Disease Based Eligibility Assessment for Health Insurance

Dudla Anil kumar(✉), Are Santhosh, and Mule Siva Prasad Reddy

Department of Computer Science and Engineering, Lakireddy Bali Reddy College of Engineering, Mylavaram 521230, India
anil.dudla@gmail.com

Abstract. This paper focuses on revolutionizing health insurance eligibility assessments by deploying advanced machine learning algorithms. Through the analysis of extensive medical records, including historical diagnoses, symptoms, and health indicators, our automated system predicts the likelihood of heart disease, diabetes, liver disease, lung cancer and abnormal RBC counts. Using classification and clustering approaches together improves disease profiling's precision and effectiveness. This data-driven approach aims to streamline the eligibility process, significantly reducing administrative burdens for both applicants and insurance providers. The system provides an objective and personalized evaluation of an individual's health status, contributing to a fair and sustainable health insurance market. This paper model not only aids in predicting specific diseases but also serves as a tool to mitigate adverse selection risks, ensuring a more equitable distribution of insurance coverage. By fostering a comprehensive understanding of an individual's health profile. This paper significantly contributes to improving healthcare accessibility and providing financial security.

Keywords: Machine Learning · Mutual Information · Logistic Regression · KNN Algorithm · XGB Classifier · SVM

1 Introduction

Health insurance plays a significant role in mitigating financial impact of unforeseen medical expenses, serving as a vital tool for individuals to access quality healthcare services without being burdened by exorbitant costs. It stands as a cornerstone in today's intricate healthcare landscape, fostering preventive care, early diagnosis, and effective management of chronic conditions. By promoting a healthier population and alleviating the strain on healthcare systems, health insurance provides a crucial sense of financial security, enabling individuals to seek necessary medical treatments without the looming threat of crippling debt. The determination of eligibility for health insurance involves a comprehensive and multifaceted evaluation process, taking into account an individual's medical history, age, lifestyle choices, and existing health conditions [1]. This meticulous assessment is paramount for maintaining a fair and sustainable health insurance market, preventing adverse selection, and ensuring the equitable allocation

P. Pareek et al. (Eds.): IC4S 2024, LNICST 599, pp. 250–265, 2025.
https://doi.org/10.1007/978-3-031-77081-4_20

of coverage. This paper is dedicated to refining the health insurance eligibility process through the innovative application of advanced machine learning techniques, particularly in predicting specific diseases that significantly influence an individual's insurability and overall health. Focused on key health indicators such as heart disease, diabetes, abnormal red blood cell (RBC) counts, urine analysis, liver disease, and lung disease, our model employs sophisticated algorithms to analyze medical data. By accurately predicting heart disease, contributing to early diabetes detection, predicting liver diseases, and assessing the likelihood of lung diseases, This paper aims to contribute significantly to an individual's eligibility assessment for health insurance coverage. This intricate interplay between health insurance, disease prediction, and eligibility assessment is a testament to our commitment to enhancing the efficiency, objectivity, and fairness of health insurance evaluations, ultimately contributing to improved access to healthcare services and financial protection for individuals. So, our innovative approach combines advanced machine learning methodologies with in-depth medical analysis to provide a holistic assessment of health insurance eligibility, ultimately contributing to a more efficient, fair, and sustainable healthcare insurance landscape.

2 Literature Survey

This model does not have any standard base paper, but the sub parts of our paper has ML prediction techniques, as long as the accuracy is concerned we are using some of the best ML algorithms and we are taking some base papers for reference.

In the realm of cardiovascular health, the imperative to diagnose heart diseases accurately has led researchers to explore intelligent systems fueled by machine learning. This study, authored by Chaimaa Boukhatem, delves into the realm of predictive modeling for heart diseases, leveraging electronic health data [1]. Boukhatem's work not only contributes valuable insights into the predictive modeling landscape for heart diseases but also underscores the significance of machine learning in harnessing critical health factors for accurate diagnosis. This research serves as a beacon in the ongoing quest for effective early detection and management of heart diseases.

Pronab Ghosh's research on efficient cardiovascular disease (CVD) prediction via machine learning stands as a pivotal exploration into a prevalent health concern. The study emphasizes the crucial role of early diagnosis in mitigating CVDs and reducing mortality rates [2]. The proposed model integrates diverse methods, utilizing efficient data collection, pre-processing, and transformation techniques across a combined dataset from various sources. A thorough assessment of the model's results is made possible by creative hybrid classifiers, careful application of ML measures, and advanced feature selection. Ghosh's work provides valuable insights into effective heart disease prediction, contributing significantly to the landscape of predictive modeling for cardiovascular health.

Ishan Sen's research on predictive healthcare makes use of ML algorithms to look into lung ailment diagnosis right away [3]. This study delves into conditions such as Asthma, Allergies, COPD, bronchitis, emphysema, and lung cancer, emphasizing the importance of timely prediction for proactive health management. A variety of machine learning techniques are used in the study, such as Bayesian networks, Random Forest,

Logistic Models, Bagging, and Logistic Regression [4]. The objective is to evaluate their effectiveness in predicting lung diseases. Sen's work contributes valuable insights to the landscape of predictive modeling for lung diseases, illuminating the various uses of ML algorithms in this domain. This research serves as a foundation for healthcare professionals, guiding them in the early detection and proactive management of lung diseases.

