Predicting Insurance Premium for Life Insurance Applicants using Machine Learning

Subhash Kumar Wary[1], Akher Uddin Ahmed[1], Birhang Borgoyary[1], Prasanta Baruah[1*], Pankaj Pratap Singh[1[0000-0003-4079-4485]]

[1] Department of Computer Science and Engineering, Central Institute of Technology Kokrajhar (Deemed to be University, under MoE, Govt. of India) Kokrajhar, Assam, India -783370
*Email: p.baruah@cit.ac.in

Abstract

This study explores an application of machine learning (ML) to predict the insurance premiums for life insurance applicants. Traditional methods often face challenges in accurately assessing risk due to the complexity and diversity of applicant data. By leveraging ML algorithms, we analyze a wide range of applicant attributes, including demographic details, medical history, lifestyle factors, and financial indicators. In this research, we employ models such as linear regression (LR), decision trees (DT), random forests (RF), and gradient boosting (GB), with feature engineering techniques used to enhance the accuracy of predictions. Our findings indicate that ML models significantly outperform traditional methods in terms of risk assessment and pricing strategies for insurers. Specifically, the LR model achieved an accuracy of 74%, demonstrating moderate performance. The RF regressor, with an accuracy of 83%, exhibited better predictive capabilities, particularly in handling complex datasets and capturing nonlinear relationships. GB, however, outperformed all other models with an accuracy of 86%, showcasing its strong predictive power. The ability of gradient boosting to iteratively improve the performance of weaker models contributed to its superior results. This research advances underwriting practices by demonstrating the potential of ML to revolutionize risk assessment and pricing strategies in the insurance industry, to provide precise and data-driven decision-making tools for insurers.

Keywords: Gradient Boosting Regressor, Linear Regression, Machine Learning, Predictive Modelling, Random Forest Regressor

1 Introduction

The insurance industry plays a crucial role in mitigating financial risks and providing economic security to both individuals and businesses. Among the various insurance products, life insurance stands out as a vital safeguard against unforeseen events, ensuring the financial well-being of beneficiaries. However, determining insurance premiums, the cost of coverage, is a multifaceted process influenced by numerous factors. Life insurance relied on models of statistics using data which are old. Some of the traditional approaches set the foundation of all the insurance pricing and this

struggles to capture the risk profiles. The data which are growing in voluminous amount introduces new challenges for insurers. Advancements in machine learning (ML) techniques contributed in the improvement of the accuracy and efficiency from the voluminous data. It is capable in analyzing huge amounts of dataset which then uncovers different patterns. In this study, we developed predictive models by utilizing diverse datasets with diverse applicant information [1]. The old ways of determining life insurance premiums sometimes inadequately assess the underwriting decisions. The diverse data sources provide improved risk assessment. With ML techniques, we tried to explore by developing the many predictive models accurately enhancing with fairness. We intend to improve the underwriting efficiency with different ML algorithms. The models performance was rigorously evaluated using different metrics like the Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2).

2 Literature Survey

Risk assessment is an important way in the life insurance to deal with the challenges that lies ahead. Earlier, the process has relied on actuarial methods and now since ML techniques develops more robust, efficient and precise results it is now easy to navigate through huge sets of data. The ML approaches helps the insurers categorize applicants effectively which then improves the premiums with the individual risk levels. Many different ML algorithms were evaluated like Random Trees (RF), Linear Regression (LR) and Gradient Boosting (GB) [2][5][6][7]. The ML algorithms provide several advantages in life insurance and depicting the primary improvements in the accuracy of risk prediction. The ML models have the capability of handling large datasets and finding the complexity of the relationships between applicant characteristics which are difficult to understand with older methods [12].

ML algorithms manages to streamline the underwriting process by automation of the many tasks which reduces the time and cost in finding out the premium calculation and also increases the fairness and transparency in underwriting which reduces the human decision making still with these benefits there exist several challenges. All the ML models are extensively dependent on the quality and availability of data. The data which are incomplete will definitely undermine the effectiveness of even the most advanced algorithms. The concerns relating to the ethics regarding the fairness and also the bias in the ML models are also severe since there is a risk that the models may unintentionally discriminate against certain demographic groups [3]. All these studies collectively underscores the importance of the ML algorithms in the advancements of the underwriting practices like the enhancement of risk assessment (see table 1).

