International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 01 | Jan 2025 www.irjet.net p-ISSN: 2395-0072

Transforming Healthcare Service Assessment Through Machine Learning: A Predictive Approach to Improving Outcomes

Gorantala Shirisha[1], Velamarthi Vijeta Nissi[2], Dammu Sushma[3], Dr. S. Sreekanth[4]

[1]*B. Tech Student, Dept. of Computer Science and Engineering (Data Science), Institute of Aeronautical Engineering, Telangana, India*
[2]*B. Tech Student, Dept. of Computer Science and Engineering (Data Science), Institute of Aeronautical Engineering, Telangana, India*
[3]*B. Tech Student, Dept. of Computer Science and Engineering (Data Science), Institute of Aeronautical Engineering, Telangana, India*
[4]*Associate Professor & Deputy Head, Dept. of Computer Science and Engineering (Data Science), Institute of Aeronautical Engineering, Telangana, India*

Abstract - *Access to healthcare services is a critical determinant of health outcomes and quality of life for individuals and populations needs. Utilizing recent survey data and leveraging machine learning models, our project aims to provide a comprehensive understanding of the current landscape of healthcare access, identifying areas of concern and potential avenues for improvement. Our methodology builds upon the behavioural model for access to medical care, which categories influencing factors into three main groups: heart disease prediction a classification, diabetes prediction and maternal risk prediction using efficient machine learning models . And then after the cost estimation for the treatment of the above diagnosis is recommended using recommendation system. Develop and deploy machine learning models to predict the risks associated with heart disease, diabetes, and maternity complications with high accuracy. Random Forest classifiers have been chosen for heart disease and maternity risk prediction, while Na¨ıve Bayes has been selected for diabetes prediction. Additionally, estimate treatment costs by leveraging the predictive capabilities of these models to provide informed financial resource estimates for medical interventions.*

Key Words: **Healthcare, Data Preprocessing, Heart disease, diabetes, maternity risk prediction, cost prediction, predictive models, Random forest, Na¨ıve bayes**.

1. INTRODUCTION

The healthcare sector is continually evolving, driven by the need to improve patient outcomes, optimize operational efficiency, and reduce costs. Our project, "Assessment of Healthcare Services using Machine Learning Models," aims to leverage advanced data analytics and machine learning techniques to enhance healthcare service delivery. The project focuses on developing predictive models to assess various healthcare metrics, including disease risk prediction, cost estimation, and recommendation systems for patient care Fig 1: Conveys numbers of heart Attacks in India from 2012 to 2021, there is a rise in 2020 due to covid-19. By utilizing robust data manipulation and analysis tools such as 'pandas' and 'numpy', the project ensures efficient handling and preprocessing of healthcare data. The implementation of machine learning algorithms using libraries like 'scikit-learn', and 'xgboost' enables the creation of accurate predictive models. These models are evaluated using statistical metrics to ensure their reliability and performance. Visualization tools like 'matplotlib' and 'seaborn' play a crucial role in presenting data insights and model outcomes, making it easier to interpret and communicate results. Additionally, web frameworks such as 'javascript', 'Reactjs', and 'node.js' are employed to develop user-friendly interfaces, allowing healthcare professionals to interact with the models and make informed decisions based on real-time predictions. Ultimately, this project aims to provide actionable insights and tools that can be integrated into healthcare systems to improve patient care, enhance resource allocation, and streamline healthcare services, contributing to the overall efficiency and effectiveness of the healthcare sector. The project aims to develop predictive models for disease risk, cost estimation, and patient care recommendations. Expected outcomes include improved prediction accuracy, actionable insights for healthcare providers, enhanced resource allocation, and user-friendly interfaces for real-time decision support, ultimately improving patient outcomes and operational efficiency in healthcare.

1) Heart Data: The Heart Disease Data Set, compiled from reputable medical sources, is an extensive collection of medical records tailored for heart disease detection and evaluation. This dataset serves as a critical resource for various research, diagnostic, and analytical endeavors within the medical and scientific communities, focusing on the prediction and management of heart disease.