Ketan Gupta's exploration into liver disease prediction through Machine Learning is a significant stride in proactive healthcare [5]. This study delves into the extensive datasets from liver patients, emphasizing the application of supervised learning from the UCI Repository. Gupta highlights the richness of information derived from medical examinations of liver patients, envisioning its potential for shaping future improvements in patient care [6]. The study focuses on utilizing historical and classified patient data to predict future outcomes, fostering a foundation for advancements in proactive management of liver conditions. The results underscore Gupta's dedication to improving predictive modeling for liver diseases, offering valuable insights for the ongoing enhancement of healthcare practices.

Abdullah Alqahtani's research addresses the rising global prevalence of kidney stones and its significant impact on individuals' well-being. Kidney stones can lead to complications such as ureter blockage, urinary tract infections, and kidney damage, underscoring the need for early detection. Leveraging Improved Modified Binary Particle Swarm Optimization (Improved MBPSO) for feature selection and Sigmoid functions for precise predictions with binary values, the research proposes an efficient approach. The classification process is refined by Improved Modified XGBoost, updating loss functions for effective learning [7]. Internal comparisons with Decision Tree and Naïve Bayes validate the proposed system's efficacy, highlighting its potential contribution to advancing predictive models in healthcare.

3 Proposed Methodology

The proposed methodology employs a combination of ML algorithms to determine the eligibility of an individual for health insurance. The process involves the prediction of several primary diseases, and based on these predictions along with additional parameters, we derive the final status of a health insurance application. Each disease prediction utilizes a distinct dataset and employs a unique machine learning model.

3.1 Data Gathering

HEART: Utilizing the "heart.csv" dataset, which encompasses 14 attributes, our paper focuses on predicting the presence of heart disease. The crucial "target" field in this dataset indicates the occurrence of heart disease in patients, represented by integer values 0 (no disease) and 1 (disease). The dataset, containing instances of 0 and 1, underwent meticulous data cleaning procedures, including the handling of missing values, correction of inaccuracies, and the transformation of dataset into a workable format. Duplicates and outliers were also addressed. This refined dataset now stands as a robust foundation for training and validation purposes in our endeavor to predict heart disease and assess

eligibility for health insurance. Random forest algorithm is used for this dataset which helps us to get good accuracy and applied data cleaning methods to handle missing values.

KIDNEY STONE DETECTION BASED ON URINE ANALYSIS: The "kidney.csv" dataset, which included 79 urine specimens, was used for estimating the existence of kidney stones by performing urine analysis [Fig. 1]. The analysis aimed to discern potential relationships between certain physical characteristics of urine and the formation of CaC2O4. Preceding model training, the dataset underwent meticulous preprocessing to ensure its suitability for predicting kidney stone presence based on urine analysis [8]. Logistic Regression [9], KNeighborsClassifier [10] and SVC algorithms are used for this dataset to attain 91.66% accuracy (Table 1).

Table 1. kidney stone prediction based on urine analysis dataset.

Attribute	Range
Gravity	1–1.04
Ph	4.76–7.94
Osmo	187–1236
Cond	5.1–38
Urea	10–620
Calc	0.17–14.3
Target	0 or 1

LIVER: We utilized the "liver disease prediction.csv" dataset, comprising 11 attributes, to facilitate liver disease prediction. This dataset encompasses a total of 583 records, with 416 pertaining to patients with liver conditions and 167 to those without. The focal point of our predictions is the "Dataset" variable, denoting the presence of liver disease with values of 0 or 1. This dataset underwent careful preprocessing to ensure its integrity and suitability for training and validation in our liver disease prediction model. The data preprocessing technique label encoding is used to convert categorical attribute into numerical ones and other data cleaning techniques used to deal with outliers and missing values. And SVC algorithm is better compared to other ML algorithms for this dataset.

DIABETES: For diabetes prediction, we employed the "diabetes.csv" dataset which is designed to predict, based on diagnostic measurements, whether a patient has diabetes. Comprising 9 attributes, the focal point of our predictions is the "Outcome" variable, representing the presence of diabetes with values of 0 or 1. The dataset, consisting of carefully curated diagnostic information, underwent thorough preprocessing to ensure its reliability and suitability for training and validation in our diabetes prediction model.

LUNG: We have used dataset named "survey lung cancer.csv" from kaggle to predict lung cancer for a person. This dataset contains 310 records with 16 attributes. The target variable named "lung_cancer" specifies whether person suffers from lung cancer or not which is either yes or no. Duplicates and null records were removed by data preprocessing

techniques. The data preprocessing technique label encoding is used to convert gender attribute into numerical value.

