International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Predictive Model of Disease Prognosis in Healthcare

Prof.Milind Kamble
Department of *Electronics and Telecommunication*
Vishwakarrma *Institute of Technology,Pune*
Milind.kamble@vit.edu

Soumil Joshi
Department of *Electronics and Telecommunication*
Vishwakarrma *Institute of Technology,Pune*
soumil.joshi211@vit.edu

Shravani Karbhajane
Department of *Electronics and Telecommunication*
Vishwakarrma *Institute of Technology,Pune*
shravani.karbhajane21@vit.edu

Rutuja Khandagale
Department of *Electronics and Telecommunication*
Vishwakarrma *Institute of Technology,Pune*
rutuja.khandagale22@vit.edu

***Abstract*—The Examination and modeling of a healthcare dataset to predict disease prognoses at the patient level. The dataset encompasses diverse features such as age, sex, general health, checkup history, exercise habits, smoking history, and the presence of diseases like heart disease, skin cancer, other cancers, diabetes, and arthritis. Exploratory data analysis reveals significant variations in disease distribution, emphasizing the influence of health and lifestyle factors on disease risk. Missing data underscores the need for effective imputation methods in healthcare predictions. Machine learning, particularly XGBOOST,Random Forest, Desion Tree is employed for disease prognosis prediction, employing rigorous training and validation through a train-test split strategy. Evaluation metrics include confusion matrix, ROC curve, and Precision- Recall curve, offering a comprehensive view of model performance. The study highlights the significance of multiple metrics in imbalanced classification tasks, showcasing machine learning's potential in healthcare for patient risk assessment, healthcare planning, and clinical strategy formulation. Future work may focus on model refinement, class balancing strategies, and incorporating additional patient data for more accurate disease prognosis predictions.**

Keywords—Correlation,Healthcare,XG-Boost,Covaariance,Univariate.

I.INTRODUCTION

Dynamic prediction models have emerged as a powerful tool in healthcare research, allowing for the utilization of evolving data to enhance the accuracy of predictions over time. The ability to adapt to changing landscapes and incorporate new information sets dynamic models apart from traditional static models. By continuously updating past knowledge with current data, these models offer a more nuanced understanding of health outcomes and enable real-time adjustments for improved predictions.In recent years, there has been a growing interest in the development and validation of dynamic prediction models to address the complexities of healthcare scenarios. Researchers have explored various approaches, including discrete model updating and functional varying coefficient models, to capture the dynamic nature of health data. While these methods show promise, challenges in model validation and evaluation persist, requiring innovative solutions to ensure the reliability and effectiveness of dynamic prediction models[1].The study focuses on developing machine learning predictive models for Coronary Artery Disease (CAD) using diagnostic datasets from hospitals in Kano State, Nigeria. With the increasing prevalence of non- communicable diseases like CAD, accurate predictive models are essential for early detection and intervention to improve healthcare outcomes[2].The paper discusses the challenges and importance of accurate cardiovascular disease risk prediction models. It highlights concerns regarding the validity of existing models, particularly when applied to diverse populations. Inaccurate risk estimation can lead to both overestimation and underestimation of risks, impacting healthcare costs and patient outcomes. The study emphasizes the need for population-specific models to improve risk classification and inform policymakers and health authorities effectively. Efforts are underway to develop more precise and tailored risk prediction tools through

extensive observational data and validation studies in various regions[3].

II.LITERATURE SURVEY

The smart town initiative integrates artificial intelligence, data analysis, and communication technologies to enhance sustainability. Within this framework, the focus is on predicting and diagnosing diseases in smart healthcare. Utilizing advanced data analysis techniques, such as machine learning algorithms, enables early disease detection and improved patient care. However, challenges arise from incomplete medical data and regional variations in disease patterns, affecting prediction accuracy. To address this, a system is proposed that predicts disease occurrences and recommends treatment sequences based on waiting times, demonstrated through a case study on cerebral infarction. diagnosis are also discussed[4].

