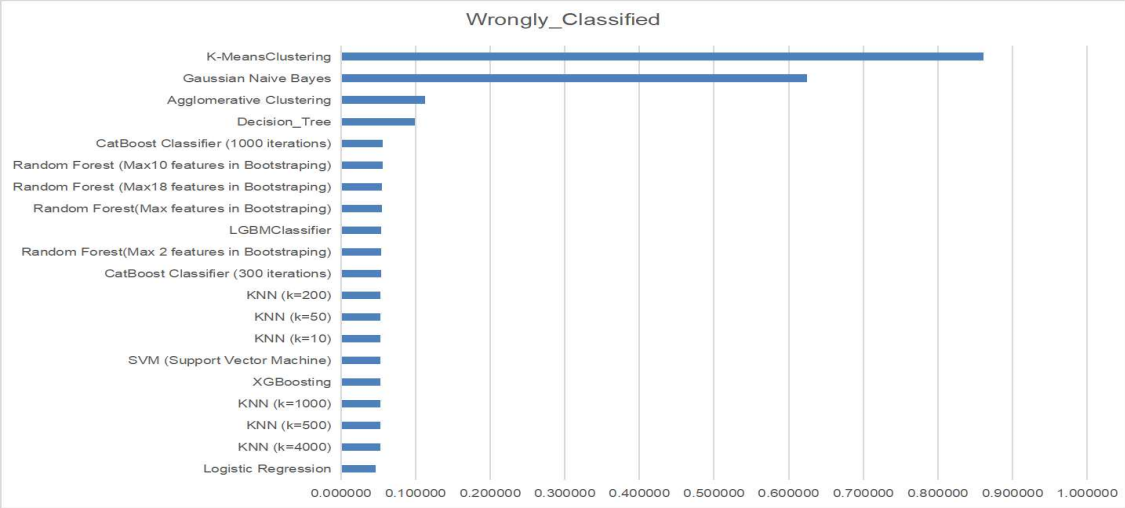
VI. Feature Engineering:

Addressing missing and null data: In the context of BMI, a considerable proportion of the records contain BMI values that are unclear or uncertain. As a result of the significant number of entries containing BMI values that are not empty, the empty spaces in the table have been filled with the average of these values. As a result, the model's ability to generate precise predictions is enhanced and streamlined.

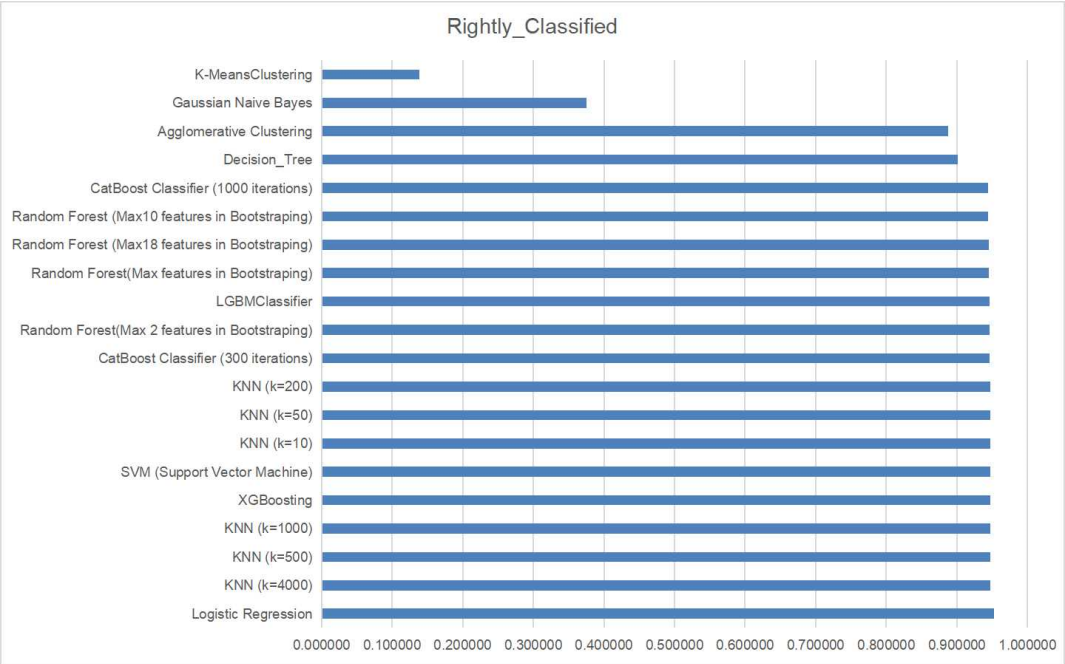
One Hot Encoding is a technique that transforms categorical data variables into a format suitable for machine learning algorithms. Adopting this format can enhance a model's capacity to generate precise forecasts and categorizations. During this stage, the categorical data is converted into binary vectors, which are a form of numerical data. This process generates additional features that are influenced by the total number of unique values present in the categorical feature.

