Medical insurance cost prediction using Machine Learning

Diaa Salama AbdElminaam[1], Mohamed Khaled[2],
Karim Ahmed[3], Rana Mohamed Abulkassem[4], Sara Yasser[5], Youssef Mohamed[6]
Faculty of Computer Science
Misr International University, Cairo, Egypt
diaa.salama[1], mohamed.kmohamed[2],
karim2110342[3], rana2105525[4], sara2100416[5], youssef2109450[6]{@miuegypt.edu.eg}

***Abstract*—The accuracy of prediction of medical insurance costs has become very important in healthcare industry. In this paper , we provide a machine learning-based approach to develop a predictive model for medical insurance. Six machine learning models were explored and compared, such as Naïve Bayes, Decision Tree, Random Forest, Support Vector Machines, Logistic Regression, and k-Nearest Neighbors. By investigating two datasets about patients' medical conditions, ages, and gender, these models were trained and tested to predict the future medical expenses. The performance of the models is evaluated through many experiments, and the results show their effectiveness in accurately estimating medical insurance costs. Through extensive experiments on all models and comparing their accuracy, it was found that the Random Forest model emerged as the best model in accuracy for estimating the cost of the medical insurance in both datasets. This research shows the potential of machine learning to enhance the accuracy and efficiency of medical insurance prediction. In conclusion, the accuracy and reliability of machine learning-based models in predicting medical insurance costs remain an active area of research. While the current models offer promising results, there are several potential limitations that need to be addressed in future studies.**

Keywords: Medical insurance price prediction; Machine Learning; Regression; Random forest; Decision tree; Linear Regression; K-Nearest Neighbor;Support Vector Machines; SVM ;KNN ;Naïve Bayes.

I. INTRODUCTION

In recent years, healthcare insurance cost prediction has become as a crucial challenge for both healthcare providers and policy makers.The inability to accurately estimate medical insurance costs leads to significant financial problems for both healthcare providers and insurance companies. This is primarily due to the complexity of healthcare costs, which are influenced by various factors such as demographic characteristics.

By incorporating advanced machine learning algorithms, we can improve the accuracy and efficiency of cost prediction for medical insurance.However, there are several challenges that need to be addressed to make these models effective in predicting medical insurance costs. One major thing is getting a large and diverse enough dataset that accurately represents the medical insurance cost-prediction problem.

To achieve the most accurate model for predicting medical insurance costs, we collected two datasets with their information to train the models on them and to compare each model to find the one with the highest accuracy in predicting the medicalinsurance cost. After collecting the datasets, we preprocessed the data by cleaning and organising it to ensure its quality. Additionally, we split the datasets into training and evaluating sets to evaluate the performance of each model effectively.

After training the models on the collected datasets, we evaluated their performance using various metrics such as mean squared error, R-squared values, mean absolute error, and root mean square error.This helped us to find the most accurate model in predicting the medical insurance cost and the results showed that the Random Forest model highlights the most accurate predictions for medical insurance costs for both datasets, as it has the highest R-squared value and the lowest mean absolute error, root mean square error, and mean squared error values.

Fig. 1. Medical insurance cost prediction using Machine Learning

The contributions made to this topic are:

- The Medical insurance cost prediction with machine learning.

- The testing of 6 machine learning algorithms.
- The use of 2 datasets, for a total of 3,760 entries with different sets of features spanning between 6 &10 features

The remaining sections in this paper are ordered as the following; related work is discussed in the second section. Moreover, The third section indicates the proposed methodology of the paper, it consists of dataset description and used algorithms. The results of the proposed algorithms can be found in the fourth section and their analysis. The conclusion is located in the fifth section.Finally an acknowledgement towards all the supporting figures of this research is present in the sixth section.

II. Related Work

The field of medical insurance prediction has been extensively studied, with numerous researchers investigating various aspects and achieving significant findings. In our research, we have analyzed several papers that have contributed to our understanding of medical insurance prediction. These papers will be properly referenced in the references section.

