Classification and prediction of healthcare data using Supervised Learning

Achanta, Chandrika Bhargavi*,
[1]*School of Information Technology, University of Cincinnati, Ohio, USA*

*** Chandrika Bhargavi Achanta, Email address: Achantci@mail.uc.edu**

Abstract

Health insurance plays a major role in helping people financially to procure health care benefits who fall ill suddenly at once. It covers financial support for the medical tests and other major health related expenses for who pay the insurance amount over a certain period of time. Predictive analysis is used to know the insights of the given health insurance data throughout the period of time using various machine learning algorithms. However, some of the steps include exploration data analysis and various regression techniques follows linear regression, Decision tree, K fold cross validation, Random Forest, logistic regression for the analysis of health insurance data. The study gives out classification and predictive analysis of the health insurance data by using the five supervised learning models. We performed data cleaning and preprocessing using Exploration data analysis, data visualization, feature engineering, training and testing data, model evaluation on dataset for preprocessing and cleaning the data. Through various plot representations we analyzed the overall health data in an efficient manner. We inspected the performance of each model with the metrics f1 score, precision, recall and support by stacking all the implemented models obtained the results with high accuracy 97% for Random Forest showing the best performance with the related models.

Keywords: Health care, Prediction, Data analysis, Insurance, Evaluation

1. Introduction

In terms of disease detection, predictive analytics, individualized care, drug discovery, and administrative effectiveness, machine learning has a considerable impact on healthcare. It can aid in the early diagnosis of diseases, the identification of people who are at high risk of contracting them, the customization of treatment regimens for specific patients, the acceleration of medication development, and the detection of inefficiencies in the healthcare system. The ability of machine learning to enhance patient outcomes, lower costs, and boost efficiency is ultimately what makes it important in the field of healthcare [1].

Organizations rely on this data to make critical decisions for the expansion of their businesses because the introduction of digital technologies has increased the availability of data from a variety of sources. Many industries, including business, management, healthcare, and government, have been touched by big data [6]. It may take scientific or artificial means to obtain hidden samples and significant insights in the vast amount of gathered data. To uncover these values, a new prototype called Data Intensive Scientific Discovery or Big Data Analytics has arisen.

The main objective of the research is the medical cost prediction and analysis of insurance data through various machine learning algorithms to find the accurate model in determining the insights of the given data for future analysis. It helps the government sector and health care particularly in provision of certain schemes and support to people in rural areas through financial support. The analysis enhances the implementation perspective of various health organizations via provision of resources through insurance and other commodities [8].

Several studies are showing the need of improvement in the health sector with basic needs provision and support in wide range. The struggle for quality hospital care for rural residents never ends well. The government has made some progress, but they fall of the short standard [1]. The significance and contribution of the study attempts to

gauge how well various machine learning models perform on health sector. Moreover, the most accurate model for the analysis of the obtained data is the random forest.

The scope of the taken objective includes exploring huge data and determining the performance of each model through analysis and evaluation. Limitations are using the linear regression model the accuracy is quite low compared to other implementation models. In this research and paper work, we are now working on an effective method in our study to identify those who have a habit of smoking and using medication furtherly for health discrepancies. Data analysis approach can predict the health measures cost taking the health insurance and body measures data as the input.

2. Methodology

The dataset is obtained from an online platform namely Kaggle and is related to the health care sector and charges covered by the health insurance in various aspects of health issues and health degradation. Each data cell with the variable existence has its own uniqueness. The data set which taken as sample is of size 9366, it constitutes the strength of the raw data source given as input. The charges attribute is key in the entire data as it depicts the overall insurance charges covered by the children. As it is the target variable it is sufficient to train the data model and acquire the insights from the data. The CSV file containing the data is being uploaded to the execution platform from the machine location and it is studied by using the python library namely pandas: Data = pd.read_csv('source data') where source data is either path of file location in the machine or the filename which is uploaded directly to the execution platform.

The act of choosing, modifying, and extracting the features from the raw data in order to produce accurate predictions is known as feature engineering. However, to capture the underlying patterns and relationships in the data and increase the accuracy of the machine learning model, feature engineering is the crucial phase in the machine learning workflow. Model based feature selection technique is used to train the model on the dataset and used for selecting the certain features that are most important for training and testing process.

Data scaling is the process of normalizing the data to distribute the data uniformly after splitting the data into train data and test data. Log transformation technique is used for scaling the data in certain range. It reduces the impact of missing data and outliers for further implementation of code without any discrepancies. Moreover, the scaling technique enhances the data analysis criteria with certain variables and values of the given dataset.

As shown below fig.1 depicts the scaled data by taking the categorical variable sex as female and male before splitting the dataset into training and testing data.

