International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Machine Learning Approaches to Estimating Health Insurance Expenses

Dr. Sheelesh Kumar Sharma , Md Naved Khan , Mohd Zaid Saifi , Mohd Sanif Khan

ABSTRACT

This study compares the performance of three machine learning models, XGBoost, Artificial Neural Networks (ANN), and Decision Trees, for a specific task.Provide some quantitative results or comparisons to support the claim that ANN performs the best among the three models. We also present a detailed analysis of the models, their strengths and weaknesses, and the factors that contribute to their performance. Our study contributes to the growing literature on machine learning and highlights the importance of selecting the appropriate model for a given task.

Based on the comparison of XGBoost, ANN, and Decision Tree models, it was found that the ANN model has out performance in the prediction task. The ANN model demonstrated a higher accuracy score and lower root mean squared error (MSE) compared to the other models. This indicates that the ANN model is more capable of accurately predicting the target variable compared to the other models. In addition to the comparison of the models, the study also explored the importance of feature selection and hyperparameter tuning in improving the performance of the models. The results showed that selecting relevant features and optimizing hyperparameters can significantly enhance the performance of the models.

Overall, the study highlights the potential of using ANN models in predictive tasks and emphasizes the importance of careful feature selection and hyperparameter tuning to achieve optimal performance

Keywords: machine learning models, XGBoost, Artificial Neural Networks, Decision Trees, performance, quantitative results, accuracy score, root mean squared error, feature selection, hyperparameter tuning, predictive tasks, optimal performance.

INTRODUCTION

Health insurance is one thing that every individual must have. It provides individuals with access to medical care when they require it, without worrying about the high costs for healthcare. In addition to providing financial coverage for medical expenses, health insurance also offers financial stability to individuals, as it protects individuals from unexpected medical charge that can be financially challenging. Health insurance plans also promote preventive care, which helps individuals stay healthy and detect health problems early. This is because health insurance often covers preventive care services such as vaccinations, health screenings, and check-ups. Preventive care is an important part of maintaining good health, as it can help individuals avoid developing serious health problems that require expensive medical treatment.Knowing that you have coverage can reduce the stress and anxiety associated with worrying about medical bills and unexpected health issues. This can be particularly important for individuals who have chronic health conditions or who have a family history of medical problems. Finally, in many countries, having health insurance is a legal requirement. Failure to have health insurance can result in significant financial penalties or even legal consequences. Overall, health insurance is an important component of healthcare systems, providing individuals with access to medical care and financial security

The use of artificial intelligence (AI) in health insurance can have many advantage,that including:

1. Improved Risk Assessment: AI algorithms can analyze vast amounts of data and identify patterns that human underwriters may miss, enabling insurance companies to more accurately assess the risk of insuring an individual. This can lead to more accurate pricing of policies and better decision-making when it comes to accepting or rejecting applications for insurance.
2. Fraud Detection: AI can help detect fraudulent claims by analyzing data from multiple sources and identifying anomalies and patterns that suggest fraudulent activity. This can help insurance companies save money and prevent false claims.
3. Personalized Recommendations: AI can analyze a person's medical history, lifestyle, and other factors to make personalized recommendations for coverage and wellness programs.
4. Improved Customer Service: AI can be used to automate routine tasks such as claims processing, allowing insurance companies to provide faster and more efficient service to their customers.
5. Improved Health Outcomes: By analyzing data on individual health histories, AI can help insurance companies identify individuals who are at risk of developing certain health conditions and provide them with proactive care and support to prevent these conditions from developing or becoming more severe.

Health insurance is influenced by a different parameter. Age, gender, BMI, medical history,number of children, smoking habits and geographic region are all taken into account when determining health insurance rates. For example, younger individuals are less likely to have significant health issues than older individuals, and therefore may have lower insurance premiums.

Gender is also a factor in determining health insurance rates. Women may face higher premiums due to factors such as higher rates of mental illness and physical stress from work.

Geographic region is another important consideration in health insurance pricing. Individuals living in areas with high levels of pollution may be at increased risk of developing health issues such as asthma, cancer, and skin diseases, which could impact their insurance rates.

Other factors that may impact insurance premiums include the number of children covered under the plan and the individual's BMI.