In this paper [1], the authors discussed and compared machine learning models for The american predicition of medical insurance cost highlighting on the importance of machine learning in healthcare and insurance cost prediction. The study evaluates algorithm performance in terms of accuracy and computational time, focusing on the need to consider both factors when assessing algorithms. The research confirms the value of machine learning in healthcare and its potential to improve patient experience and help in cost prediction for insurance providers.

In this research [2], authors use many machine learning approaches for the prediction on medical insurance costs, aiming to explain the key determinant factors using SHAP and ICE plots. They compare the approaches' effectiveness and find that XGBoost showed the best model for the prediction of the medical insurance cost than the others. The study shows new ideas and has ramifications for the insurance industry and actuarial modeling in healthcare.

In this study [3], machine learning models, especially linear regression, decision tree regression, and gradient boosting regression, were used to predict healthcare insurance costs. The used a medical insurance dataset to train and test thier models. The models were trained using the training data and evaluated based on testing data. Gradient boosting regression was found to provide the highest accuracy, with an R-squared value of 86.86. This research highlights the potential benefits of using machine learning in the insurance industry, such as improving accuracy, reducing human effort, and enhancing profitability. However, restrictions in predicting medical insurance costs are also admitted, along with the need for further improvement in this area.

This paper [4] addresses the important task of predicting health insurance costs in India, where high-quality healthcare costs are rising. Using machine learning regression models, including polynomial regression, the study obtains an 80.97 percent accuracy and an RMSE of 5100.53. The results focuses the possibility of predictive modeling for accurate cost estimation in real-world insurance field.

This study [5] concentrates on the financial load of medical insurance premiums and aims to promote prediction using various machine learning algorithms. By exploring people's medical history the study confirms the importance of employing machine learning algorithms in the healthcare industry. Using a dataset, the methodology covers preprocessing, splitting into training and testing sets, and training four models—SVM, random forest, linear, and ridge regression. Results show that all models predict premiums reasonably well, with SVM and random forest regression outstanding linear and ridge regression. The results offer valuable insights for insurance field.

The study [6] introduces a machine learning-based regression scope to predict health insurance premiums, utilizing a dataset with 1300 entries and seven columns. The model, implemented in Python, achieved a high overall accuracy of 92.72 percent. The primary objective was to evaluate model accuracy and imagine the relationships between charges and many factors like smoking, region, children, BMI, sex, and age.

In this study [7] the author states the challenge of predicting medical insurance costs, specifically for people with rare diseases, aiming to reduce treatment cost. Using machine learning and deep learning techniques, the study introduces a new methodology for valuing insurance costs. The purpose covers predicting insurance costs based on many factors, using nine regression models such as XGBoost and Gradient Boosting. The dataset is split into training and test data, and the study deduces that XGBoost and Gradient Boosting Regression are the most accurate models, achieving over 86 precent accuracy.

This research [8] the author use machine learning regression approaches to predict health insurance cost, aiming to provide accurate cost predictions to ease the financial stress of medical health. By comparing many regression models, the study underscores the effectiveness of Polynomial Regression in predicting insurance cost. The methodology involves using a dataset from Kaggle, setting the data, and changing categorical values to numerical values using label encoding. The study concludes by pointing to the significance of choosing suitable approaches based on dataset features.

The purpose of this document [9] is to investigate the application of machine learning in predicting medical insurance costs. This research uses a dataset from Kaggle consist of many factors related to health insurance cost, including age, BMI, number of children, gender, smoking status, and region. The paper discusses the importance of using developed statistical methods, machine learning algorithms, and deep neural networks to accurately predict health insurance costs. Additionally, it addresses the challenges faced by the insurance field, such as missing data and bias in simple replacement methods. In summary, this document aims to participate to outstanding research efforts in the field of insurance claim estimation and demonstrate the potential of machine learning to significantly reduce the people efforts involved in policy making.

This study [10] predicts health insurance prices for people using three regression algorithms and compares their effectiveness. The purpose is to help people make informed decisions about their health insurance needs. The Gradient Boosting Regression model was found to be the most accurate, with age and smoking status being important factors. This research provides a valuable scope for people to make informed decisions related to their health insurance needs.