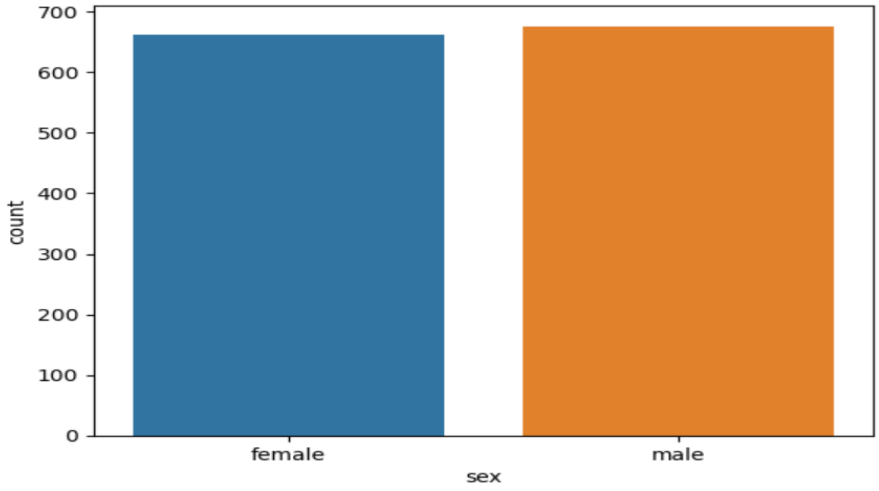

Fig 1. Scaling dataset based on sex

Data visualization plays an important role in exploration data analysis, feature engineering, model selection and evaluation. Different graphic representations and plots are used for exploration data analysis. Model selection and evaluation get processed through confusion matrix, precision recall curves, etc. Likewise, it helps in understanding the data, identification of patterns and extracting insights.

Medical cost prediction and analysis of insurance data can be processed by using various supervised learning implementations which includes Random Forest, decision tree, Logistic regression, linear regression and K fold cross validation. In the correlation analysis giving the variable charges as an independent variable and other attributes as dependent variables as x and y in training and testing of evaluation criteria.

The below figure (fig 2) depicts the correlation analysis by taking independent and dependent variables which includes charges and age of the children attributes which has been given through the health insurance dataset. The joint plot representation showing the variation of charges with respective growth of children by taking the age variable using the library matplotlib.

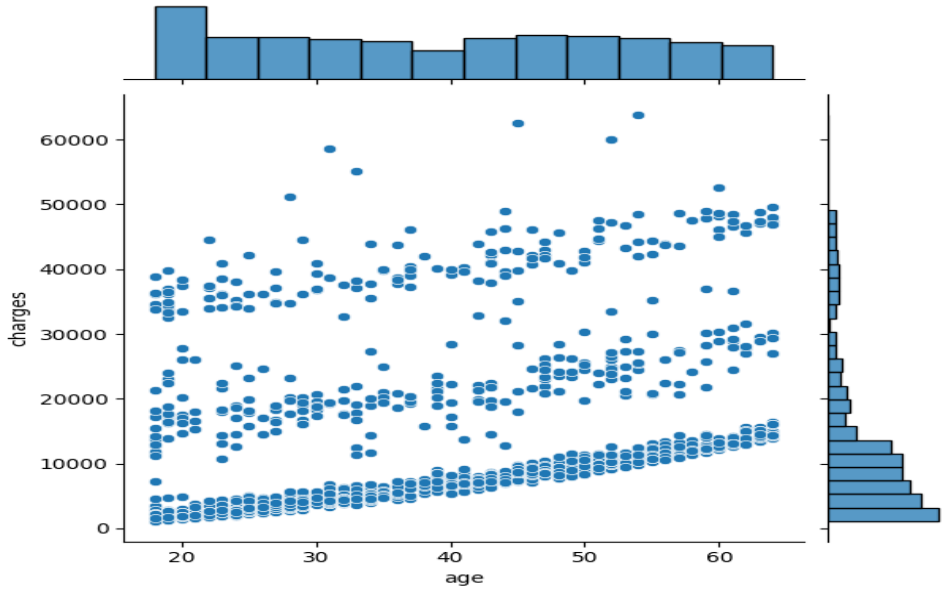

Fig 2. Correlation analysis of children health insurance

The below graphic representation (Fig 3) labelled with charges and density as X and Y labels demonstrates the rise and fall of charges on each individual as per the density given by the X axis of the graph. The line drawn over the graph represents the prediction of Y-test variable in linear regression model.