Table 1: Related Work

Authors & Year	Objectives	Methodology	Results
Junedi and Mauritsius (2020) [11]	to predict the risk level of life insurance applicants using ML.	Three ML classifiers as SVM, Naive Baye's & RF are used.	Precision using RF, SVM using a linear kernel & NB are 0.85, 0.72 and 0.49 respectively.

Chang et al. (2022) [18]	to develop an automated diabetes diagnosis system using ML models.	It discusses SMOTE and ADASYN for dealing with imbalanced datasets & used DT, RF, NB, KNN and LR.	5 types of classifiers (DT, RF, NB, KNN and LR) for predicting diabetes and RF achieved highest accuracy.
Kaushik et al. (2022) [4]	to predict health insurance premiums by using the concept of AI and ML in healthcare.	The training and evaluation of an ANN and LR based regression model is done to predict health premiums.	ANN based regression model for predicting health insurance premium with accuracy of 92.72%.
Singh et al. (2017) [13]	To classify and predict the different level of diabetes in the patients	Data mining approach	Comparison with association rule mining based approach
Baruah et al. (2023) [14]	to improve risk assessment in life insurance industries of the applicants using predictive analytics	Geographical Information Systems (GIS) and ML approaches	ML approaches are applied to predict the applicants risks on both the datasets.
Baruah and Singh [15]	Review on role of Risk prediction for customers in insurance companies.	Risk classification is done based on their risk levels by grouping of customers with ML	underwriters role is to calculate the premium based on the calculated risk of customers

3 Methodology

Our work employs ML techniques to predict life insurance premiums, utilizing a structured and systematic approach to data collection, preprocessing, and model development. Data for this study is sourced from publicly available online databases, ensuring a diverse and comprehensive dataset. The research adopts a positivist paradigm, which emphasizes empirical analysis and hypothesis testing, aligning with the objective of predicting premiums using ML algorithms. This approach enables a detailed exploration of patterns and relationships within the dataset, facilitating improved decision-making in insurance underwriting.

3.1 System Architecture

The data is collected and the steps of preprocessing is done capturing the key applicant attributes such as age, gender, BMI, children, smoker, region, and charges. In the preprocessing stage which includes tasks such as cleaning, feature engineering, and standardization to ensure compatibility with the ML models (see fig. 1). After the data preparation we deployed the three ML algorithms to develop predictive models like

the LR, RF Regressor, and GB Regressor. Each of the algorithms was trained and evaluated on the dataset with the GB Regressor emerging as the most accurate achieving an accuracy of 86%. The hyperparameter tuning and cross-validation techniques are applied to optimize the model's performance and ensure generalizability. The final model is integrated into an interactive platform also allowing insurance professionals to input applicant data and receive real time predictions.

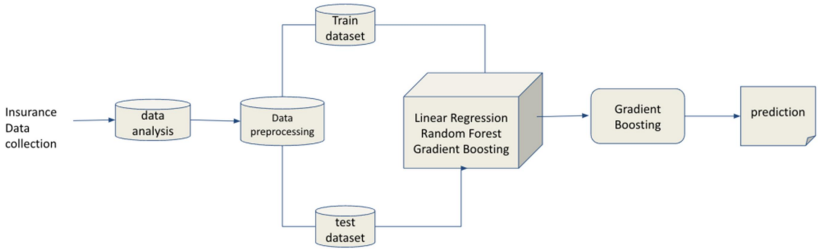

Fig. 1: System Architecture

3.2 Data Collection and Preparation

The dataset [16] which is used in this study consist of applicant data including all the features such as age, sex, BMI, children, smoker, region, and insurance charges (see table 2). Some key attributes behavior are shown in the figs. 2-5. The data is comprised of a mix of nominal continuous and discrete variables which necessitates preprocessing to ensure compatibility with ML models. The process such as data cleaning is done to find where missing values are addressed using appropriate imputation techniques and outliers are identified and managed.