VII. Combined Analysis:

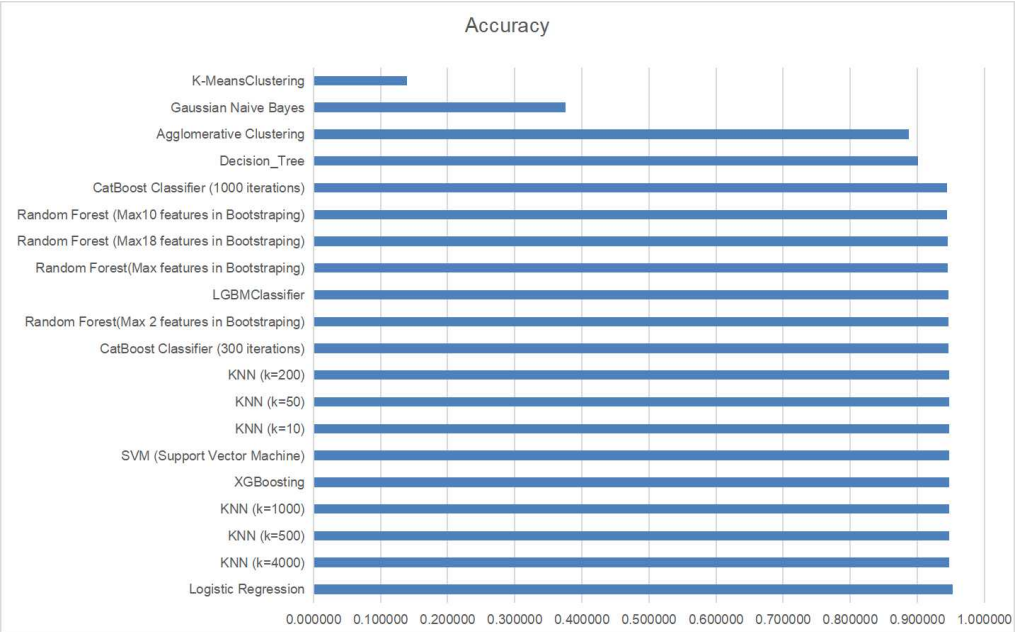
- **Wrongly classified:**



- **Rightly Classified:**



Accuracy:



- **Final Analysis**

	ML Classification Algo	Rightly_Classified	Wrongly_Classified	Accuracy	Recall	Specificity
5	Logistic Regression	0.958904	0.041096	0.958904	0.023810	0.998980
16	KNN (k=4000)	0.947162	0.052838	0.947162	0.000000	1.000000
11	KNN (k=10)	0.947162	0.052838	0.947162	0.000000	1.000000
15	KNN (k=1000)	0.947162	0.052838	0.947162	0.000000	1.000000
14	KNN (k=500)	0.947162	0.052838	0.947162	0.000000	1.000000
13	KNN (k=200)	0.947162	0.052838	0.947162	0.000000	1.000000
17	SVM (Support Vector Machine)	0.947162	0.052838	0.947162	0.000000	1.000000
12	KNN (k=50)	0.947162	0.052838	0.947162	0.000000	1.000000
3	CatBoost Classifier (300 iterations)	0.946184	0.053816	0.946184	0.037037	0.996901
6	Random Forest (Max 2 features in Bootstraping)	0.946184	0.053816	0.946184	0.037037	0.996901
8	Random Forest (Max 10 features in Bootstraping)	0.944227	0.055773	0.944227	0.018519	0.995868
1	LGBM Classifier	0.944227	0.055773	0.944227	0.018519	0.995868
9	Random Forest (Max 18 features in Bootstraping)	0.944227	0.055773	0.944227	0.000000	0.996901
7	Random Forest (Max 5 features in Bootstraping)	0.944227	0.055773	0.944227	0.018519	0.995868
2	CatBoost Classifier (1000 iterations)	0.944227	0.055773	0.944227	0.037037	0.994835
4	XGBoosting	0.941292	0.058708	0.941292	0.037037	0.991736
0	Decision_Tree	0.911937	0.088063	0.911937	0.129630	0.955579
19	Agglomerative Clustering	0.887476	0.112524	0.887476	0.500000	0.909091
18	K-Means Clustering	0.861057	0.138943	0.861057	0.537037	0.879132
10	Gaussian Naive Bayes	0.375734	0.624266	0.375734	0.981481	0.341942