The paper [11] presents a study on medical insurance cost prediction using machine learning algorithms. It emphasizes the importance of predicting medical prices and the potential benefits for patients and policymakers.It uses the Random Forest Regression algorithm to predict medical prices using the Medicare payment dataset.The study compares the performance of machine learning algorithms such as Regression Tree, Random Forest Regression, Linear Regression, and Gradient Boosted Decision Trees.in conclusion it was concluded that the Gradient boosted regression trees model hit the best model in predicting the medical insurance price.

The paper [12] presents the study that aims to estimate healthcare insurance costs using computational intelligence, big data tools, and healthcare charges datasets.The research use Spark, a big data tool, in conjunction with ML algorithms (linear regression,polynomial regression) to estimate healthcare insurance prices, highlighting the effectiveness of the gradient-boosted tree regression model.The results of the study indicate that the polynomial regression model outperforms the linear regression model in accurately estimating healthcare insurance prices.

The paper [13] presents the significance of the field of population cost prediction through public healthcare datasets, it focuses on the enhancement of the population health management and accountability. The authors divide the data into three parts (training,testing and validation) and use claims and surveys to evaluate the performance of the algorithms used. The dataset is analyzed using Regression Tree, M5 Model Tree and random Forest algorithms and root mean squared error, mean absolute error, and prediction error quantiles are used in measuring their the performance. The results collected from these models that the M5 Model Tree provided the most accurate predictions.

The article [14] discusses the usage of the machine learning algorithms in medical insurance price predictions,highlighting the downsides of the traditional methods. Two datasets are used in this study containing discrete and continuous variables. The authors used a hybrid machine learning approach which is a combination of Network Structure Learning and Hybrid Regression. The Hybrid Regression is a combination of 3 algorithms, which are Long Short-Term Memory,Random Forest and Support Vector Regression. The results of the study shows that the hybrid models have better performance than that of the single models. In addition to that the authors plan to improve the performance of the hybrid models.

The article [15] explains the influence of the distinct Regression models on the health insurance price prediction. Its purpose is to establish a new strategy for using the machine learning in medical insurance cost prediction. A dataset of seven attributes is used and divided into training and testing parts. The authors use nine regression models and compare their accuracy by computing the Mean Absolute Error, Root Mean Squared Error, R-squared value, and Mean Squared Error. The outcome of the study reveals that XGBoost Regression, Gradient Boosting Regression, and Random Forest Regression provided the most accurate cost predictions.

The paper [16] highlights the study of hove machine learning plays a crucial role in the improvement of the medical insurance industry. The paper aims to make fast insurance charges prediction and to improve the accuracy of estimating the price of insurance by developing a system named ML Health Insurance Prediction System. The authors use various regression models,which are Multiple Linear Regression, Ridge Regression, Simple Linear Regression, Lasso Regression, and Polynomial Regression. The dataset is split into testing and training parts and evaluated using Root Mean Squared Error, R-squared value, and accuracy metrics. The outcome of the study reveals that the Polynomial Regression model achieved the highest accuracy. The authors suggest to apply more optimized techniques and feature selection methods to improve the accuracy of the predictive models.

III. Proposed Methodology

Before training the model, many models were used and each model was studied widely to understand its performance on the datasets. The diagram below shows the different steps that the datasets went through to obtain the results.