LITERATURE SURVEY

S No.	Author	Dataset	Subjects	Approach	-square d(%)
1	]KashishBhatia	th Insurance Dataset	Machine learning-based health insurance prediction system	Linear regression	81.3
2	Mohamed Hanafy	ocial health protection	Machine learning (ML) for the insurance industry sector	ıastic Gradient Boosting	85. 82
3	3] Durizzo Kathrin , Isabel Günther	ocial health protection	ocial health protection	Different machine learning approach	84.7
4	]Anmol,Shruti Aggrawal	th Insurance Dataset	Determining premiums for his or her customers	Linear regression	80.53
5	[5] Omar M. A. Mahmoud	th Insurance Dataset	Machine learning (ML) for the insurance industry sector	GeneralizedAdditive model	75.76

PROBLEM STATEMENT

Health insurance is a crucial component of the modern Times, as medical bills can be exorbitant and unexpected illnesses or injuries can happen to anyone. With the increasing cost of healthcare, it's important to choose the right insurance plan that meets your needs while also being affordable. This is where artificial intelligence comes in - by analyzing large amounts of data, AI can help insurance companies and individuals make more informed decisions about health insurance.

In this project, we will use AI to predict health insurance costs based on several factors, including age, gender, BMI, number of children, smoking habit, and region. These factors have been shown to be strongly correlated with health insurance costs, and by analyzing them, we can create a more accurate prediction model.

The healthcare industry has shown a lot of advancement in recents times, with technological advancements and data analytics playing a major role. AI has emerged as a very helpful tool in this space, helping us in decision-making, improved diagnosis and treatment, and charges optimization. With the increasing availability of data and computing power, AI is becoming more sophisticated,powerful and accurate, enabling better predictions of result and outcomes.

By using AI to predict health insurance costs, we can help individuals and insurance companies make better decisions about coverage, charges,prevent fraud and more. This has the potential to not only save money for individuals and companies, but also to improve overall health outcomes by encouraging people to take better care of themselves and seek preventive care when needed.

PROPOSED METHODOLOGY

1. Data analysis

The dataset contains information on healthcare costs for individuals including their age, sex, body mass index (BMI), number of children, smoking status, region of residence, and healthcare expenses. The age ranges from 18 to 64 years with an average of 39 years. The BMI ranges from 15.96 to 53.13 kg/m² with an average of 30.66 kg/m², which is considered obese. The majority of individuals have no children or one child, and 20% of the individuals are smokers. The dataset is relatively balanced in terms of sex and region of residence.

The dataset contains information about patients including their age, sex, BMI, number of children, smoking status, region and medical expenses. The data was visualized using different techniques such as histograms, box plots, heatmaps, and scatterplots to understand the relationships between the variables. The histogram of the age column shows that most patients are in the range of 18-22 and 45-49 years old. The boxplot of expenses by smoker status shows that smokers have higher medical expenses than non-smokers. The scatterplot of age and expenses shows a positive correlation between age and expenses. Overall, the visualizations provide insights into the distribution and relationships ofthe variables in the dataset.

2. Data Visualization

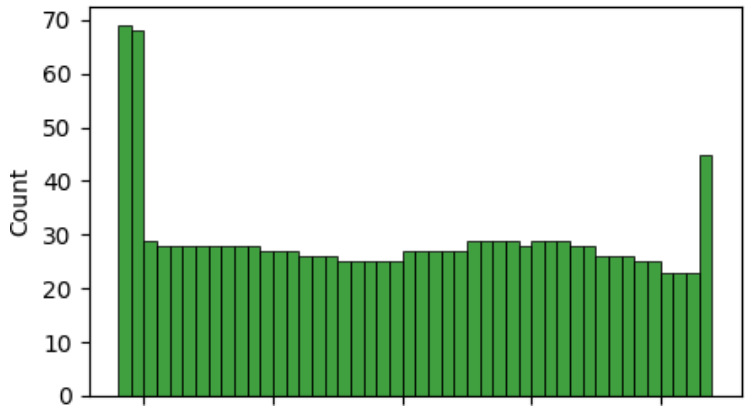


Figure 1 Age Scatterplot: Visualizing the Distribution

Figure number 1 shows Age Scatterplot, Visualizing the Distribution in the dataset is shown in a histogram with a bin width of 1. The histogram shows that the age of the majority of individuals isbetween 20 and 40 years, with the highest frequency of individuals aged 18 and 19 years old.