Fig. 8: Analysis of all models

VIII. Conclusion:

Our investigation thoroughly explored innovative approaches to assess the risk of heart strokes by conducting a detailed analysis of both supervised and unsupervised learning models. The primary goal was to address the urgent need for enhanced early detection and prevention strategies concerning heart strokes, a major global health concern.

Our research findings reveal that supervised learning models, relying on labeled data, demonstrated impressive predictive accuracy in evaluating the risk of heart strokes. They showcased exceptional proficiency in leveraging patient data to make well-informed forecasts. A comprehensive comparative analysis was carried out on various supervised and unsupervised models, assessing their performance, computational efficiency, accuracy, and other relevant parameters.

In conclusion, the accuracy of models like KNN, SVM, Random Forest, and XGBoosting was notably superior to other models. Consequently, these supervised learning algorithms have been chosen over unsupervised learning models.

IX. Future Outlook:

Our project on heart stroke risk assessment unfolds promising avenues for future research and practical implementation, reflecting the evolving landscape of healthcare technology. The exploration of hybrid models, amalgamating supervised and unsupervised approaches, emerges as a potential breakthrough. This integration could yield more reliable and precise risk assessment tools, offering a holistic understanding of stroke likelihood.

Feature engineering, a critical aspect of supervised models, holds the key to further performance enhancement. By identifying and selecting pertinent features from patient data, predictive accuracy can be significantly improved. The project's future scope extends to the utilization of deep learning techniques on expanding healthcare datasets. Neural networks and other deep learning models have the potential to extract complex patterns from patient data, leading to more nuanced risk assessments.

Real-time risk assessment systems represent a transformative prospect. Continuous monitoring of patient health data using machine learning models could enable prompt interventions, significantly reducing the impact of heart strokes. The future also beckons advancements in personalized medicine, tailoring risk assessment and intervention plans based on individual patient profiles. This tailored approach could mark a significant stride in healthcare, ushering in an era of precision medicine.

Explainable AI emerges as a crucial consideration in the future of healthcare technology. Models offering interpretable explanations for their predictions are essential for ensuring transparency and reliability. This development is particularly pertinent in healthcare, where trust and understanding of AI-powered tools are imperative for widespread acceptance.

Clinical validation stands as an essential step in bridging the gap between research and practical application. Comprehensive clinical trials and validations are necessary to evaluate the practical efficacy of the developed models. Collaboration among data scientists, clinicians, and healthcare institutions is vital for this translational process, ensuring that the models align with real-world healthcare scenarios.

Ethical considerations continue to gain prominence as AI-based risk assessment tools become more prevalent. Addressing concerns related to privacy, consent, and responsible AI practices is imperative for the ethical deployment of these tools in healthcare settings. A focus on responsible AI research is crucial in navigating the ethical challenges posed by the intersection of technology and healthcare.

In summary, the potential for future research in utilizing machine learning models for heart stroke risk assessment is vast. The outlined future scope encompasses hybrid models, advanced feature engineering, deep learning techniques, real-time risk assessment, personalized medicine, explainable AI, clinical validation, and ethical considerations. Ongoing collaboration among researchers, healthcare professionals, and technology experts will be pivotal in fully harnessing the potential of these innovative methods, ushering in a new era of precision healthcare.