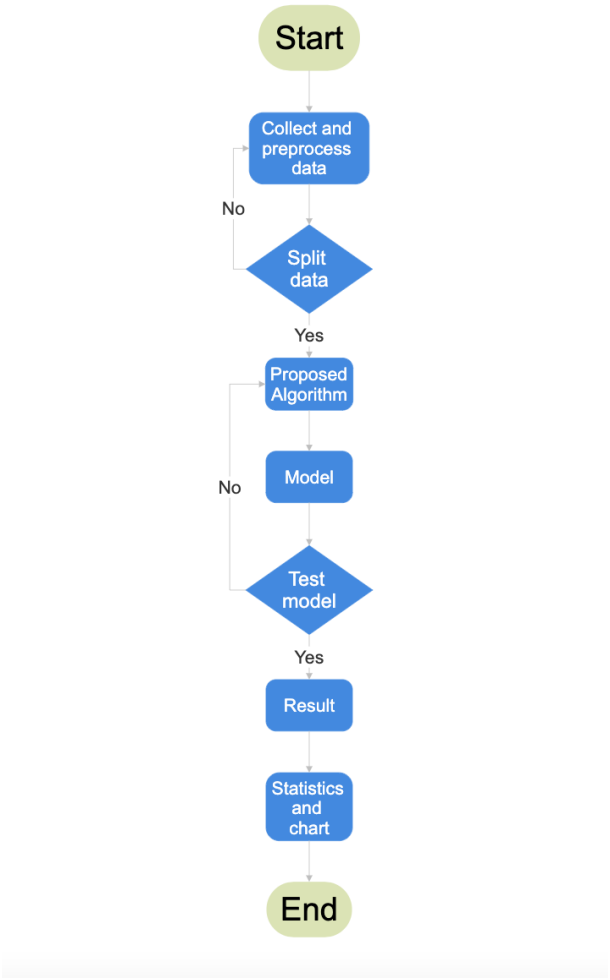

Fig. 2. Medical insurance cost prediction process

A. Datasets Descriptions

The first dataset consists of 6 features, and it has 2,773 records. The dataset is splitted into two sections: 70% for training, and 30% for evaluating A detailed description of the features can be found below.

Charges is the target representing the cost of the medical insurance of each record. Age is the age of the patients that can effect on the medical insurance cost as the old people pay more than the children because typically they need more medical care. Sex is the gender of the patient that has an effect on the medical insurance cost as women policyholders are generally seen to be more expensive for the insurance providers as compared to men. BMI is an abbreviation for Body mass index which is computed by multiplying a person's weight in kg by the square of their height.Children represents the number of children . Smoker represents whether the patient is a smoker or not. Region represents the place where the patient live, and finally the target which is Charges (the Medical insurance cost).

TABLE I
FEATURES OF DATASET 1

Feature	Type	Values
Age	Numerical	From 18 to 64
Sex	Classification	Female or Male
BMI	Numerical	From 15.96 to 53.13
Children	Numerical	From 0 to 5
Smoker	Classification	Yes or No
Region	Classification	northwest or northeast or southwest or southeast

The second dataset consists of 10 features, and it has 986 records.The dataset splitted into two partitions: 80% for training, and 20% for evaluating. A detailed description of the features can be found below.

PremiumPrice is the target, representing the expected price for each patient based on their features. Age represents the age of each patient as it affects the expected medical insurance price. Diabetes represents whether the patient suffers from diabetes or not. BloodPressureProblems represent whether the patient has blood pressure problems or not. AnyTransplants represents whether the patient has had a transplant before or not. AnyChronicDiseases represents whether a person suffers from any chronic diseases or not. Height represents each person's height in centimeters. Weight represents each person's weight in kilograms. KnownAllergies represents whether the patient knows what type of allergy he or she has or not. HistoryOfCancerInFamily represents whether the patient has any family member who suffers from cancer or not. NumberOfMajorSurgeries represents the number of major surgeries performed by each patient.

TABLE II
FEATURES OF DATASET 2

Feature	Type	Values
Age	Numerical	From 18 to 66
Diabetes	Classification	Yes or No
BloodPressureProblems	Classification	Yes or No
AnyTransplants	Classification	Yes or No
AnyChronicDiseases	Classification	Yes or No
Height	Numerical	From 145 to 188
Weight	Numerical	From 51 to 132
KnownAllergies	Classification	Yes or No
HistoryOfCancerInFamily	Classification	Yes or No
NumberOfMajorSurgeries	Numerical	From 0 to 3

B. Used Algorithms

The mentioned datasets were passed into 6 different Machine Learning algorithms which were Support vector machines (SVM), Naïve Bayes, Decision tree, K-Nearest neighbors (KNN) , Logistic regression, Random forest For each of the algorithms there were results and statistics generated, these statistics were: Mean square, Root mean square, R squared, and Mean absolute error. Afterwards, the results were charted and compared. The results, charts, and the discussion of the results can be found later in the paper.