Data mining in healthcare has immense potential for uncovering hidden patterns in medical data, aiding in disease diagnosis and prognosis. However, medical data is often complex, large, and distributed, making manual processing by physicians impractical. Automated Disease Predictive Models are thus needed to accurately predict diseases with minimal effort. The use of data mining in the Indian healthcare sector shows promising growth. This paper provides a systematic overview of Data Mining Techniques, their Applications, and the current state of healthcare in India[5].

Given the escalating medical costs and burden of cardiovascular disease, we devised ASHRO, an integrated healthcare resource consumption predictive model. It incorporates patient behaviors and assesses its connection with clinical outcomes. Our study drew data from extensive sources including health insurance claims and health check-ups, focusing on patients hospitalized for circulatory system diseases. Utilizing random forest learning for adjustment and multiple regression analysis for ASHRO score construction, we evaluated model performance through discrimination and calibration tests. Comparing mortality rates over 48 months, ASHRO showed significant associations with healthcare resource utilization and patient outcomes[6]

In recent years, predictive modeling in healthcare has surged in popularity, driven by increasingly sophisticated tools with higher accuracy. We illustrate this through a case study showcasing how artificial intelligence (AI) enhances predictive modeling, improving healthcare quality and efficiency. MEDai, Inc. provides analytical tools for predictive modeling, utilized by Sentara Healthcare to identify members who could benefit from preventive measures. Unlike traditional methods like rule- based systems or regression techniques, AI accommodates the complexity of medical data, yielding significantly higher accuracy. Comparing R2 values, a measure of predictive model accuracy, traditional techniques typically range from 0.10 to 0.15, while AI techniques implemented at Sentara achieve an R2 value of 0.34. These predictions inform data mining and analysis, enabling examination at subgroup or individual member levels to identify risk factors and optimize medical caredelivery[7].

Healthcare analytics increasingly relies on predictive modeling using electronic health records (EHRs). To streamline this process, we developed PARAMO, a platform that supports key tasks like cohort and feature construction, cross-validation, feature selection, and classification. PARAMO constructs a task dependency graph, schedules tasks in a topological order, and executes them in parallel using Map-Reduce. We tested PARAMO on datasets from Geisinger Health System, Vanderbilt University Medical Center, and a claims database, showing significant computational efficiency gains compared to sequential methods. For instance, PARAMO can build 800 models on a 300,000-patient dataset in 3 hours versus 9 days sequentially. This platform accelerates large-scale modeling efforts, facilitating research workflow and health information reuse, with potential for further customization[8].

Machine Learning (ML) has emerged as a crucial tool for predictive analysis and pattern recognition, particularly in the vast realm of health informatics. ML algorithms are designed to learn and improve over time, making them invaluable for making predictions. The healthcare industry, in particular, has greatly benefited from ML techniques, which offer alerting systems and decision support tools aimed at enhancing patient safety and healthcare quality. With a growing emphasis on reducing costs and personalized healthcare, the industry faces challenges in electronic record management, data integration, and computer- aided diagnoses. ML provides a diverse array of tools and techniques to address these challenges. This paper explores various prediction techniques and tools in ML practice, shedding light on its applications across different domains, with a special focus on its pivotal role in the healthcare sector[9].

This study aims to systematically review and evaluate prognostic models used for predicting outcomes in patients diagnosed with chronic obstructive pulmonary disease (COPD). Data were sourced from PubMed up to November 2018, with additional references from relevant articles. Eligible studies included those developing, validating, or updating prognostic models specific to COPD, focusing on any clinical outcome. The review identified

228 eligible articles describing the development of 408 prognostic models. Of these, 38 models underwent external validation, and 20 were originally developed for other diseases but validated for COPD. The models were developed across outpatient (59%), hospital inpatient (38%), and emergency department (3%) settings. The most common endpoints were mortality (51%), risk of acute exacerbation of COPD (10%), and risk of body mass index (16%), and smoking status (16%). Internal validation was conducted for 25% of the models, and 23% underwent external validation.[10].