3.2 System Architecture

The architecture of the proposed system is designed to provide a comprehensive view of the methodology. This architecture outlines the intricate details of how the system operates:

1. Disease Prediction Modules: The Disease Prediction Modules in our system are intricately crafted for each primary health concern, utilizing dedicated datasets tailored to the specifics of heart disease, diabetes, kidney stones, lung diseases, and liver diseases [Fig. 2]. Within these modules, advanced machine learning models analyze the respective datasets to predict the likelihood of the associated disease. This modular approach ensures precision and specialization in predicting individual health conditions.
2. Integration Module: Concurrently, the Integration Module acts as a pivotal component where predictions from each disease module are seamlessly amalgamated. In addition to disease-specific parameters, crucial factors such as age, salary, and red blood cell (RBC) count are factored in. This holistic integration forms the cornerstone for the final health insurance application status prediction [Fig. 1]. The modular design allows for independent assessment within each disease module, ensuring focused predictions, while the cohesive integration guarantees a comprehensive prediction for overall health insurance eligibility.

This architecture facilitates a modular and systematic approach, allowing for the independent prediction of each disease while ensuring a cohesive integration for the final eligibility assessment of health insurance [Fig. 1].

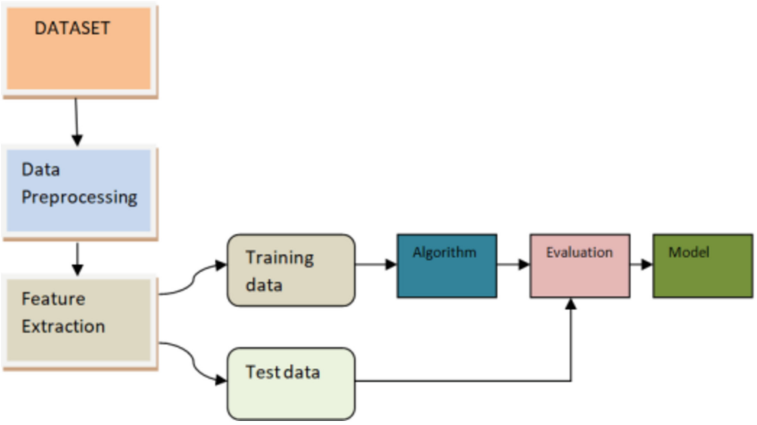

Fig. 1. Model Architecture

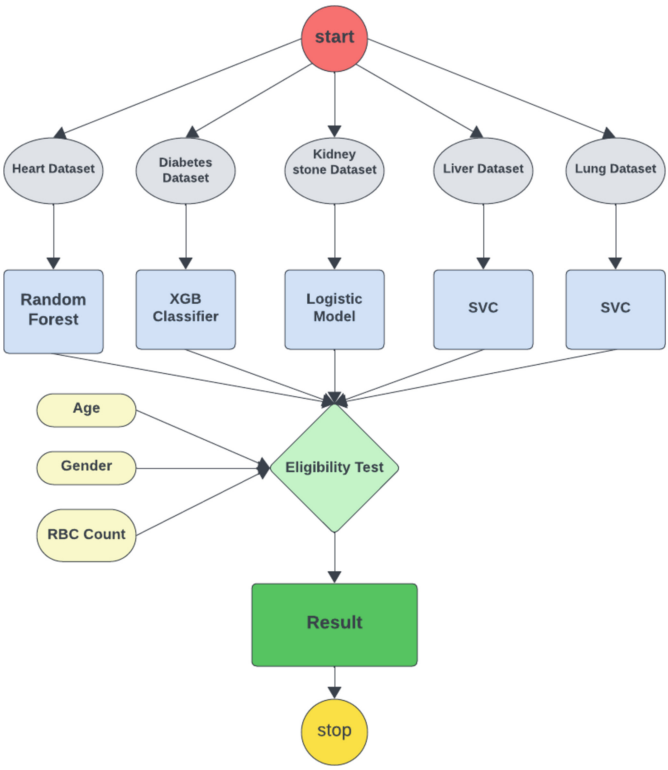

Fig. 2. Workflow of model

3.3 Scalarization in Data Preprocessing

Scalarization is a critical preprocessing step essential for bringing numerical features onto a standardized scale, fostering consistency, and optimizing the performance of machine learning algorithms. This process becomes particularly pivotal when dealing with features exhibiting disparate ranges, magnitudes, or units.

Min-Max Scaling:

One widely adopted scalarization technique is Min-Max Scaling, designed to rescale each feature's values to a predefined range, commonly between 0 and 1. The Min-Max Scaling formula is articulated as follows:

$$Xscaled = \frac{X - Xmin}{(Xmax - Xmin)} \tag{1}$$

- X reflects the feature's initial value.
- X_{min} is the feature's minimal value.
- X_{max} signifies the maximum value of the characteristic.

Standardization (Z-score Normalization): Standardization is another common scalarization method that aims to modify data so that it has a mean(ν) of 0 and a standard

deviation(σ) of 1. The Standardization formula is as follows:

$$Xstandardized = \frac{(X - \nu)}{\sigma} \tag{2}$$

In this context:
- X signifies the original value of the feature.
- ν indicates the feature's average.
- σ denotes the feature's standard deviation (SD).