1) **Support Vector Machines (SVM):**

SVM, or Support Vector Machine, is a supervised machine learning algorithm used for regression and classification tasks. SVMs works by identifying the ideal hyperplane that splits all data points of one class from those of the other classes with the maximum margin. By converting he data into a higher-dimensional space where it becomes linearly separable, procedures as kernel functions allows SVMs to handle non-linearly separable data. The decision function involves finding the closest point to the hyperplane.

$$f(x) = W^*X + b \tag{1}$$

Definition: W represents the weight vector, X represents the input features, b represents the bias term, and f(x) represents the predicted output.

2) **Naïve Bayes:**

Naive Bayes is a probabilistic machine learning algorithm used for classification and regression problems. It works on the principle of Bayes' theorem, which calculates the conditional probability of an event occurring given the probability of another event that has already occurred. In the context of Naive Bayes, the 'event' we are trying to predict is the class of an input, given the input's features.

$$P(C|X) = (P(X|C) * P(C))/P(X) \tag{2}$$

Definition: P(C|X) represents the probability of class C given the input features X, P(X|C) represents the probability of observing features X given class C, P(C) represents the prior probability of class C, and P(X) acts as the evidence.

3) **Logistic Regression:**

Logistic regression is a statistical method used for examining a dataset in where a result is determined by identified by one or more independent variables,or features. It belongs to the class of machine learning algorithms known as logistic regression models.The output is mapped the range [0,1] by applying the logistic function to the linear combination of input features.

$$P(y = 1|X) = 1/(1 + e^{-}(W * X + b)) \tag{3}$$

Definition: P(y=1|X) represents the probability of class 1 given the input features X, W represents the weight vector, X represents the input features, b represents the bias term, and e is the base of the natural logarithm.

4) **K-Nearest Neighbors (KNN):**

KNN (K-Nearest Neighbors) is a non-parametric supervised machine learning algorithm used for classification and regression. The underlying concept of KNN is to predict the class or value of a new data point based on the class or value of its K-nearest neighbors in the training data. KNN works by computing the distance between the new data point and all the points in the training set. The distance can be calculated using various distance metrics, such as Euclidean distance, Manhattan distance, or Hamming distance. The K-nearest neighbors are then determined based on the smallest distances. Finally, the predicted class or value is calculated based on the majority class or average value of the K-nearest neighbors.

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i \tag{4}$$

$\hat{y}$: represents predicted value for the target variable.
k : represents the number of nearest neighbors.
y_i : represents the actual target values of the i-th nearest neighbors to the new data point.

$\sum_{i=1}^{k} y_i$: Add up the target values of the k nearest neighbors.

$\frac{1}{k}$: Take the average by dividing the sum by k.

The "Euclidean equation" is used in that model to calculate the distance between consecutive points:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{5}$$

5) **Random Forest:**

Random Forest is a machine learning algorithm that can be used for both classification and regression tasks.

It revolves around the concept of constructing multiple decision trees at training time and then combining them together to make predictions. This approach offers several advantages over single decision trees, such as improved robustness, better performance, and reduced risk of over fitting.

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \tag{6}$$

Where:
- $\hat{y}$ represents the predicted output (target variable).
- N is the total number of decision trees in the Random Forest.
- $f_i(x)$ is the prediction from the i-th decision tree.

6) **Decision Tree:**

Decision tree is a supervised learning algorithm used for both classification and regression tasks. It operates by repeatedly dividing the input features into subsets based on their values. This process continues until the decision tree reaches a stopping condition, Each internal node in the decision tree represents a test on a feature. If the feature value is less than or equal to a threshold, the test passes, and the sample is assigned to the left child node. Otherwise, the test fails, and the sample is assigned to the right child node.at each leaf node, the decision tree algorithm determines the majority class.This majority class or average target value becomes the prediction for the samples in the corresponding leaf node.