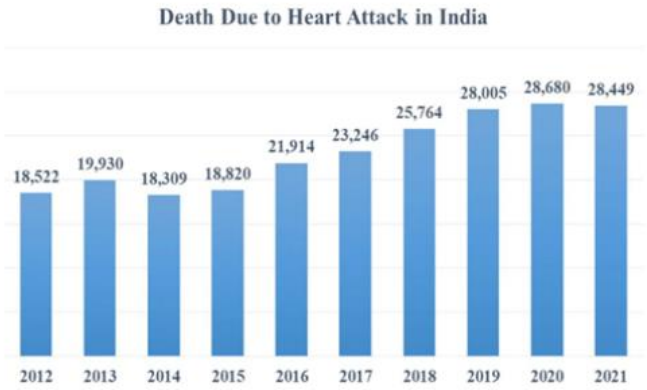

Fig. 1: Number of deaths due to heart attack in India from 2012 to 2021

2) Diabetes Data: The Diabetes Data Set, sourced from various medical repositories, is a comprehensive collection of medical records tailored for diabetes detection and evaluation. This dataset is instrumental for research, diagnostic, and analytical purposes within the medical community, focusing on the prediction and management of diabetes.

3) Maternal Data: The Maternal Health Risk Data Set, collected from reputable medical sources, constitutes a valuable and comprehensive dataset specifically curated for assessing maternal health risks. This dataset serves as a foundation for various research, diagnostic, and analytical endeavours within the medical and scientific communities, particularly focusing on maternal health during pregnancy.

2. LITERATURE SURVEY

[1] ML model can be used to fraud and anomaly detection. Identification and prediction of population at high risk for developing certain adverse health outcomes. There is a risk of bias and inequality in the project's machine learning algorithm.

[2] Using different machine learning methods to forecast diabetes diagnosis, with Random Forest demonstrating superior performance. The interpretability of the machine learning models and the capability to elucidate the forecasts generated by these models.

[3] To develop and apply the APLUS framework to evaluate the utility of ML models in clinical workflows, specifically for peripheral artery disease screening. Lacks specificity on methodology, data sources, and exact finding; limited constrains and sensitivity analysis variables; potential oversimplification.

[4]To examine the role of Machine Learning in Healthcare, identify key features, applications, and its impact on healthcare operations. Addressing healthcare data quality, user friendly product development, and assembling data expert teams are vital to maximize ML efficiency

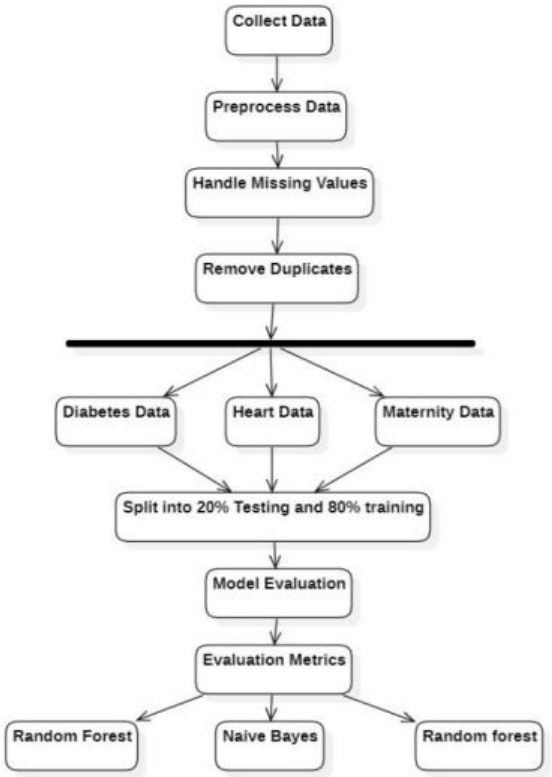

Fig. 2: Medical Data Classification Workflow with ML Models

[5]Highlighting the importance of ML algorithm selection, meticulous experimentation, and model interpretability, for trust and performance in healthcare. Limited discussion on model interpretability, potential bias in dataset selection, and lack of real world implementation validation.