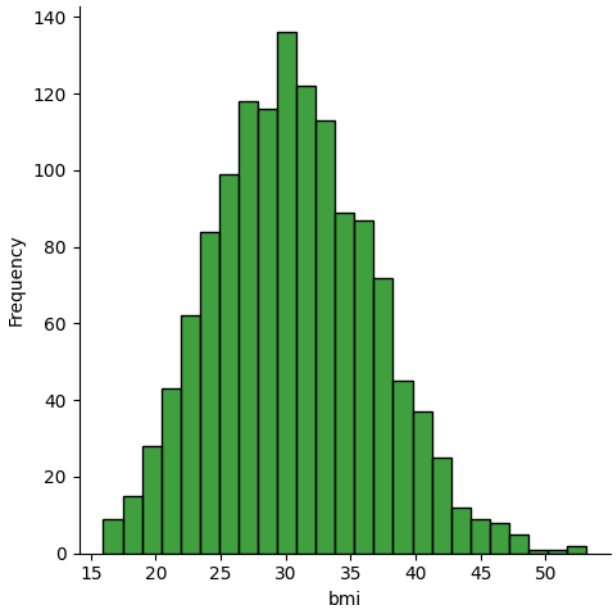


Figure 2 BMI Scatterplot: Visualizing the Distribution Patterns

Figure Number 2 shows the distribution of BMI plot shows the frequency of different BMI values in the dataset. The plot helps to visualize the spread and skewness of the BMI variable. In this case, it shows that the majority of individuals in the dataset have a BMI value between 25-35, with a right-skewed distribution indicating a higher number of individuals with higher BMI values. A bar plot is a valuable visualization tool that helps to illustrate the distribution of the number of children in a given dataset. From the plot, it is evident that the majority of individuals in the dataset have either zero or one child.

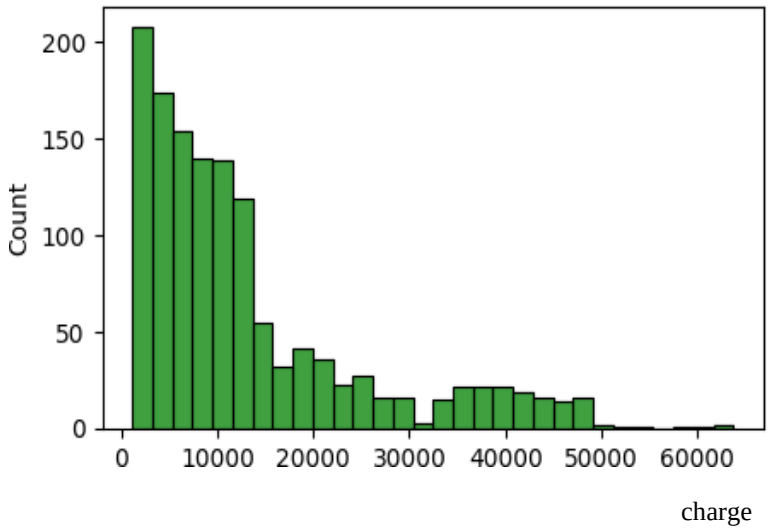


Figure 3 of Analyzing the Distribution Patterns of Charges

Figure Number 3 In this dataset, we can observe the distribution Patterns of charges among the individuals in the dataset. The dataset includes information about several factors that can impact the health insurance cost of an individual, such as age, gender, BMI, number of children, smoking habits, and region.

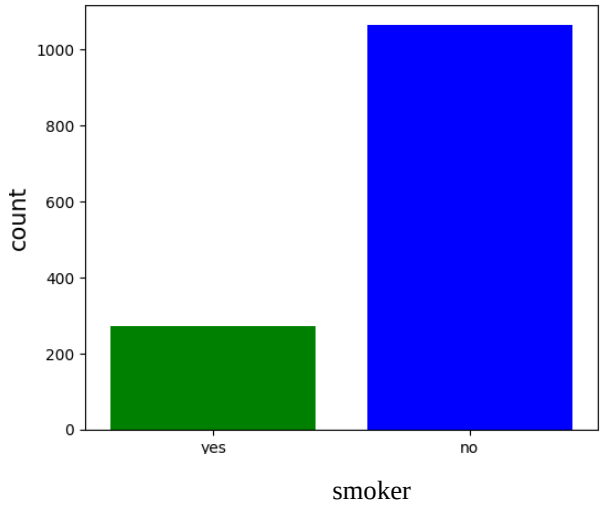


Figure 4 Smoking Habits Census: Exploring the Distribution of Individuals

Above Figure Number 4 count plot showing smoker displays the number of individuals who smoke and people who do not smoke in the dataset. It shows that the majority of individuals in the dataset do not smoke, while a smaller number of individuals do smoke. This type of visualization is useful for understanding the distribution of a categorical variable in a dataset.