Definition: Suppose S is a set of instances, A is an attribute, S_v is the subset of S with $A = v$, and $\text{Values}(A)$ is the set of all possible values of A, then

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} | \frac{|S|}{S|} \cdot \text{Entropy}\,(S_v) \tag{7}$$

C. *Performance Metrics*

Mean Squared Error (MSE) is average of squared differences between predicted and actual values, measuring prediction precision. Root Mean Squared Error (RMSE) is square root of MSE, providing a explainable scale for prediction errors. R-squared is proportion of variance in the dependent variable explained by the model, assessing fit. Mean Absolute Error (MAE) is average absolute differences between predicted and actual values, gauging prediction accuracy.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8}$$

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{10}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{11}$$

IV. Results and Analysis

The results collected from the six models, Support vector machines (SVM) , Naïve Bayes, Logistic Regression, Random Forest, k-Nearest Neighbor (KNN) , Decision Tree are shown below.

The following results were conducted from the first dataset.

TABLE III
STATISTICS OF ALGORITHMS WITH 70/30 DATA SPLIT

Model	MSE	RMSE	R-squared	MAE
Logistic Regression	2.99227e+08	17298.2	-0.912377	11066.1
Random Forest	1.11543e+07	3339.81	0.928712	1558.37
Decision Tree	2.54704e+07	5046.82	0.837217	2976.92
K-NN	8.17281e+07	9040.36	0.477671	5102.32
SVM	1.74587e+08	13213.1	-0.115796	6644.38
Naïve Bayes	3.69391e+07	6077.75	0.76392	2039.7

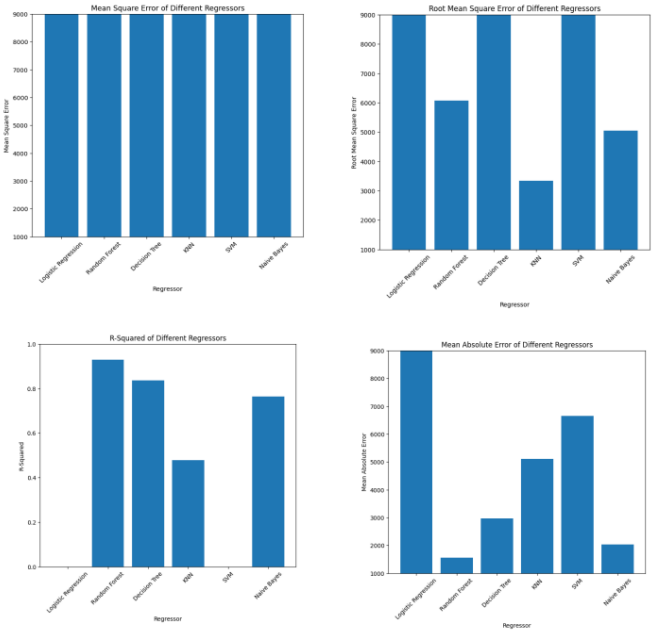

Fig. 3. First dataset performance chart with data split

Random Forest is the dominant algorithm as it has the lowest mean absolute error percentage of 19.0069. Naïve Bayes and SVM algorithms have slightly close mean absolute percentage error of 23.939 and 28.6665 respectively.Both Decision Tree and K-NN algorithms have a higher mean percentage error than that of the Naïve Bayes and SVM, where the Decision Tree has 42.1369 mean absolute percentage error

of and the K-NN has 62.3224 mean absolute percentage error. Logistic Regression is considered the worst algorithm as it has the highest mean absolute percentage error of 120.597.

The following results are from the second dataset.