References:

1. Prediction of Heart Stroke Using Support Vector Machine Algorithm(Research Gate)
2. Comparative Analysis and Implementation of Heart Stroke Prediction using Various Machine Learning Techniques(IJRASET)
3. anaconda.org/conda-forge/vadersentiment
4. kaggle.com/datasets/Early_Stroke_Prediction_Using_Machine_Learning
5. anaconda.org/anaconda/nltk

INDIVIDUAL CONTRIBUTION REPORT:

"Heart Stroke Risk Assessment: A Comparative Study of Supervised and Unsupervised Learning Models"

KANDALA ABHIGNA
20051694

Abstract: To address the global heart stroke crisis, our study develops risk assessment and intervention strategies. In our thorough comparative research of heart stroke risk assessment supervised and unsupervised learning models, we use labeled data for predicting accuracy in supervised learning and explore data-driven patterns in unsupervised models. We give healthcare practitioners and researchers actionable insights from real-world patient datasets. Our study optimizes patient care and shifts healthcare practices to reduce heart strokes worldwide by bridging theory and practice.

Individual contribution and findings: As a member of the project group, my role was to contribute to the implementation part of the project, specifically using XGBoosting, Agglomerative Clustering, and Gaussian Naive Bayes. My primary contribution was to design and develop the mentioned models and integrate them with the rest of the project components. In terms of technical skills and experience, I gained a deep understanding of algorithms and their applications in data analysis. I used various libraries and tools, including Python, Keras, and TensorFlow, to design and develop the model. I also conducted several experiments to tune the hyperparameters and optimize the model's performance.

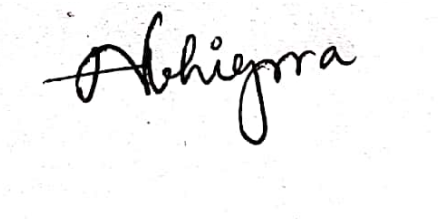
Individual contribution to project report preparation: My responsibility as a member of the project group was to contribute to the project's implementation in project report preparation. My primary contribution was to document the implementation details and performance evaluation of the unsupervised models.

Individual contribution to project presentation and demonstration: My responsibility as a team member was to contribute to preparing the slides of the models used and Implementation of the project.

Full Signature of Supervisor:

……

Full signature of the student:



INDIVIDUAL CONTRIBUTION REPORT:

"Heart Stroke Risk Assessment: A Comparative Study of Supervised and Unsupervised Learning Models"

RASAGNA T
20051696

Abstract: To address the global heart stroke crisis, our study develops risk assessment and intervention strategies. In our thorough comparative research of heart stroke risk assessment supervised and unsupervised learning models, we use labeled data for predicting accuracy in supervised learning and explore data-driven patterns in unsupervised models. We give healthcare practitioners and researchers actionable insights from real-world patient datasets. Our study optimizes patient care and shifts healthcare practices to reduce heart strokes worldwide by bridging theory and practice.

Individual contribution and findings: As a member of the project group, my role was to contribute to the data preprocessing part of the heart stroke prediction project using SVM (Support Vector Machine), and LGBM Classifier algorithm. My primary contribution was to design and develop the data preprocessing pipeline, which involved data cleaning, normalization, and feature engineering. I obtained a thorough understanding of data preparation methods and how they are used in time-series data analysis in terms of technical knowledge and experience. To clean and normalize the data, I utilized a variety of libraries and programs, including Python, Pandas, and NumPy. Additionally, I ran several tests to design fresh features that might enhance the SVM, and LGBM Classifier algorithm model's performance and accuracy.

Individual contribution to project report preparation: My responsibility as a team member was to contribute to the documentation of the Result Analysis and Comparative Analysis of the project.

Individual contribution to project presentation and demonstration: My responsibility as a team member was to prepare the slides for the Comparative Analysis of the project.

Full Signature of Supervisor: Full signature of the student:

…….  ……..

INDIVIDUAL CONTRIBUTION REPORT:

"Heart Stroke Risk Assessment: A Comparative Study of Supervised and Unsupervised Learning Models"

KANKANALA L S V S NAGAMANI CHARAN
2005028

Abstract: To address the global heart stroke crisis, our study develops risk assessment and intervention strategies. In our thorough comparative research of heart stroke risk assessment supervised and unsupervised learning models, we use labeled data for predicting accuracy in supervised learning and explore data-driven patterns in unsupervised models. We give healthcare practitioners and researchers actionable insights from real-world patient datasets. Our study optimizes patient care and shifts healthcare practices to reduce heart strokes worldwide by bridging theory and practice.