The significance of scalarization lies in its ability to promote uniformity among feature scales, contributing to improved algorithm convergence, enhanced model generalization, and more effective machine learning outcomes.

3.4 Mutual Information

In our quest to optimize feature selection and enhance model performance, we harnessed the power of Mutual Information. This statistical measure served as a cornerstone in evaluating the relevance and interdependence between individual features and the target variable, providing crucial insights into the predictive capabilities of each attribute.

$$I(A; B) = \sum a \in A \sum b \in Bp(a, b)log(p(a, b)/p(a)p(b)) \tag{3}$$

In this formula, I(A;B) represents the Mutual Information between variables A and B, and p(a, b), p(a), and p(b) are probabilities associated with their respective events. Our methodology involves systematically computing Mutual Information scores for each feature, quantifying the amount of information gained about the target variable. Features with higher Mutual Information scores were deemed more influential in predicting health insurance eligibility and disease presence.

Integration into Feature Selection:

Mutual Information played a pivotal role in our feature selection process. Features exhibiting substantial Mutual Information were prioritised, guiding us in assembling a refined set of attributes that demonstrated heightened relevance to our prediction tasks.By incorporating Mutual Information into our methodology, we not only strengthened the interpretability of our predictive models but also achieved a more streamlined and effective feature set. This technique stands as a testament to our commitment to precision and clarity in optimizing the performance of our algorithms for health insurance eligibility and disease prediction.

3.5 Random Forest Algorithm

An influential ensemble learning technique widely employed for classification and regression tasks is the Random Forest algorithm. It builds upon the principles of decision trees, yet introduces a powerful concept of aggregation. Similar to a Decision Tree, a Random Forest comprises multiple trees, each created through a random subset of the training data and a arbitrary subset of features. These trees operate independently, and their predictions are combined through a process called bagging (Bootstrap Aggregating).The data partitioning occurs recursively, with each tree dividing the dataset into

smaller subsets based on attributes that result in the highest information gain. However, in Random Forest, this division is governed by a random subset of features, injecting diversity into the individual trees [7]. As the Random Forest algorithm descends the tree from internal nodes to terminal nodes, decisions are made based on the test findings. The final prediction is a result of aggregating predictions from all the trees, offering robustness and reducing overfitting. To quantify the impurity and guide the splitting process, Random Forest employs metrics such as Gini Impurity and Entropy. Entropy evaluates the impurity of the dataset, whereas Gini Impurity estimates the probability that a randomly chosen data point would belong to the incorrect class.

$$\text{Gini Impurity Formula} : G = 1 - \sum_{i=1}^{c} pi^2 \tag{4}$$

$$\text{Entropy Formula} : E = -\sum_{i=1}^{c} Pi * log \quad 2(pi) \tag{5}$$

The overarching objective of the Random Forest algorithm is to enhance predictive accuracy and resilience against overfitting by introducing randomness in the construction of individual decision trees. This randomness, coupled with the collective wisdom of diverse trees, contributes to the robustness and effectiveness of the Random Forest model.

3.6 Logistic Regression

Logistic Regression plays a pivotal role in binary classification tasks, utilizing a probabilistic framework to estimate the probability of an instance belonging to a specific class. The model expresses this probability (P(Y = 1)) through the logistic function:

$$P(Y = 1) = \frac{1}{(1 + e) - (\beta 0 + \beta 1 \times X 1)} \tag{6}$$

During the training phase, the algorithm refines parameters via Maximum Likelihood Estimation to maximize the likelihood of observing the provided outcomes. Logistic Regression establishes a decision boundary, typically a hyperplane, effectively delineating instances of distinct classes within the attribute space [Fig. 3] [5].To prevent overfitting, regularization mechanisms like L1 (Lasso) or L2 (Ridge) are often introduced, governed by the regularization parameter. Evaluation indicators including as accuracy, recall, precision and F1-score offer insights into the model's performance on validation or test data. The ROC curve visually illustrates the trade-off between true positive and false positive rates [11]. Logistic Regression proves instrumental across varied domains, including healthcare, finance, and marketing. Its versatility shines in tasks such as disease prediction, credit risk assessment, and customer churn prediction, owing to its interpretability and efficacy. In essence, Logistic Regression establishes a robust foundation for addressing binary classification challenges, merging simplicity with potent predictive capabilities.

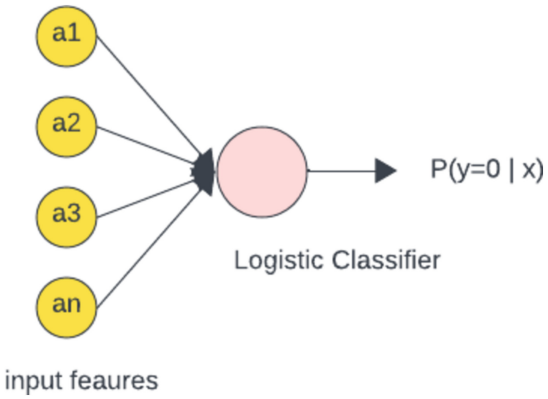

Fig. 3. Logistic Regression

3.7 Support Vector Machine (SVM) Algorithm

In the classification and regression arena, the Support Vector Machine (SVM) algorithm emerges as a robust contender, leveraging geometric principles for nuanced predictions. SVM aims to find the best hyperplane that maximises the margin between instances of various classes in the feature space [11] [Fig. 4]. Crucial to its adaptability is the kernel trick, introducing non-linearity through functions like the polynomial, RBF kernels and linear.