TABLE IV
STATISTICS OF ALGORITHMS WITH 80/20 DATA SPLIT

Model	SVM	Naïve Bayes	Logistic Regression	Random Forest	KNN	Decision Tree
Accuracy	0.666667	0.146465	0.590909	0.914141	0.580808	0.878788
Mean Square Error	1.772727e+07	4.037374e+07	1.156061e+07	6.313131e+06	2.727273e+07	1.154545e+07
Root Mean Square Error	4210.376792	6354.033158	3400.089125	2512.594538	5222.329679	3397.860289
R-squared	0.584284	0.053211	0.728897	0.851953	0.360438	0.729252
Mean Absolute Error	1989.898990	4646.464646	1479.797980	636.363636	2626.262626	939.393939

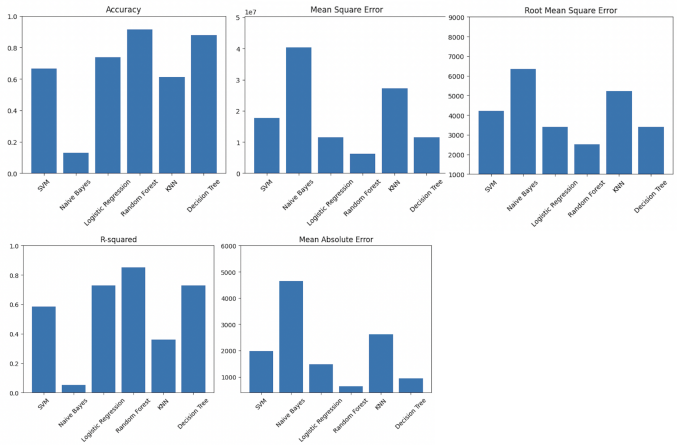

Fig. 4. Plot for Second dataset results

TABLE V
CROSS-VALIDATION SCORES FOR EACH MODEL USING K-FOLD

	$-RMSE$	$-MSE$	$R2$	$-MAE$
SVM	−4421.389	−19548681.541	0.498754	−1970.588
Naïve Bayes	−8241.474	−67921906.693	−0.741577	−6102.434
Logistic Regression	−3773.640	−14240365.111	0.634864	−1548.681
Random Forest	−3030.273	−9437119.675	0.767099	−743.407
KNN	−4973.458	−24735294.117	0.365765	−2449.290
Decision Tree	−4234.984	−16384381.338	0.5157097	−1249.492

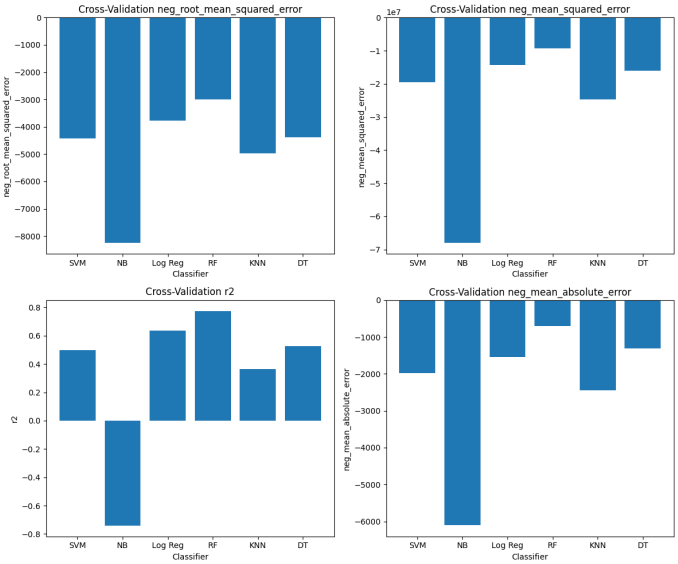

Fig. 5. Plot for cross validation results

After looking at the results of the models on the second dataset, it's clear that the Random Forest model constantly excellence others in accurately predicting medical insurance prices. In contrast, the Naïve Bayes model showed less efficiency. Therefore, considering the results from the second dataset, the Random Forest model appears to be a more credible choice for accurate predictions in the context of medical insurance pricing.