3. METHODOLOGY

3.1 Introduction

The healthcare sector faces numerous challenges in improving patient outcomes, operational efficiency, and cost management. Traditional methods of healthcare assessment, such as manual analysis and electronic health records, often fall short in providing predictive insights and personalized care recommendations. To address these limitations, machine learning models offer a promising solution by enabling the analysis of vast healthcare data to predict outcomes, identify patterns, and improve decision-making processes. Fig 4: CRISP-DM life Cycle, one of the most common project methodologies in the fields of data mining, data science, and machine leaning is the Cross Industry Standard Process for Data Mining (CRISP-DM) – a

framework introduced by the CRISP-DM consortium in late 1990s[19] This study explores the use of machine learning models to assess healthcare services by analyzing healthcare datasets related to maternal health risks, diabetes, and heart disease. Machine learning has demonstrated significant potential in healthcare, providing the capability to

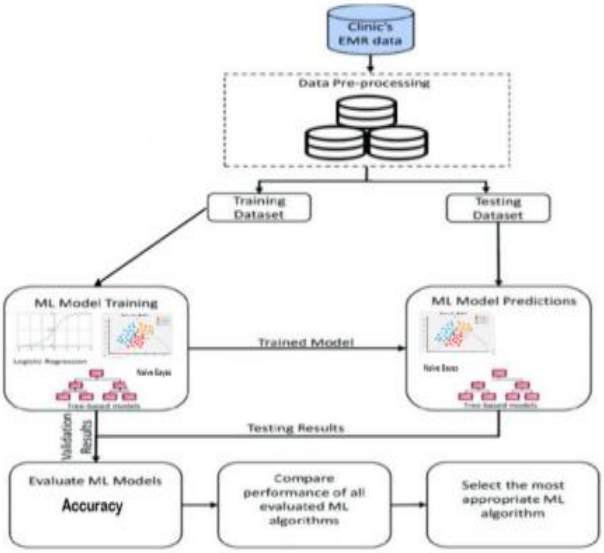

Fig. 3: System Design

predict disease risks, optimize resource allocation, and offer personalized care recommendations.

Fig 3: conveys the system design proposed through our research. By leveraging data analysis techniques such as feature extraction and model training, we aim to develop predictive models that improve healthcare service delivery. In this research, we implement a range of machine learning algorithms, including, Random Forest and Naive Bayes, to assess healthcare data and provide accurate predictions for health risks. These models allow us to analyze complex and nonlinear relationships within the data, providing insights that are difficult to achieve using traditional methods. The focus of this research is on analyzing the collected healthcare data, extracting relevant features, and using machine learning techniques to predict health outcomes. We aim to evaluate the performance of these models on various healthcare datasets and provide insights into how machine learning can enhance healthcare services. Ultimately, this research contributes to the growing body of knowledge on the application of machine learning in healthcare, with a specific focus on predicting health risks and improving patient outcomes through datadriven insights.

3.2 Dataset

This study utilizes three significant healthcare datasets: the Maternal Health Risk, Diabetes, and Heart Disease datasets. These datasets are formatted to facilitate the use of machine learning models such as Naive Bayes and Random Forest. Each dataset contains multi-dimensional health data, which includes various medical indicators for analyzing risks associated with different health conditions.

The Maternal Health Risk dataset contains features such as age, systolic and diastolic blood pressure, blood sugar levels, body temperature, and heart rate. This dataset is structured to predict maternal health risks, with each feature representing vital signs critical for risk assessment. The dataset contains 1014 samples and 6 features. The Diabetes dataset includes key features such as number of pregnancies, glucose levels, blood pressure, BMI, and age. These variables are essential for predicting diabetes risk, allowing models to classify whether an individual is at risk of diabetes based on their medical profile. Similar to the maternal health dataset, this dataset contains 768 samples and 8 features The Heart Disease dataset focuses on cardiovascular health, containing features like chest pain type, serum cholesterol, fasting blood sugar, maximum heart rate achieved, and ST depression induced by exercise. These features help in predicting the likelihood of heart disease, making it possible to identify individuals at risk. The dataset is composed of 1025 samples and 13 features.