SVM's decision functions for these kernels encapsulate essential implementation formulas:

1. Linear Kernel SVM Decision Function:

$$Formula : f(x) = Bo + B1x1 + B2x2 + \ldots + Bnxn \tag{7}$$

2. Polynomial Kernel SVM Decision Function:

$$Formula : f(x) = \left(\gamma.\left(x, x'\right) + r\right)^d \tag{8}$$

3. Radial Basis Function (RBF) Kernel SVM

$$\text{Decision Function} Formula : f(x) = \exp\left(-\gamma.x - x'||^2\right) \tag{9}$$

The training process involves optimizing the hyperplane to separate classes and maximize the margin for enhanced generalization [11] [Fig. 4]. Notably, SVM supports classification of several classes with strategies such as one-vs-one or one-vs-all. Its versatile applications span image recognition, text classification, and bioinformatics. In summary, Support Vector Machines, with their geometric acumen and kernel-induced adaptability, navigate intricate data landscapes, transcending linear boundaries to discern complex patterns.

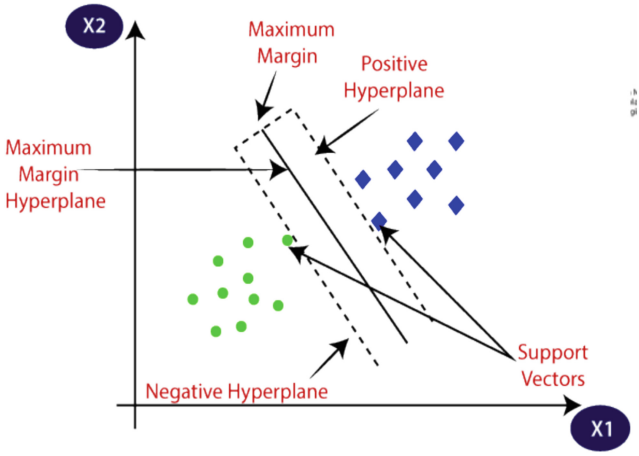

Fig. 4. SVM Analysis

3.8 K-Nearest Neighbors (KNN) Algorithm

In the domain of proximity-based learning, the K-Nearest Neighbors (KNN) algorithm surfaces as an intuitive and adaptable technique for classification and regression tasks. Guided by the fundamental principle of proximity, KNN categorizes instances by considering the majority class of their 'k' nearest neighbors in the attribute space [10]. The selection of 'k' and the distance metric, often using Euclidean distance, profoundly influences the algorithm's behavior. In classification scenarios, the predicted class is determined through majority voting among the nearest neighbors, while regression tasks involve calculating either the average or a weighted average of target values [11]. Specifically, in weighted voting for regression, the implementation involves introducing a weightage based on inverse-distance weighting:

$$\hat{Y} = \frac{\sum_{i=1}^{k} \left(\frac{1}{di}\right) * Yi}{\sum_{i=1}^{k} 1/di} \tag{10}$$

This formula encapsulates the weighted impact of each nearest neighbor, where $\_d_i\_$ denotes the distance to the i^{th} neighbor, and y_i signifies the target value associated with that neighbor. While KNN's simplicity and interpretability render it applicable across diverse domains, such as recommendation systems and anomaly detection, considerations for scalability become imperative with larger datasets due to the necessity of storing the entire dataset. In essence, K-Nearest Neighbors offers a straightforward yet potent strategy for proximity-based learning, finding equilibrium between interpretability and predictive prowess.

3.9 XGBoost Algorithm

In the realm of gradient boosting frameworks, XGBoost (Extreme Gradient Boosting) stands out as a powerful and efficient algorithm known for its speed and performance.

Using a sequential training methodology, XGBoost creates a group of ineffective learners, typically decision trees [12]. Each subsequent tree corrects errors made by its predecessors, with the final prediction being a weighted sum of the predictions from all the trees. XGBoost introduces features that enhance model robustness and control complexity. Regularization terms, including L1 and L2 regularization on leaf weights, prevent overfitting and improve generalization [10]. The algorithm optimizes tree construction by incorporating a "max depth" parameter, limiting the depth of each tree to prevent overly complex structures and contribute to faster training times. At the core of XGBoost lies its objective function, a measure of the model's performance. This objective function incorporates a loss term quantifying the difference between predicted and actual values, as well as a regularization term for each tree.

$$Objective = \sum_{i=1}^{n} Loss(\hat{y}, yi) + \sum_{k=1}^{K} \cap(fk) \tag{11}$$

This formula encapsulates the comprehensive objective function, where Loss($\hat{y}$, yi)) measures the prediction accuracy, and Ω (f_k) represents the regularization term for the kth tree. XGBoost offers insights into feature importance, aiding in the interpretation of each feature's impact on model predictions. Its support for parallel and distributed computing accelerates the training process, making it particularly advantageous for large datasets. Widely applicable across various domains, including finance, healthcare, and technology, XGBoost's versatility, transparent feature importance, and robust performance solidify its standing as a state-of-the-art tool in the machine learning landscape.