V. CONCLUSION

In conclusion, The prediction of the medical insurance cost can be a main reason for helping many families avoid suffering, by using machine learning it proved to be a great approach to predicting its cost. After conducting a thorough evaluation, it was determined that the Random Forest model demonstrated the highest accuracy among all six models in the 2 datasets we provided . This finding indicates that the Random Forest model is the most suitable choice for predicting medical insurance costs. Moreover The utilization of machine learning models can provide valuable insights and improve decision-making in various fields, including healthcare. The Random Forest model, with its high accuracy, emerges as the most suitable choice for predicting medical insurance costs.

VI. ACKNOWLEDGMENT

As Misr International University students, we would like to show our gratitude the staff of the faculty computer science for giving us the chance to learn at this high-quality educational institution.We would like to give a special thanks to Prof. Ayman Nabil, dean faculty of Computer Science, Dr. Diaa Salama and Eng. Mohamed Khaled for their wonderful efforts this semester and their guidance in helping us achieve our goals.

References

[1] C. M. Barbosa, D. J. P. Alencar, F. d. S. S. Borges, F. P. Cavalcante, F. D. F. Brilhante, J. G. Portela, L. F. L. d. Costa, L. M. C. Fernandes, L. A. Fontenele, M. A. N. Diniz *et al.*, "Machine learning in healthcare management for medical insurance cost prediction," *OPEN SCIENCE RESEARCH II*, vol. 2, no. 1, pp. 1323–1334, 2022.

[2] U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," *Machine Learning with Applications*, p. 100516, 2023.

[3] S. Sushmita, R. Sengupta, and S. Bandyopadhyay, "Predicting individual healthcare costs: A comprehensive review," *Computers and Informatics*, vol. 3, no. 1, p. 1250124, 2023.

[4] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved k-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503–1514, 2012.

[5] M. A. H. Salman, M. A. Mosleh, and S. A. S. Sajid Ullah, "A computational intelligence approach for predicting medical insurance cost," *Mathematical Problems in Engineering*, vol. 2021, no. 2021, p. 1162553, Dec 2021.

[6] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine learning-based regression framework to predict health insurance premiums," *International Journal of Environmental Research and Public Health*, vol. 19, no. 13, p. 7898, 2022.

[7] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting motor insurance claims using telematics data—xgboost versus logistic regression," *Risks*, vol. 7, no. 2, p. 70, jun 2019. [Online]. Available: https://www.mdpi.com/2219-7947/7/2/70

[8] Y. A. Christobel and S. Subramanian, "An empirical study of machine learning regression models to predict health insurance cost," *Webology*, vol. 19, no. 2, 2022.

[9] M. Thorat, "International journal of science, engineering and technology, an open access journal," *International Journal of Science, Engineering and Technology*, vol. 11, no. 3, p. 4, 2023.

[10] N. Bhardwaj and R. Anand, "Health insurance amount prediction," *Int. J. Eng. Res*, vol. 9, pp. 1008–1011, 2020.

[11] N. Jayasekara, S. Karunanayake, S. Tilakaratne, U. Jayawardana, S. Ranasinghe, and D. Senarath, "Implementing multilevel linear regression models in apache spark: A step-by-step guide," *Computers and Informatics*, vol. 3, no. 1, p. 1250124, 2023.

[12] J. Cubanski and J. Neuman, "The facts on medicare spending and financing," *The Henry J. Kaiser Family Foundation*, 2017.

[13] S. Sushmita, S. Newman, J. Marquardt, P. Ram, V. Prasad, M. De Cock, and A. Teredesai, "Population cost prediction on public healthcare datasets," May 18–20 2015.

[14] S. Kaushik, A. Choudhury, N. Dasgupta, S. Natarajan, L. Pickett, and V. Dutt, "Ensemble of multi-headed machine learning architectures for time-series forecasting of healthcare expenditures," vol. 214, pp. 199–216, 2020.

[15] T. Dietterich, "Machine learning models for predicting health insurance costs: A study on medical cost personal datasets," *arXiv: Machine Learning eXchange*, vol. abs/2104.08716, pp. 1–34, 2021.

[16] V. Roth, "The generalised lasso," *IEEE Transactions on Neural Networks*, vol. 15, pp. 16–28, 2004.