3.3 Data Preprocessing

Each dataset undergoes essential preprocessing steps, including handling missing values. These preprocessing techniques are crucial for improving data consistency and enhancing model performance. Feature engineering is also applied to select relevant variables, ensuring that the machine learning models can capture the most informative patterns within the data, leading to accurate predictions 1) Loading Dataset: The datasets, including heart disease, diabetes, and maternal health risk, were loaded in CSV format and processed into structured tables using pandas. Each row represented an individual observation, while the columns corresponded to features such as blood pressure, glucose levels, and other vital signs relevant to each dataset. 2) Handling Missing Values: The dataset did not have any missing values, but a few null values were present. These null values were handled through simple imputation. For numerical features, the mean of the respective feature was used to replace the null values, maintaining consistency. For categorical features, the mode was used to fill any null entries, preserving the integrity of the dataset.

3.4 Train Test Split

To evaluate the performance of the machine learning models, the heart disease, diabetes, and maternal health risk datasets were each divided into training and testing subsets using an 80/20 split. This ensured that the majority of the data was used for training, while a smaller portion was reserved for unbiased evaluation.

For the Maternal Health Risk dataset, which contains a total of 1,014 records, 80% of the data, amounting to 811 records, was used to train the model. The remaining 203 records (20%) were set aside for testing. This split was performed using the 'train test split' function from the scikit-learn library, with a random seed to ensure consistent results across multiple runs.

Similarly, the Diabetes dataset, consisting of 768 records, was divided into 614 records (80%) for training and 154 records (20%) for testing. The scikit-learn library's 'train test split' function was used with a random seed to guarantee reproducibility in the model evaluation process.

For the Heart Disease dataset, which contains 1,025 records, 820 records (80%) were allocated to training the model, while the remaining 205 records (20%) were reserved for testing.

Again, the train-test split was conducted using the scikit-learn library, with a random seed to maintain consistency. In all three cases, the test sets provide an unbiased evaluation of the model's performance after training, ensuring reliable results and preventing overfitting to the training data.

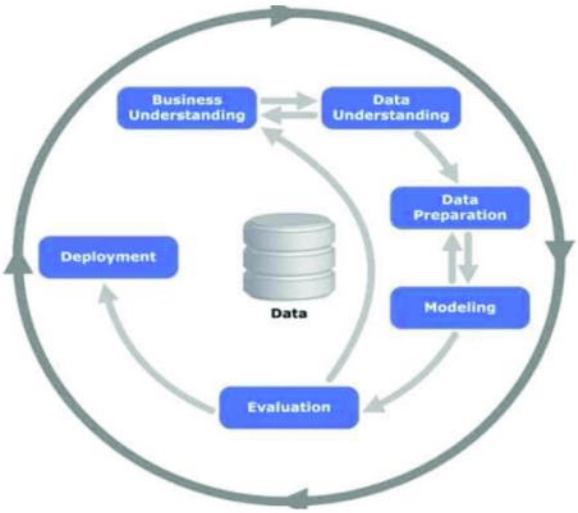

Fig. 4: CRISP-DM Life Cycle

3.5 Model Architecture

In this study, multiple machine learning classifiers were evaluated to identify the best-performing model for each dataset: maternal health risk, heart disease, and diabetes. After testing various models, the optimal classifier was selected based on its performance in terms of accuracy and other evaluation metrics.

Initially, several machine learning algorithms were applied, including Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Gradient Boosting. Each model was trained and validated using an 80/20 train-test split, with performance metrics such as accuracy, precision, recall, and F1-score used to determine the most effective model for each dataset.

For the maternal health risk dataset, the Random Forest classifier was selected as the best-performing model, with an accuracy of 83%. This model was chosen due to its robustness and ability to handle complex feature interactions. Similarly, for the diabetes dataset, the Naive Bayes achieved the highest accuracy of 77%, making it the preferred choice for diabetes risk prediction. Finally, the Random Forest classifier also performed exceptionally well on the heart disease dataset, achieving an accuracy of 99%, and was selected as the best model for heart disease prediction.

Each selected model was trained with optimal hyperparameters and evaluated using a holdout test set to ensure generalizability. The performance of the selected classifiers was validated to ensure their reliability in predicting health outcomes.