Individual contribution and findings: As a member of the project team, I was responsible for contributing to the project's implementation phase, with a particular focus on designing and developing the Random Forest and CatBoost Classifier model. Working closely with my colleagues, I made sure that the design aligned with the project's goals and requirements. This experience helped me gain valuable technical expertise, and I employed various tools and technologies to create a high-quality Random Forest and CatBoost Classifier model that met the project's standards for accuracy and effectiveness. Overall, my involvement in this project provided me with a unique opportunity to expand my knowledge and skills in the field of machine learning, and I take pride in being part of such a successful and fulfilling project.

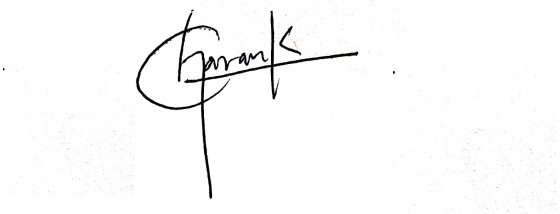
Individual contribution to project report preparation: My responsibility as a team member was to contribute to the documentation of the conclusion and the future scope of the project.

Individual contribution to project presentation and demonstration: My responsibility as a team member was to contribute to prepare the slides of the introduction and the problem statement used in the project.

Full Signature of Supervisor:

……

Full signature of the student:



INDIVIDUAL CONTRIBUTION REPORT:

"Heart Stroke Risk Assessment: A Comparative Study of Supervised and Unsupervised Learning Models"

VINAYA KAMAL D N
2005907

Abstract: To address the global heart stroke crisis, our study develops risk assessment and intervention strategies. In our thorough comparative research of heart stroke risk assessment supervised and unsupervised learning models, we use labeled data for predicting accuracy in supervised learning and explore data-driven patterns in unsupervised models. We give healthcare practitioners and researchers actionable insights from real-world patient datasets. Our study optimizes patient care and shifts healthcare practices to reduce heart strokes worldwide by bridging theory and practice.

Individual contribution and findings: In the project group, I had the responsibility of contributing to the implementation phase of the project. Specifically, my focus was on the design and development of the KNN and K-Means Clustering model. Collaborating with my team members, I ensured that the design was aligned with the project's objectives and requirements. Through this process, I acquired valuable technical skills and knowledge. I utilized a range of tools and technologies to successfully create the KNN and K-Means Clustering model, and worked diligently to ensure that it met the project's standards for accuracy and effectiveness. Overall, this project provided me with a unique opportunity to expand my knowledge and skills in the field of machine learning, and I am proud to have been a part of such a successful and rewarding endeavor.

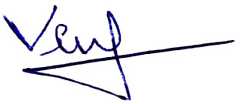
Individual contribution to project report preparation: My responsibility as a team member was to contribute to the documentation of the introduction and the basic concepts/literature review of the project.

Individual contribution to project presentation and demonstration: My responsibility as a team member was to contribute to prepare the slides of the algorithms being used in the project, the comparison of the algorithms being used and the basic concepts/literature review of the project.

Full Signature of Supervisor:

……

Full signature of the student:



Originality report

COURSE NAME
Project

STUDENT NAME
KANDALA ABHIGNA

FILE NAME
Report_finaldraft (7).docx

REPORT CREATED
Dec 10, 2023

Summary

Flagged passages	26	13%
Cited/quoted passages	0	0%

Web matches

ijrpr.com	22	10%
worldleadershipacademy.live	2	2%
leewayhertz.com	1	0.5%
medium.com	1	0.3%

1 of 26 passages
Student passage FLAGGED

is a record of bonafide work carried out by them, in the partial fulfillment **of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT Deemed to be**…

Top web match

is a record of bonafide work carried out by them, in the partial fulfilment **of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT Deemed to be**…

KIIT Deemed to be University - World Leadership Academy https://www.worldleadershipacademy.live/public/database/projects/2/1605450-1605334-1605371-1605525-1605450-KaranKhanna-AbhijitVasu-NilanjanGiri-SayanSaha-AshutoshJoshi.pdf

2 of 26 passages
Student passage FLAGGED