Table 2: Insurance Dataset

Sl. No.	Age	Sex	BMI	Children	Smoker	Region	Charges
1	19	Female	27.9	0	Yes	Southwest	16884.92
2	18	Male	33.77	1	No	Southwest	1725.552
3	28	Male	33	3	No	Southwest	4449.462
4	33	Male	22.705	0	No	Northwest	21984.47
5	32	male	28.88	0	No	Northwest	3866.855
6	31	Female	25.74	0	No	Southeast	3756.622
7	46	Female	33.44	1	No	Southeast	8240.59
8	37	Female	27.74	3	No	Northwest	7281.506
9	37	Male	29.83	2	No	Northeast	6406.411
10	60	Female	25.84	0	No	Northwest	28923.14
11	25	Male	26.22	0	No	Northeast	2721.321
12	62	Female	26.29	0	Yes	Southeast	27808.73
13	23	Male	34.4	0	No	Southwest	1826.843
14	56	Female	39.82	0	No	Southeast	11090.72

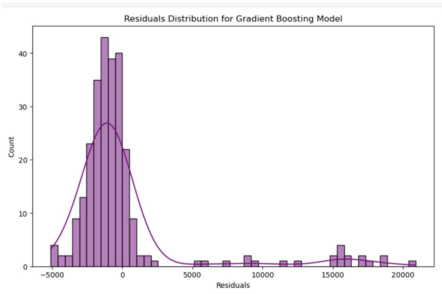

Fig. 2: Region

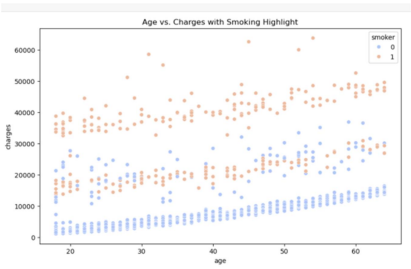

Fig. 3: Smoker Highlight

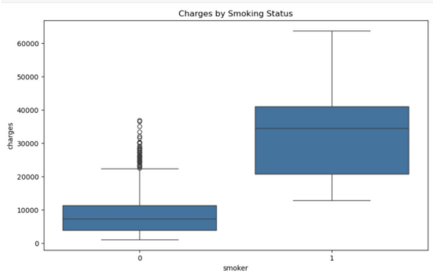

Fig. 4: Charges by Smoking Status

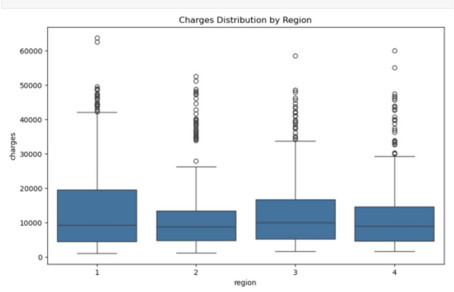

Fig. 5: Region

3.3 Model Development

Three ML algorithms LR, RF Regressor, and GB Regressor are implemented. Each model is evaluated based on its performance in predicting premiums extensively and results are shown accordingly up to the mark with the training and testing datasets. The LR model achieved an accuracy of 74%, highlighting its moderate performance. It serves as a useful baseline model and its limitations such as sensitivity to outliers. The RF Regressor is another ML model which proved to be more effective than LR after achieving an accuracy of 83%. The GB Regressor model that we implemented emerged as the best algorithm after achieving an accuracy of 86%. This model operates by sequentially building a series of weak learners and each one correcting the errors of the previous model allowing it to capture complex patterns in the data. The ensemble nature of GB enables it to handle noise and outliers effectively which makes the most suitable choice for our study. With hyperparameter tuning and cross-validation the performance of the model is further optimized which demonstrated its ability to provide accurate and reliable predictions for life insurance premiums.

4 Implementation and Results

4.1 Implementation Details

After implementing the various ML models for predicting insurance premiums involved several key steps and also containing applicant attributes such as age, gender, BMI, smoking status, and region. The dataset underwent preprocessing, including handling missing values, detecting and addressing outliers, and ensuring compatibility with the ML algorithms. Three algorithms are selected for training and they are LR, RF Regressor, and GB Regressor. All these algorithms are trained on a portion of the dataset and evaluated using performance metrics such as accuracy, R-squared, and mean absolute error. The hyperparameter tuning and cross-validation techniques were applied to optimize the model's performance to ensure robustness against overfitting and improving its generalizability to unseen data. We then deployed the models in a user-friendly application interface and where insurance professionals can input applicant details and receive real time predictions for insurance premiums.

4.2 Results

The ML models are evaluated based on their accuracy. The quantitative values of the models are shown in the table 3.