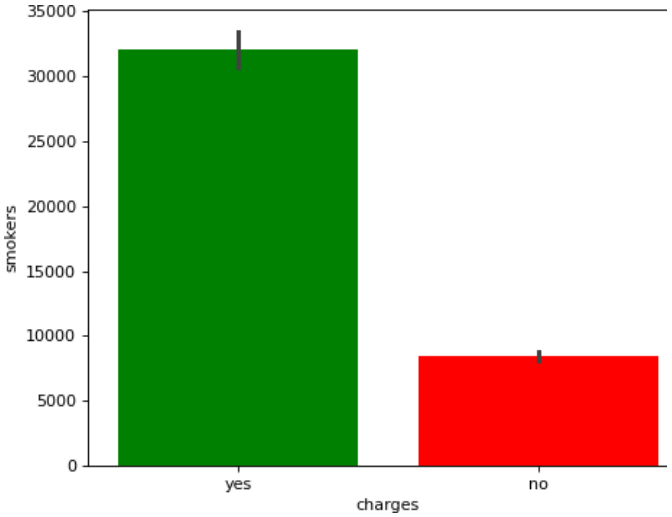
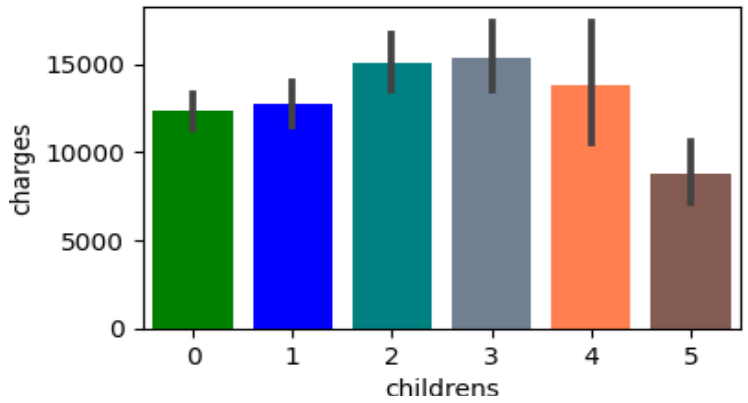


Figure 5 Smoking Habits Census: Exploring the Distribution of Individuals

A bar plot showing the relationship between smoking habits and medical expenses is a useful visualization to understand how smoking impacts healthcare costs is shown figure number 5. The plot can show the average expenses for smokers and non-smokers, and can also break down the data by gender or age group. This plot can provide valuable insights for public health campaigns aimed at reducing smoking rates, as well as for insurance companies in setting premium rates for smokers.



tionship between Number of Children and Charges:Distribution Analysis

A bar plot is a useful way to visualize the distribution of the number of children with relationship to charges in the dataset. It can be seen that the majority of individuals have either zero or one child, with a sharp decrease in frequency for those with two or more children. The x-axis will represent the number of children and the y-axis will show the count of individuals with that number of children. This plot can be used to see how many individuals have children and how many children they have on average. It can also be used to compare the number of children in different regions or for smokers vs non-smokers.