3.10 Performance Metrics

i. Accuracy: It fundamental metric in our paper, provides an overarching assessment of our predictive models. It shows the percentage of accurately anticipated cases out of all the occurrences that were assessed. A great degree of accuracy score symbolizes the model's proficiency in making accurate forecasts across all classes, offering a comprehensive understanding of its overall reliability. However, it is essential to interpret accuracy in the context of the specific paper requirements, as it may not be the sole determinant of model effectiveness in scenarios with imbalanced classes.

$$Accuracy = No\ of\ Correct\ Predictions/Total\ number\ of\ Predictions \tag{12}$$

ii. Precision: Precision is a critical metric, particularly in scenarios where false positives can have significant consequences. By calculating the ratio of true positive predictions to the total of true positives and false positives, it evaluates the accuracy of positive forecasts. A high precision score signifies the model's capability to minimize false positive instances, showcasing its accuracy in identifying positive cases.When it comes to applications like fraud detection or medical diagnosis, where reducing false positives is essential, precision is very useful.

$$Precision = TruePositives/(TruePositives + FalsePositives) \tag{13}$$

iii. Recall (Sensitivity): Recall is a statistic that assesses the model's capacity to identify every positive case in the dataset. It is sometimes referred to as sensitivity or true positive rate. The ratio of true positive forecasts to the total of true positives and false negatives is what's being measured. A high recall score means that the majority of positive events are successfully identified by the model, showcasing its sensitivity to the presence of the target condition. In applications like illness diagnosis, where finding every positive instance is crucial, recall is especially crucial.

$$Recall = TruePositives/(TruePositives + FalseNegatives) \tag{14}$$

iv. F1-Score: It is a comprehensive metric that strikes a balance between recall and precision, providing a nuanced evaluation of a model's performance. It considers both false negatives and false positives, offering a more holistic view of the model's effectiveness. The F1-score is particularly valuable when precision and recall need to be balanced, providing a single metric that encapsulates the model's ability to make accurate positive predictions while avoiding false negatives and false positives [13, 14]. A good F1-score signifies a model that better in both recall and precision, achieving a well-rounded performance.

$$F1\_Score = (2xPrecisionxRecall)/(Precision + Recall) \tag{15}$$

v. Specificity: Specificity is a crucial performance metric that focuses on the model's capacity to identify negative instances correctly. The ratio of true negative forecasts to the sum of true negatives and false positives is what being measured. A high specificity score indicates the model's proficiency in avoiding false alarms for negative cases, making it particularly relevant in scenarios where minimizing false positives is imperative [15, 16]. This metric provides a complementary perspective to recall and is valuable in applications where correctly identifying true negative instances is crucial.

$$Specificity = TrueNegatives/(TrueNegatives + FalsePositives) \tag{16}$$

vi. Area Under the ROC Curve (AUC-ROC): The AUC-ROC curve is a pivotal metric in our paper, providing a nuanced assessment of the discriminatory power of our binary classification models [17]. This graphical representation delineates the trade-off between the true positive rate (sensitivity) and the false positive rate across various classification thresholds. The AUC-ROC curve encapsulates the model's capacity to differentiate between negative and positive instances, with a higher area under the curve indicating superior discriminative capabilities. Its comprehensiveness lies in considering the entire spectrum of sensitivity and specificity, offering valuable insights into the model's overall performance beyond traditional metrics. A higher AUC-ROC score signifies enhanced discriminatory ability, making it an indispensable tool for evaluating and comparing different classification models.

vii. Confusion Matrix: The confusion matrix is a pivotal tool that visually represents the performance of a classification model. It offers a thorough analysis of predictions that are false positive, false negative, true positive, and true negative [Fig. 5]. This matrix is instrumental in gaining insights into the model's strengths and weaknesses, enabling a more nuanced evaluation of its performance. By analyzing the confusion

matrix, stakeholders can identify specific areas of improvement and refine the model to enhance its predictive capabilities. The confusion matrix serves as a foundation for deriving various performance metrics including precision, specificity, accuracy, recall, and the F1-score [17].

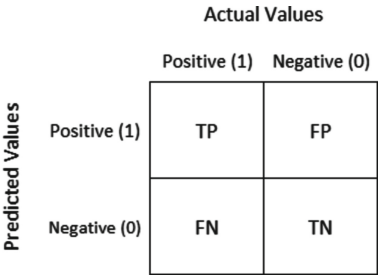

Fig. 5. Confusion Matrix

These performance metrics collectively provide a comprehensive evaluation of the machine learning algorithms implemented in predicting primary diseases for health insurance eligibility. Adjustments and interpretations can be made based on the specific goals and priorities of your paper.