Table 3. Accuracy Score of Different ML Algorithms

Approaches	Accuracy Score
Linear Regression (LR)	74%
Random Forest (RF) Regressor	83%
Gradient Boosting (GB) Regressor	86%

4.3 Analysis

The deployment of the ML model is achieved using Flask a lightweight Python-based web framework [8] enabling seamless integration with the trained ML model. When the data is processed it is implemented into the model which produced premium predictions and sent back to the frontend for user display (see fig. 6). The frontend of the application is developed [9]. The prediction process was streamlined to ensure real time feedback. The backend powered by the trained ML model returned the predicted premium in real time interface. The integration allowed for an efficient and cohesive user experience where inputs and outputs flowed smoothly between the frontend and backend [10]. The GB Regressor marks superiority in the performance with an accuracy of 86% outperforming both the LR and RF Regressor models. The model’s sequential boosting mechanism allowed it to correct residual errors from previous iterations which enhanced its ability to model complex patterns and relationships from the data. The ensemble nature made the GB regressor resilient to noise and outliers which further improves its predictive accuracy.

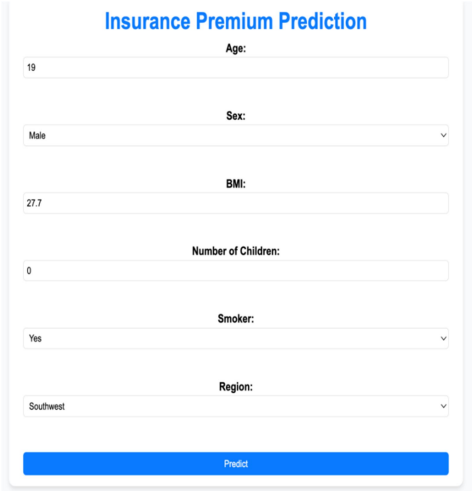

Fig. 6. Interface created using Flask and Python

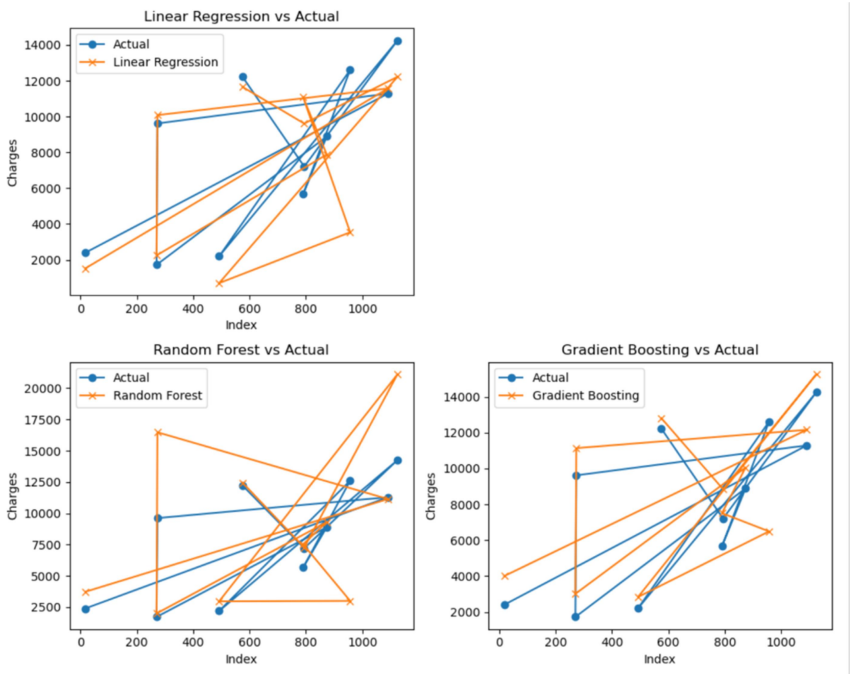

Fig. 7: Comparison of three ML Models Performance

4.4 Model Performance and Evaluation

All the three ML models were evaluated using standard performance metrics such as the Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. The LR model achieved an accuracy of 74% indicating moderate performance in premium prediction (see figures 7-9). The RF Regressor achieved accuracy up to 83% making it as an effective model. However, the GB Regressor demonstrated superior performance and achieved an accuracy of 86%. Its iterative process of improving weak learners allowed it to capture intricate patterns in the data, making it the most accurate model for premium prediction in our study. The GB Regressor's performance achieved 86% accuracy and is pivotal role in the prediction of insurance premiums.. This attributed to its sequential boosting approach which allowed the model to iteratively correct errors leading to a more precise prediction.