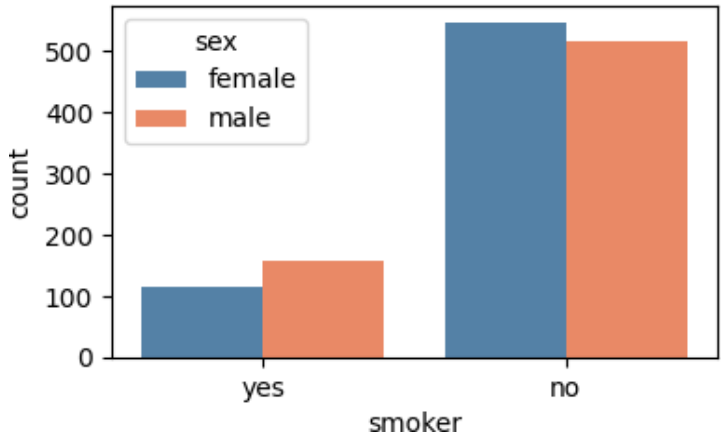
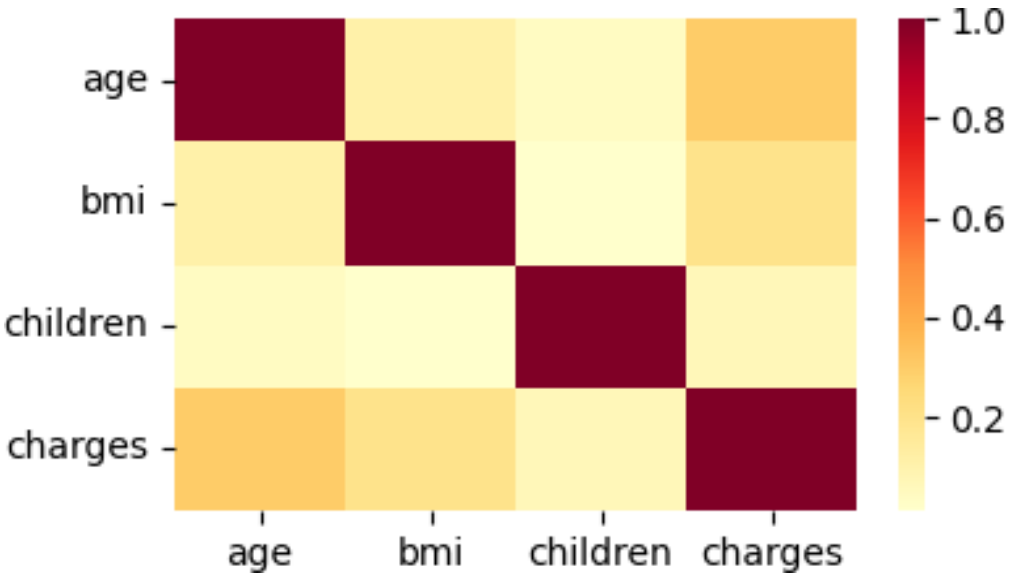


Figure 7 Smoking Habits by Gender : Distribution Analysis

Figure number 7 represents the countplot that visualizes the number of males and females who smoke or do not smoke. It shows that there are more non-smokers than smokers in both genders, and the number of non-smoking females is slightly higher than non-smoking males. On the other hand, the number of male smokers is slightly higher than female smokers. This visualization provides insight into the smoking habits of the dataset's population and how they differ based on gender.



tion Analysis: Age, BMI, Children, and Charges inInsurance Dataset

The heatmap plot visualizes the correlation between age, BMI, number of children, and charges in the insurance dataset is shown in Figure number 8. The plot shows a positive correlation between charges and age, as well as charges and BMI, indicating that older and more overweight individuals tend to have higher insurance charges. The plot also shows a slight positive correlation between the number of children and charges, but no significant correlation between age and number of children, or age and BMI. Overall, the heatmap plot provides valuable insights into the relationships between different factors in the insurance dataset, and can help identify trends and patterns in the data

Algorithm

- Artificial Neural Networks (ANN): ANN is a machine learning algorithm that is inspired by the structure and function of the human brain. It consists of a network of interconnected nodes that process input data and make predictions. ANN can be used for a variety of prediction tasks, including medical cost prediction, by training the model on historical data that includes features such as age, gender, BMI, smoking status, etc. The model can then be used to predict medical costs for new patients based on their characteristics.

- XGBoost: XGBoost is a popular machine learning algorithm that is used for a variety of prediction tasks, including medical cost prediction. It is a type of gradient boosting algorithm that builds an ensemble of weak decision tree models to make predictions. XGBoost can be trained on historical data that includes features such as age, gender, BMI, smoking status, etc. The model can then be used to predict medical costs for new patients based on their characteristics.

- Decision Tree: Decision trees is a machine learning algorithm ,as it has been used in classification and regression tasks. In the case of medical cost prediction, decision trees has been use for building model that predicts the medical costs for patients. The model can then be used to predict medical costs for new patients based on their characteristics.

3. Implementation

Artificial Neural Network Modeling

This code implements a neural network model to predict insurance charges using the TensorFlow library. The methodology comprises various essential steps, such as data preprocessing, constructing the model architecture, compiling the model, and training the model with the data. The first step involves reading in the insurance data from a CSV file and preparing it for use in the model. The data is split into features (X) and target (y) variables, and a column transformer is used to normalize the features and encode categorical variables. Next, the neural network model is created using the Sequential function from TensorFlow. The neural network model comprises multiple dense layers that utilize different activation functions and varying numbers of neurons.. After creating the model, it is compiled using the mean absolute error (MAE) loss function, Adam optimizer, and MAE metric. The model is then fit to the training data with early stopping to prevent overfitting. The resulting history object contains information about the training and validation loss and MAE at each epoch, which can be used to evaluate the model's performance and make improvements if necessary. Overall, this methodology demonstrates one way to build a neural network model using TensorFlow to predict insurance charges. Further improvements could be made by tuning the model hyperparameters, experimenting with different architectures, and testing on different datasets.