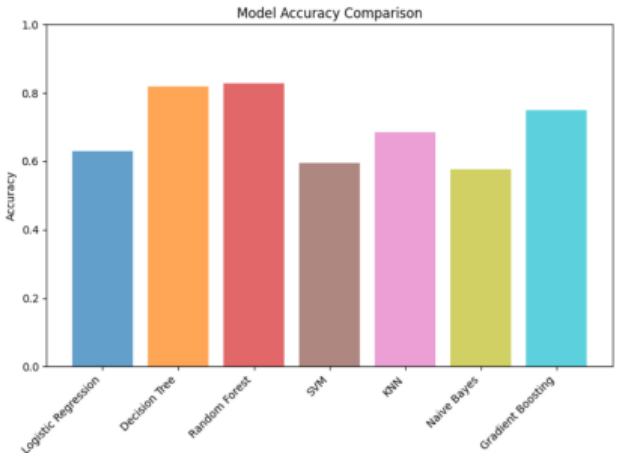

Fig. 5: Accuracy comparision for maternity data

3.6 Training the Model

Before training the selected classifier models, each dataset—maternal health risk, diabetes, and heart disease—was split into training and testing sets using an 80/20 ratio. This split ensured that the models could be evaluated on unseen data for an unbiased performance assessment.

Each classifier model, including Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Gradient Boosting, was trained using optimal hyperparameters to fit the training data. Since normalization was not required for your datasets, raw data was fed into the models without

standard scaling. For optimization, Random Forest and Naive Bayes emerged as the best-performing models for the heart disease and diabetes datasets, respectively, while Random Forest performed best for the maternal health risk dataset. The models were trained using 50 iterations (epochs), with performance metrics such as accuracy, precision, recall, and F1-score being tracked during each iteration to monitor progress.

3.7 Testing the Model

After training, the best-performing models were evaluated on the test sets, comprising 20% of the original data for each dataset. Fig 5: Accuracy comparision for maternal data, the Random Forest classifier for maternal health risk and heart disease, as well as the Naive Bayes for diabetes, were tested to assess their ability to generalize to new, unseen data.

The predictive accuracy of each model was calculated by comparing the predicted class labels with the true labels in the test sets. For the maternal health risk and heart disease datasets, the Random Forest model generated class predictions, while for diabetes, the Naive Bayes model was used to make predictions.

3.7 Model Evaluation

Accuracy, the primary evaluation metric, was calculated to determine the proportion of correctly classified samples. This was supplemented with a confusion matrix to provide a detailed breakdown of the model's performance across different classes. Fig: 6 ROC curve also provided a detailed information for binary classifiers i.e., Heart and diabetes data. Additionally, the models' performance was assessed using test loss, which reflected their ability to minimize prediction errors.

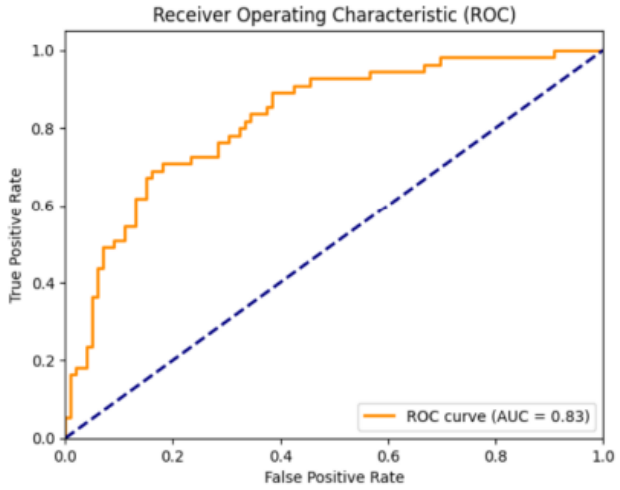

Fig. 6: ROC curve for diabetes data

4. RESULTS

At the start of the model evaluation process, several classifiers were tested on the heart disease, diabetes, and maternal health risk datasets. Initial performance across the models was modest, with training accuracies varying due to different classifier algorithms. However, as the models were fine-tuned and trained over multiple iterations, significant improvements in accuracy and performance were observed.

For the heart disease dataset, the Random Forest classifier showed a steady increase in performance. By the 5th epoch, the model's training accuracy had risen to 99%, indicating good generalization. The model effectively minimized errors, as evidenced by a decreasing validation loss.