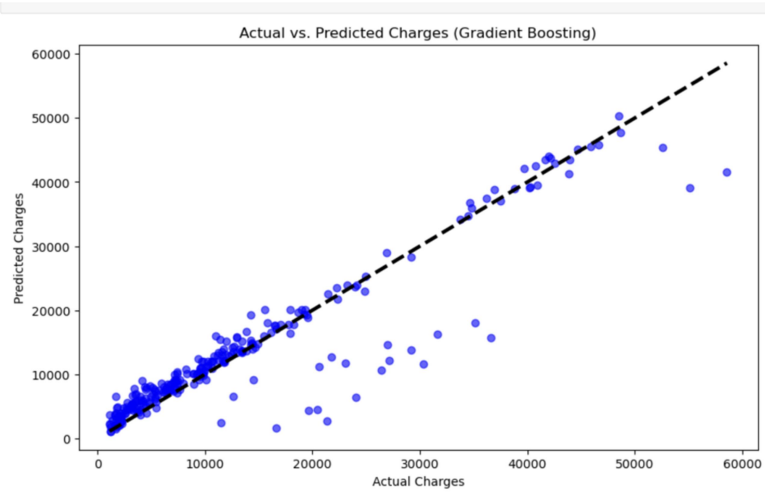

Fig. 8. Actual *vs.* Predicted Charges

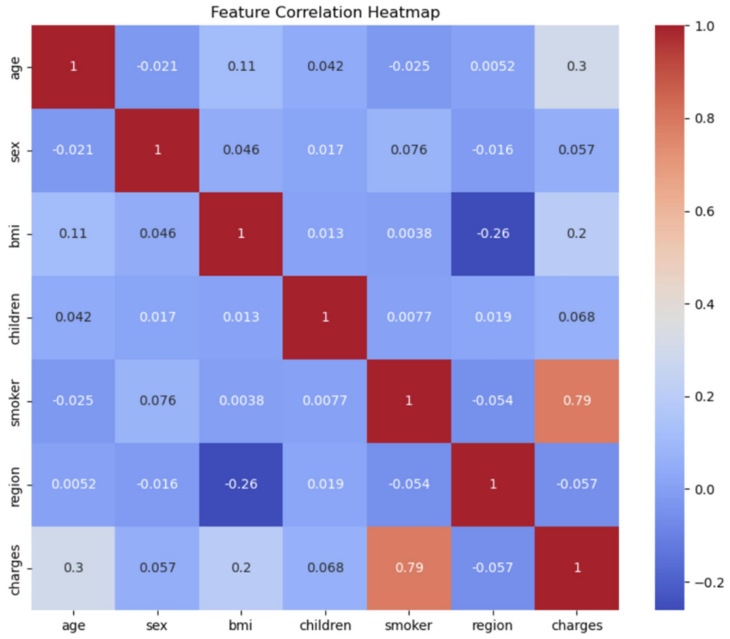

Fig. 9. Confusion Matrix (Heat map)

Fig. 10. R2 Scores for Different Regression Models

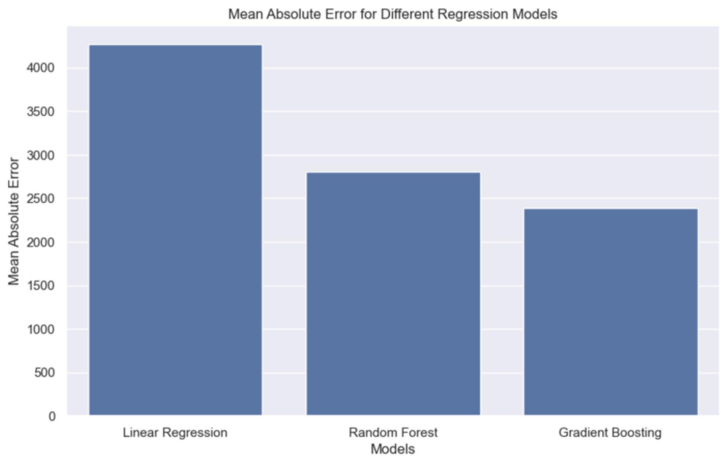

Fig. 11. Mean Absolute Error for Different Regression Models

R-squared scores and Mean Absolute Errors for each regression model are shown in the figures 10 and 11 respectively and predicting the accuracy high of the GB Regressor.