4 Results

In this experimental setup, we adopted a rigorous dataset division strategy, allocating 70% of the data for model training and reserving the remaining 30% for testing. This partitioning allowed us to train our machine learning algorithms on a substantial portion of the dataset, ensuring robust learning while maintaining a stringent evaluation on unseen data to assess predictive performance (Table 2).

To visually interpret the comparative effectiveness of our algorithms, we meticulously crafted five distinct graphs, each dedicated to a specific disease in our study. These graphs seamlessly juxtapose the accuracy and F1-score metrics across various algorithms for each disease. The graphical representations offer a comprehensive view of algorithmic performance, aiding in the identification of superior models for specific health conditions. This approach not only enhances the interpretability of our results but also provides valuable insightful information about the nuances of algorithm performance across diverse diseases in our health prediction framework (Figs. 6, 7, 8, 9 and 10).

Table 2. Comparison of Performance metrics for all diseases

S.NO	DISEASE	ALGORITHM	ACCURACY	PRECISION	RECALL	F1 SCORE
1	Heart	Random Forest Logistic Regression KNeighborsClassifier	98.7 79.5 83.44	97.3 72.5 79.4	99.1 91.8 88.43	98.65 83.44 83.6
2	Diabetes	RandomForestClassifier XGBClassifier SVC	96.54 97.01 96.05	88.26 95.24 92.8	68.7 68.54 58.42	77.31 79.71 71.71
3	Kidney stone	Logistic Regression KNeighborsClassifier SVC	91.66 91.66 91.66	80 99.1 80	99.1 75 99.2	88.88 85.71 88.888
4	Lung	RandomForestClassifier Logistic Regression SVC	89.24 90.32 91.33	90.36 91.4 92.5	97.40 97.40 97.40	90.32 94.33 94.93
5	Liver	Random Forest GaussianNB SVC	65.5 54.4 71.7	72.8 95.2 71.72	82.6 38.4 99.2	77.4 54.79 83.53

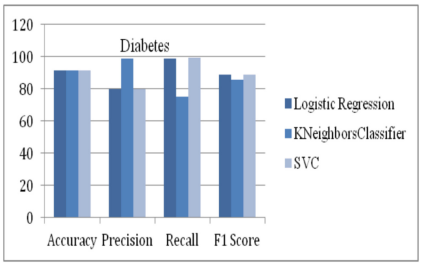

Fig. 6. Comparison graph for Diabetes

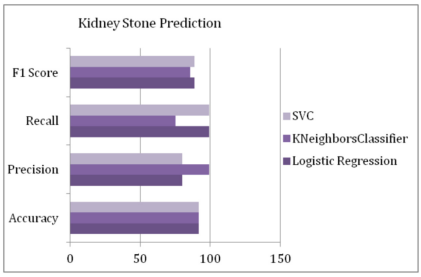

Fig. 7. Comparison graph for Kidney Stone Prediction

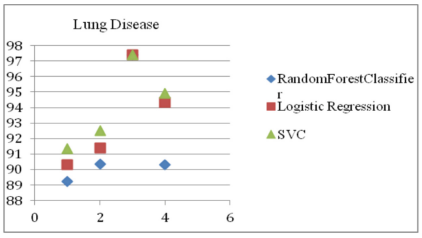

Fig. 8. Comparison graph for Lung Disease

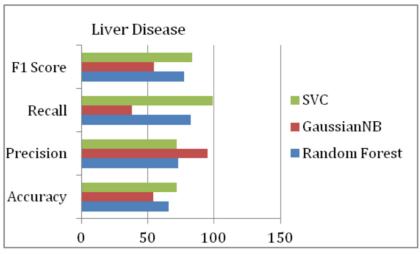

Fig. 9. Comparison graph for Liver Disease

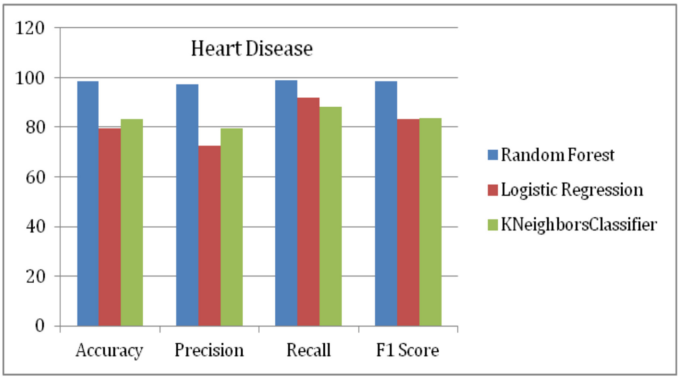

Fig. 10. Comparison graph for Heart Disease

5 Conclusion

In conclusion, This paper "Disease-Based Eligibility Assessment for Health Insurance" marks a significant advancement in the realm of health insurance evaluations. With a core objective of predicting diseases and assessing individual eligibility, our system utilizes state-of-the-art ML algorithms. This intelligent solution, distinguished by its precision, speed, and cost-effectiveness, streamlines the health insurance assessment process. By employing advanced algorithms without compromising accuracy, our paper signifies a transformative approach to healthcare decision-making. It is very helpful for health insurance companies for evaluation of health insurance application.