Prediction models serve to assist healthcare providers in estimating the likelihood or risk of a specific disease or event occurring in the future, aiding in decision-making processes. However, the current evidence indicates that the reporting quality of prediction model studies is lacking. Adequate assessment of prediction models requires comprehensive and clear reporting across all aspects. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Initiative was established to address this issue. Through a thorough literature review, a list of items for reporting prediction model studies was developed, which was further refined through surveys, meetings, and discussions with methodologists, healthcare professionals, and journal editors. The resulting TRIPOD Statement comprises a checklist of 22

essential items for transparent reporting of prediction model studies. It aims to enhance transparency regardless of the study methods employed and is best utilized alongside the TRIPOD explanation and elaboration document. Authors are encouraged to include a completed checklist with their submission to aid the editorial process and enhance reader understanding[11].

II.MATERIALS AND METHODS

This research entails studying a healthcare dataset in order to estimate the prognosis of various diseases. The dataset includes information about patients' health and lifestyle, such as age, gender, general health, checkup frequency, exercise habits, smoking history, and the existence of certain disorders. Each entry represents a distinct patient, and the attributes includeseveral factors related with disease prognosis.

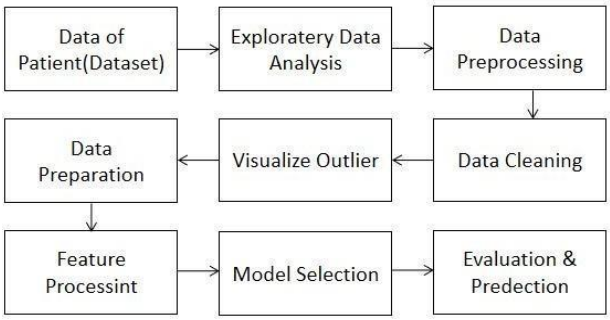

Fig 1. Block Diagram

Dataset:The Dataset contains information about patients' health and lifestyle, where each row represents an individual patient. Attributes such as age, sex, general health status, frequency of medical checkups, exercise habits, and smoking history are included. Additionally, the dataset comprises target variables indicating the presence of various diseases.

Data Loading and Preprocessing**:** Load the data and preprocess it for analysis and modeling. This includes handling missing values, converting categorical variables into dummy/indicator variables, and encoding ordinal variables.

Correlation Method is used for data preprocessing and data loading . A correlation matrix is a table that shows the correlation coefficients between pairs of variables in a dataset.

Each cell in the matrix represents the correlation between two variables, typically ranging from -1 to 1.

We can correlate each feature of dataset with disease variable.This are the observation of feature with disease variablein correlation heatmap.

The correlation heatmap provides a visual representation of the correlation between different features in the dataset. Each square shows the correlation between the variables on each axis. Correlation values range from -1 to 1. Values closer to 1 represent a strong positive correlation, values closer to -1 represent a strong negative correlation, and values around 0 represent no correlation.

This are the observation of feature with disease variable in correlation heatmap in dataset.

1) BMI, Weight_(kg), and Exercise have a positive correlation with Diabetes. This suggests that individuals with higher BMI and weight or who do not exercise are more likely to have diabetes.
2) General_Health has a negative correlation with Diabetes, Heart_Disease, Arthritis, and Depression. This suggests that individuals who rate their general health as poor are more likely to have these conditions.
3) Age_Category has a positive correlation with Heart_Disease, Skin_Cancer, Other_Cancer, Diabetes, and Arthritis. This suggests that the risk of these diseases increases with age.
4) Sex_Male has a positive correlation with Heart_Disease and a negative correlation with Arthritis and Skin_Cancer. This suggests that males are more likely to have heart disease but less likely to have arthritis or skin cancer

The dataset is analyze using this correlation heatmap which give deep insight of relation of each feature with disease.