5 Conclusion and Future Scope

Our study examined thoroughly the application of ML algorithms for predicting life insurance premiums based on applicant attributes. All the three models such as LR, RF Regressor, and GB Regressor were successfully developed predictive models capable of estimating premiums with a high degree of accuracy. The GB Regressor demonstrated the strongest predictive performance after achieving an accuracy of 86%. We will explore the potential of advanced feature engineering techniques where new insights will be derived from the already available data which could improve the models accuracy. We will address the issues with that data quality like the missing values, the outliers and imbalanced datasets for improving further. We will work on to reduce bias in the model predictions to maintain equity in the insurance practices. Finally, we will develop frameworks for real time implementation of ML models.

References

1. Shakespeare W.: An Introduction to Financial Intermediaries and Risk. An Introduction to Financial Markets and Institutions. 2010:67.
2. Chang, V., Ganatra, M.A., Hall, K., Golightly, L., Xu, Q.A.: An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. Healthcare Analytics 2, 1-14 (2022).
3. Rawat S, Rawat A, Kumar D, Sabitha AS.: Application of machine learning and data visualization techniques for decision support in the insurance sector. International Journal of Information Management Data Insights. 2021 Nov 1;1(2):100012.

4. Kaushik, K., Bhardwaj, A., Dwivedi, A.D., Singh, R.: Machine Learning- Based Regression Framework to Predict Health Insurance Premiums. International Journal of Environmental Research and Public Health 19(13), 1-15 (2022).
5. Jolly K.: Machine learning with scikit-learn quick start guide: classification, regression, and clustering techniques in Python. Packt Publishing Ltd; 2018 Oct 30.
6. Louppe G.: Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502. 2014 Jul 28.
7. Zemel R, Pitassi T.: A gradient-based boosting algorithm for regression problems. Advances in neural information processing systems. 2000;13.
8. Singh P.: Deploy machine learning models to production. Cham, Switzerland: Springer. 2021.
9. Dinh D, Wang Z.: Modern front-end web development: how libraries and frameworks transform everything.
10. Chan J, Chung R, Huang J.: Python API Development Fundamentals: Develop a full-stack web application with Python and Flask. Packt Publishing Ltd; 2019 Nov 22.
11. Junedi, H.B. and Mauritsius, T.: Risk Level Prediction of Life Insurance Applicant using Machine Learning. International Journal of Advanced Trends in Computer Science and Engineering 9(2), 2213-2220 (2020).
12. Maier M, Carlotto H, Saperstein S, Sanchez F, Balogun S, Merritt S.: Improving the accuracy and transparency of underwriting with AI to transform the life insurance industry. AI Magazine. 2020 Sep 14;41(3):78-93.
13. Singh, P.P., Das, B., Poddar, U., Choudhury, D.R., Prasad, S.: Classification of diabetic's patient data using machine learning techniques. In: Perez G., Tiwari S., Trivedi M., Mishra K. (eds) Ambient Communications and Computer Systems, AISC, vol 696, pp. 427-436, Springer Singapore (2017).
14. Baruah, P., Singh, P.P., Ojah, S.K.: A Novel Framework for Risk Prediction in the Health Insurance Sector using GIS and Machine Learning. International Journal of Advanced Computer Science and Applications 14(12), 469-476 (2023).
15. Baruah, P., Singh, P.P.: Risk Prediction in Life Insurance Industry Using Machine Learning Techniques-A Review. In: Mishra, A., Gupta, D., Chetty, G. (eds) Advances in IoT and Security with Computational Intelligence (ICAISA). Lecture Notes in Networks and Systems, vol 755, pp. 323-332 (2023).
16. [online] Dataset: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
17. Hutagaol BJ, Mauritsius T. Risk level prediction of life insurance applicant using machine learning. International Journal of Advanced Trends in Computer Science and Engineering. 2020 Mar;9(2).
18. Chang V, Kandadai K, Xu QA, Guan S. Development of a Diabetes Diagnosis System Using Machine Learning Algorithms. International Journal of Distributed Systems and Technologies (IJDST). 2022 Jan 1;13(1):1-22.