From many machine learning algorithms, we have selected best algorithm for every disease to attain better accuracy. We have faced many challenges in this paper to get proper information about procedure of insurance in accepting applications and also the factors they will consider to accept health insurance application. Apart of this, health insurance companies may consider few other diseases and criteria in the process of evaluation. Looking ahead, our future endeavors aim to extend this paper to next level and attain better accuracy.

References

1. Boukhatem, C., Youssef, H.Y., Nassif, A.B.: Heart disease prediction using machine learning. In: Computer Science 2022 Advances in Science and Engineering Technology International Conferences (ASET) (2022)
2. Ghosh, P., et al.: Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. IEEE Access **9**, 19304–19326. [9333574] (2021). https://doi.org/10.1109/ACCESS.2021.3053759
3. Sen, I., Hossain, M.I., Shakib, M.F.H., Imran, M.A., Al Faisal, F.: In depth analysis of lung disease prediction using machine learning algorithms. In: Bhattacharjee, A., Borgohain, S., Soni, B., Verma, G., Gao, XZ. (eds.) Machine Learning, Image Processing, Network Security and Data Sciences. MIND 2020. Communications in Computer and Information Science, vol. 1241. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-6318-8_18

4. Zou, X., Hu, Y., Tian, Z., Shen, K.: Logistic regression model optimization and case analysis. In: IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). https://doi.org/10.1109/ICCSNT47585.2019.8962457
5. Gupta, K., Jiwani, N., Afreen, N., Divyarani, D.: Liver disease predict. In: IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT) (2022). https://doi.org/10.1109/CSNT54456.2022.9787574
6. Anistyasari, Y., Hidayati, S.C., Harimurti, R., Ekohariadi: A random forest algorithm for predicting computer programming skill associated with learning styles. In: 6th International Conference on Vocational Education and Electrical Engineering, ICVEE 2023 (2023). https://doi.org/10.1109/ICVEE59738.2023.10348199
7. Alqahtani, A., Alsubai, S., Binbusayyis, A., Sha, M., Gumaei, A., Zhang, Y.-D.: Optimizing kidney stone prediction through urinary analysis with improved binary particle swarm optimization and eXtreme gradient boosting. Journal **11**(7), 1717 (2023). https://doi.org/10.3390/math11071717
8. Varma, K.R., Pulagam, A., Manohari, V.G., Nandini, P., Joan Niveda, J.: Chapter 13-Analyzing and Predicting Diabetes Chronic Disease Using Machine Learning Techniques. Springer (2023)
9. Mittlböck, M.: Explained variation for logistic regression – small sample adjustments, confidence intervals and predictive precision. Biometric. J. 04/2002 (2002)
10. Kumar, D.A., Jyothi, U., Mohan, K.J., Mounica, V., Nageswari, A.: Kernel-based SVM classifiers for multi-disease forecasting: a meta-analysis. In: 4th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, pp. 841–847 (2023). https://doi.org/10.1109/ICOSEC58147.2023.10275963
11. Mavroforakis, M.E., Theodoridis, S.: Support Vector Machine (SVM) classification through geometry. In: 13th European Signal Processing Conference (2023)
12. Obiora, C.N., Ali, A., Hasan, A.N.: Implementing extreme gradient boosting (XGBoost) algorithm in predicting solar irradiance. In: IEEE PES/IAS PowerAfrica (2021). https://doi.org/10.1109/PowerAfrica52236.2021.9543159
13. Taunk, K., De, S., Verma, S.: Aleena Swetapadma.: a brief review of nearest neighbor algorithm for learning and classification. In: International Conference on Intelligent Computing and Control Systems (ICCS) (2019). https://doi.org/10.1109/ICCS45141.2019.9065747
14. Singh, H., Bhatta, N.P., Tawsik Jawad K.M., Singh, H., Amsaad, F., Hopkinson, K.: ML-assisted security for the detection of DDoS attacks in connected IIoT environment: implementation and comparative analysis. In: NAECON 2023 - IEEE National Aerospace and Electronics Conference (2023)
15. Dudla, A.K., Rao Atthuluri, M., Shaik, P.S., Yalamanchili, V.S.B. An efficient approach for analyzing reviews using an ensemble technique. In: 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, pp. 1576–1581 (2023). https://doi.org/10.1109/ICACCS57279.2023.10112729(2023)
16. Chopra, S., Kalra, N., Rani, R.: Identification of cardiovascular disease using machine learning and ensemble learning. In: International Conference on Innovative Data Communication Technologies and Application (ICIDCA) (2023)
17. Kumar, D.A., Ezhilarasan, M.: Shufflenetv2: an effective technique for recommendation system in e-learning by user preferences. In: Morusupalli, R., Dandibhotla, T.S., Atluri, V.V., Windridge, D., Lingras, P., Komati, V.R. (eds.) Multi-disciplinary Trends in Artificial Intelligence. MIWAI 2023. Lecture Notes in Computer Science(), vol. 14078. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-36402-0_16