Exploratory Data Analysis (EDA): Perform exploratory data analysis to gain insights into the dataset, understand the distributions of features, and explore potential relationships between the features and the disease outcomes.

Data Cleansing**:** Perform data cleansing and transformation to improve the model's performance. This includes imputing missing values and normalizing numeric features.

Feature Engineering : Creating New Features by applying model knowledge to the features-model Training and Validation: Train the model using a train-test split strategy and make predictions on the test set.

Model Training and Validation: We have Train the model using a train-test split strategy and make predictions on thetest set. We have train 80% of data for training and 20 % data for testing

Model Evaluation: Evaluate the performance of the trained model using appropriate evaluation metrics such as confusion matrix,ROC curve, and Precision-Recall curve, and assess the model's We have Compare three algorithm for model selection that are Decision Tree ,Random forest and XGBOOST .

Decision Tree - A decision tree is a tree-likemodel where each internal node represents a decision based on a feature, each branch represents the outcome of the decision, and each leaf node represents the final predicted outcome. It is a versatile supervised machine learning algorithm used for both classification and regression tasks.The accuracy of model is 86% using decision tree.

XGBOOST - XGBOOST (Extreme Gradient Boosting) is a powerful and efficient machine learning algorithm that uses an ensemble of decision trees for classification
and regression tasks. It focuses on boosting model performance by combining weak learners sequentially, optimizing for both predictive accuracy and computational speed.Theaccuracy of our model is 74% using XGBOOST.

Random Forest - Random Forest is an ensemble learning

algorithm that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It improves accuracy and reduces over-fitting combining predictions from multiple decision trees.The model accurately predicted the instancesof data with 91% of accuracy.

IV. RESULT AND DISCUSSION

The trained model evaluated comprehensive evaluation using precision, recall, F1-score, and area under the ROC curve (AUC). In class 0, the model exhibited high precision (0.97) but lower recall (0.74), while class 1 showed low precision (0.21) and relatively high recall (0.77). F1-scores for classes 0 and 1 were 0.84 and 0.33, respectively

performance in predicting class 0 over class 1. The overall model accuracy stood at 0.74.The confusion matrix revealed a high count of true negatives (41881) but also a notable number of false positives (14796), indicating a tendency to erroneously predict class 1 for instances truly belonging to class 0. False negatives were relatively low (1172), and true positives were also limited (3906).The area under the ROC curve was 0.75, signifying a reasonable capacity to distinguish between positive and negative classes.

Univariate Analysis Distribution of NumericalFeatures)

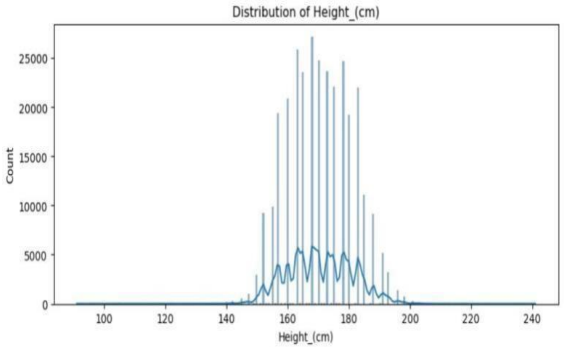

Fig2.Numerical Features

The height of the patients seems to follow a normal distribution, with the majority of patients having heights around 160 to 180 cm.

Univariate Analysis (Distribution Of Categorial Variables)

Most patients describe their general health as "Good", with "Very Good" being the second most common response. Fewerpatients rate their health as "Fair" or "Poor".