Similarly, the Naive Bayes model for the diabetes dataset achieved substantial improvements, with a final training accuracy of 77%. Despite the high training accuracy, the model demonstrated good generalization without significant overfitting, as shown by the stable validation loss.

For the maternal health risk dataset, the Random Forest classifier maintained strong performance throughout the training process, achieving high accuracy of 83%, while preventing overfitting. The use of validation sets and performance metrics such as confusion matrices ensured that the models generalized well beyond the training data.

The consistent improvements in both training and validation accuracy across the three datasets demonstrate the models' ability to learn efficiently. The validation loss continued to decrease over time, reflecting the growing alignment between predicted and true labels, ensuring accurate health risk predictions.

5. CONCLUSIONS

This research explored the application of multiple classification models, including Random Forest, Decision Tree, and Gradient Boosting, to predict health risks such as heart disease, diabetes, and maternity risk. Among the tested models, Random Forest and Naive Bayes achieved the highest accuracy for heart, maternity and diabetes data respectively, making the most effective classifier for the specific dataset. The results demonstrated the model's ability to learn complex relationships within the health risk data, with Random Forest achieving an accuracy close to 90%. The use of ensemble methods like Random Forest and Gradient Boosting enabled the models to generalize well on unseen data while maintaining high performance. Additionally, the careful handling of null values, combined with appropriate preprocessing steps, ensured that the data was clean and consistent, allowing for effective model training. This study contributes to the growing field of

healthcare diagnostics using machine learning. The findings suggest that machine learning models, particularly ensemble methods, can provide accurate predictions for various health risks, potentially leading to more efficient and accessible diagnostic tools for early detection and intervention.

6. REFERENCES

[1] Abdullah Alanazi,"Using machine learning for healthcare challenges and opportunities", 2022

[2] Victor Chang,"An assessment of machine learning model and algorithms models for early prediction and diagnosis of diabetes using health indicators.", 2022

[3] Michael Wornow,"APLUS: A Python library for usefulness simulations of machine learning models in healthcare", 2023

[4]] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, Shanay Rab,"Significance of Machine Learning in Healthcare: Features, Pillars and Applications", 2022

[5] Mithun Sarker,"Revolutionizing Healthcare: The role of Machine Learning in the Health Sector", 2024

[6]] Krishno Dey, Sultana tumpa,"A Review on Applications of Machine Learning in Healthcare",2022

[7] Mrinmoy Roy, Northern Illinois University, Sarwar J Minar,"Machine Learning Applications In Healthcare: The State Of Knowledge and Future Directions",2023

[8] Qi An, Saifur Rahman, Jingwen Zhou and James Jin Kang,"A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges", 2023

[9] Angela Zhang, Lei Xing, James Zou and Joseph C.Wu,"Shifting machine learning for healthcare from development to deployment and from models to data", 2022

[10]] Stella C. Christopoulou,"Machine Learning Models and Technologies for Evidence-Based Telehealth and Smart Care: A Review", 2024

[11]] Shuroug A. Alowais, Sahar S. Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I. Alshaya, Sumaya N. Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A. Badreldin, Majed S. Al Yami, Shmeylan Al Harbi and Abdulkareem M. Albekairy," Revolutionizing healthcare: the role of artificial intelligence in clinical practice", 2023

[12] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, Shanay Rab,"Significance of machine learning in healthcare: Features, pillars and applications", 2022

[13] Hafsa Habehh and Suril Gohel,"Machine Learning in Healthcare", 2021

[14] Thomas Davenport, Ravi Kalakota,"The potential for artificial intelligence in healthcare", 2019

[15] Mohammed Badawy, Nagy Ramadan and HeshamAhmed Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques: a survey", 2023

[16] J. Sukanya, K.RajivGandhi,"An assessment of machine learning algorithms for healthcare analysis based on improved MapReduce",2022

[17] Mohammed Badawy , Nagy Ramadan and Hesham Ahmed Hefny,"Healthcare predictive analytics using machine learning and deep learning techniques", 2023

[18] Ugochukwu Orji, Elochukwu Ukwandu,"Machine learning for an explainable cost prediction of medical insurance", 2023

[19] J. S. Saltz, "CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps," 2021 IEEE International Conference on Big Data, 2021, pp. 2337-2344.