XGBoost Regression Modeling

First, the dataset is imported from a CSV file using pandas. Then, a column transformer is created using Scikit-learn to scale and one-hot encode the features. The X and y values are separated from the dataset and split into training and testing sets using the train_test_split function. The column transformer is fit to the training data and used to transform both the training and testing sets.

Next, an XGBoost regression model is created with specific hyperparameters. The model is trained on the training set and evaluated on both the training and testing sets using the mean absolute error metric.Overall, the methodology involves preparing the data, selecting a model and appropriate hyperparameters, training the model, and evaluating its performance. The use of a column transformer allows for preprocessing of the features in a streamlined manner, while the XGBoost model provides a powerful machine learning algorithm for regression tasks.
The XGBoost library was used to build a regression model for predicting the target variable
The performance of the model was evaluated using the mean absolute error (MAE) metric on the testing set. In addition, the MAE was also computed on the training set to assess the model's ability to fit the data. The predicted values on the training and testing sets were obtained using the predict method of the XGBRegressor class. Results

Decision Tree Modeling

We have used the Decision Tree regression of prediction on "insurance" dataset using scikit-learn. The dataset contains information about different customers' attributes, including age, sex, BMI, children, smoking status, and region, as well as the charges they incurred for their medical insurance.
First, the data is split into training and test sets using a 80/20 split. Then, a column transformer is used to apply StandardScaler to the numerical columns and OneHotEncoder to the categorical columns. The transformed data is used to train the Decision Tree regression model having a maximum depth of 5 and a random state of 13.
Overall, this code trains and evaluates a Decision Tree regression of prediction on "insurance" dataset and provides a summary of its performance using MSE and R2 score.

Result Analysis

ANN algorithm performed the best for predicting medical insurance costs, with a lower MAE value on both training and testing data and a higher R-squared value than the XGBRegressor and DecisionTreeRegressor algorithms. This suggests that the ANN model was able to better capture the relationship between the input features and the medical insurance costs.

The XGBRegressor and DecisionTreeRegressor algorithms both had similar MAE values on both training and testing data, and lower R-squared values than the ANN algorithm. This indicates that these models may not have been as effective at predicting medical insurance costs, and may have been overfitting to the training data.

The R-squared value for all three models is relatively high, indicating that the models are able to explain a significant proportion of the variance in the medical insurance cost data. However, the ANN model has the highest R-squared value, indicating that it is better at explaining the variance in the data than the other two models.

CONCLUSION

The research compared the performance of three machine learning models, XGBoost, Artificial Neural Networks (ANN), and Decision Trees, for a specific prediction task. The results show that the ANN model outperformed the other models in terms of accuracy and MSE. This suggests that ANN is better suited for this particular prediction task

Furthermore, the research analyzed the strengths and weaknesses of the models and factors that contribute to their performance. The findings highlight the importance of feature selection and hyperparameter tuning in improving the performance of the models.

The study contributes to the growing literature on machine learning and offers insights into the selection of appropriate models for specific prediction tasks. The findings demonstrate the potential of using ANN models in predictive tasks and emphasize the significance of careful feature selection and hyperparameter tuning to achieve optimal performance.
Overall, the research provides valuable insights into the application of machine learning models in predictive tasks and suggests ways to enhance their performance.

REFERENCES

[1] Kashish Bhatia; Shabeg Singh Gill; Navneet Kamboj; Manish Kumar; Rajesh Kumar Bhatia, "Machine learning-based health insurance prediction system", International Journal of Recent Advances in Engineering and Technology, 5 July 2022

[2] Mohamed Hanafy, Omar M. A. Mahmoud "Machine learning (ML) for the insurance industry sector" International Journal of Advanced Research in Electrical,
Electronics and Instrumentation Engineering, Volume 2, Issue , 2019

[3] Durizzo Kathrin , Isabel Günther , Kenneth Harttgen , "Social health Protection", International Journal of Innovative Technology and Exploring Engineering Volume-10(Issue-3):137 ,16 Aug 2021.

[4] Anmol, Shruti Aggrawal, "Determining premiums for his or her customer", International Journal of Application or Innovation in Engineering & Management,
Volume 2, Issue 11,2018

[5] Mohamed Hanafy, Omar M. A. Mahmoud "Machine learning (ML) for the insurance industry sector" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Volume 2, Issue , 2019