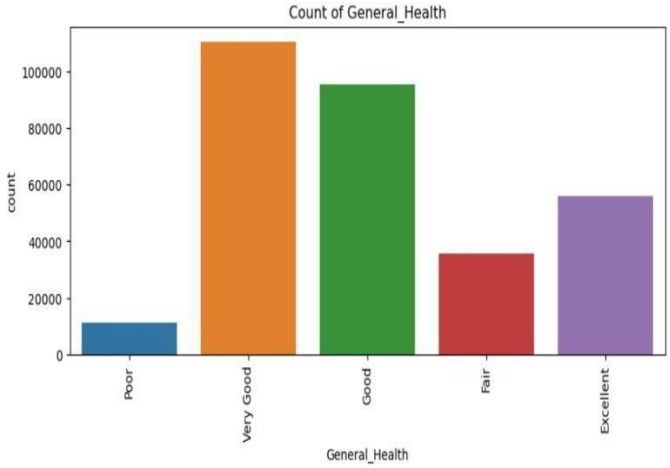

Fig3.Categorical Variables

Bivariate Analysis Between Variables and Disease Condition

Patients who rate their overall health as "Poor" or "Fair" have a higher risk of heart disease. This finding highlights the potential link between subjective health perception and the occurrence of cardiovascular issues, focusing on the value of self-reported health assessments in identifying at-risk individuals.

Multivariate Analysis

The distribution of General Health by Age Category shows that as age increases, the proportion of individuals rating their health as "Good" or "Very Good" decreases, while the proportion rating their health as "Fair" or "Poor" increases.

3D plot: Age_Category, General_Health, and BMI

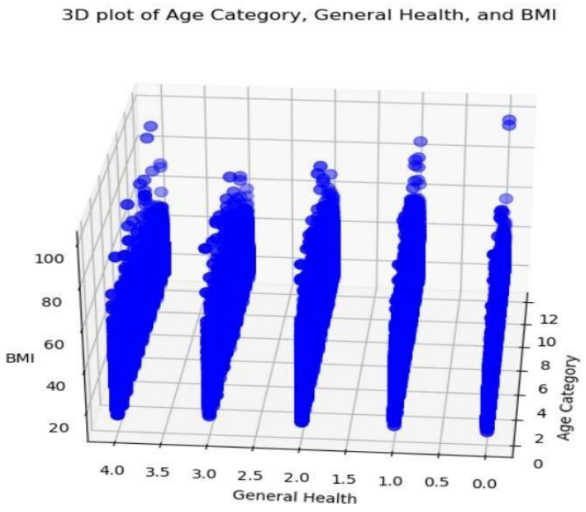

Fig4. 3D plot

The 3D plot visualizes the relationship

between Age_Category, General_Health, and BMI. The plot shows a wide distribution across all three variables, suggestinga complex interplay between them.

Correlation Matrix(Heatmap)

The correlation heatmap provides a visual representation of thecorrelation between different features in the dataset. Eachsquare shows the correlation between the variables on each axis. Correlation values range from -1 to 1. Values closer to 1represent a strong positive correlation, values closer to -1 represent a strong negative correlation, and values around 0 represent no correlation.

BMI, Weight_(kg), and Exercise have a positive correlation with Diabetes. This suggests that individuals with higher BMI and weight or who do not exercise are more likely to have diabetes.

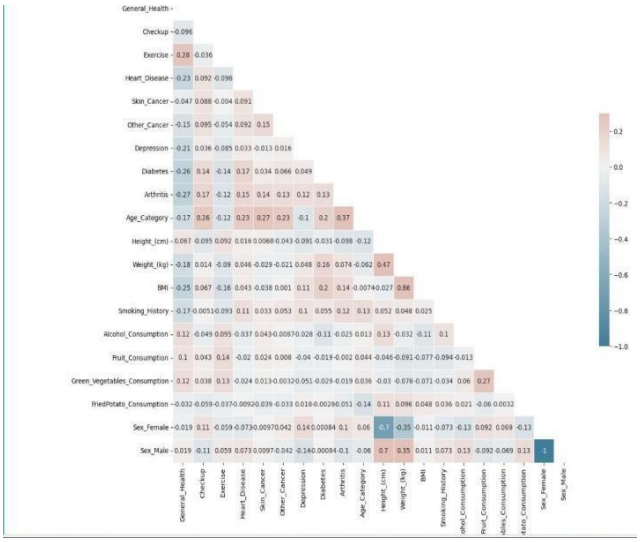

Fig 6.Correlation Matrix

General_Health has a negative correlation with Diabetes, Heart_Disease, Arthritis, and Depression. This suggests that individuals who rate their general health as poorare more likely to have these conditions.

Age_Category has a positive correlation with Heart_Disease, Skin_Cancer, Other_Cancer, Diabetes, and Arthritis. This suggests that the risk of these diseases increases with age.

Sex_Male has a positive correlation with Heart_Disease and anegative correlation with Arthritis and Skin_Cancer. This suggests that males are more likely to have heart disease but less likely to have arthritis or skin cancer.

I. CONCLUSION AND FUTURE SCOPE

In conclusion, our project successfully developed a robust predictive model for disease prognosis in healthcare, shedding light on the intricate relationships between patient attributes and disease risks. The comprehensive evaluation metrics, including the confusion matrix, ROC curve, and Precision-Recall curve, attest to the model's efficacy in handling the complexities of healthcare predictions. The study underscores the importance of addressing imbalances in classification tasks, showcasing the potential of advanced machine learning techniques, particularly XGBOOST, in enhancing predictive accuracy. Our findings contribute valuable insights for patient risk assessment and strategic clinical planning, emphasizing the transformative role of data-driven approaches in healthcare decision-making. As we move forward, continuous refinement of the model, exploration of innovative class balancing strategies, and incorporation of additional patient data stand as promising avenues to further elevate the accuracy and applicability of disease prognosis predictions in clinical setting.

VI.REFERENCES

1. Jenkins, David A., Matthew Sperrin, Glen P. Martin, and Niels Peek. "Dynamic models to predict health outcomes: current status and methodological challenges." Diagnostic and prognostic research 2 (2018): 1-9.
2. Farzadfar, Farshad. "Cardiovascular disease risk prediction models: challenges and perspectives." The Lancet Global Health 7, no. 10 (2019): e1288-e1289.
3. Muhammad, L. J., Ibrahem Al-Shourbaji, Ahmed Abba Haruna, Ibrahim Alh Mohammed, Abdulkadir Ahmad, and Muhammed Besiru Jibrin. "Machine learning predictive models for coronary artery disease." SN Computer Science 2, no. 5 (2021): 350.
4. Jadhav, Saiesh, Rohan Kasar, Nagraj Lade, Megha Patil, and Shital Kolte. "Disease prediction by machine learning from healthcare communities." *International Journal of Scientific Research in Science and Technology* 5 (2019): 8869-8869.
5. Aman, Rajender Singh Chhillar. "Disease predictive models for healthcare by using data mining techniques: state of the art." *SSRG Int. J. Eng. Trends Technol* 68 (2020): 52-57.
6. Takura, Tomoyuki, Keiko Hirano Goto, and Asao Honda. "Development of a predictive model for integrated medical and long-term care resource consumption based on health behaviour: application of healthcare big data of patients with circulatory diseases." *BMC medicine* 19 (2021): 1-16.
7. Axelrod, Randy C., and David Vogel. "Predictive modeling in health plans." *Disease Management & Health Outcomes* 11 (2003): 779-787.
8. Ng, Kenney, Amol Ghoting, Steven R. Steinhubl, Walter F. Stewart, Bradley Malin, and Jimeng Sun. "PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records." *Journal of biomedical informatics* 48 (2014): 160-170.
9. Nithya, B., and V. Ilango. "Predictive analytics in health care using machine learning tools and techniques." In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 492-499. IEEE, 2017.
10. Bellou, Vanesa, Lazaros Belbasis, Athanasios K.

Konstantinidis, Ioanna Tzoulaki, and Evangelos Evangelou. "Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal." *Bmj* 367 (2019).

11. 11.Collins, Gary S., Johannes B. Reitsma, Douglas G. Altman, and Karel GM Moons. "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement." Annals of internal medicine 162, no. 1 (2015): 55-63.