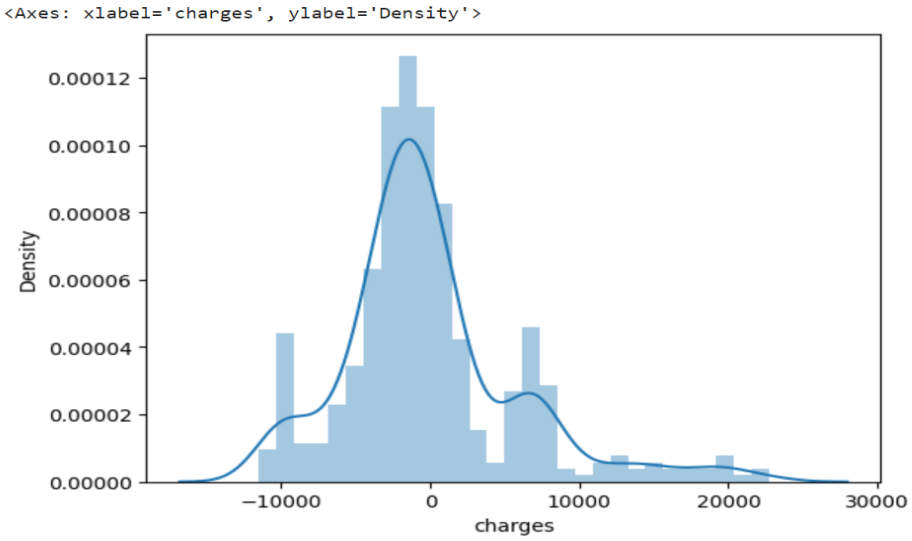

Fig 3. Charges on smoker as per the data

3. Results and Discussion

The supervised learning techniques used for classification and prediction of health data implemented the machine learning algorithms and performed well in the analysis criteria. Through linear regression the charges for the insurance are predicted and it is around 25505 approximately for the insurance data as mentioned in the below figure (fig 4).

```
pred = lr.predict([[19, 27.9,0,1]])
print('value charge to frederick insurance = ', pred)
```

```
value charge to frederick insurance =  [25505.58586529]
```

Fig 4. Prediction of health insurance charges using Linear Regression model

When examining the given data and analysis it is found that the accuracy is determined from the given metrics of the reported analysis of Random Forest model is obtained as 0.98, 0.92 f1 score of two classes. The overall accuracy of the implementation of Random Forest model is in the range of 0.97 out of 1 which is quite high compared to other models.

```
from sklearn.metrics import classification_report
```

```
print(classification_report(y_test, predictions))
```

```
              precision    recall  f1-score   support

           0       0.99      0.97      0.98       356
           1       0.89      0.94      0.92        86

    accuracy                           0.97       442
   macro avg       0.94      0.96      0.95       442
weighted avg       0.97      0.97      0.97       442
```

Fig 5. Random Forest classification and prediction with metrics

As per the obtained results stated in fig 5. of medical cost prediction and analysis through various machine learning algorithms the model implementation and analysis are performed well with high evaluation score for the Random Forest model implementation with 0.97 accuracy (out of 1) compared to other implementation models. The other two models' logistic regression and decision tree obtained the same accuracy i.e., 0.93 (out of 1) by implementation.

Model Name	Accuracy (%)
Logistic Regression	93
Decision Tree	93
Random Forest	97

Table1. Comparison of various models

4. Conclusion & Future work

In the investigation using various implementation models such as Logistic regression, Decision tree, Random Forest on the given health insurance dataset have been constructed to produce accuracies of 93%, 93%, 97% respectively. By using regression models instead of classification results, target values were predicted as exact values. While evaluating the model the mean absolute error and mean squared error were also predicted.

In future, a variety of techniques will be employed, including feature engineering, which entails creating unique features for the data to be used in algorithms, and feature selection, which involves choosing the most significant features to create the model, in order to enhance the model's ability to deliver 100% accuracy for two classes. So that the cost changes which creating a huge impact on accuracy would be minimized [9]. With the usage of unsupervised

learning techniques, prediction of high, low and medium cost is easy to detect the major health issue by taking the smoker attribute on the given source dataset [10].

References

(1) M. D. Samiul Islam, Daizong Liu, Kewei Wang, Pan Zhou, Li Yu, and Dapeng Wu. 2019. A Case Study of HealthCare Platform using Big Data Analytics and Machine Learning. In Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference (HPCCT '19). Association for Computing Machinery, New York, NY, USA, 139–146. https://doi-org.uc.idm.oclc.org/10.1145/3341069.3342980

(2) Himanshu Singh, Moirangthem Biken Singh, Ranju Sharma, Jayesh Gat, Ayush Kumar Agrawal, and Ajay Pratap. 2023. Optimized Doctor Recommendation System using Supervised Machine Learning. In Proceedings of the 24th International Conference on Distributed Computing and Networking (ICDCN '23). Association for Computing Machinery, New York, NY, USA, 360–365. https://doi-org.uc.idm.oclc.org/10.1145/3571306.3571372

(3) S. M. Hasan Mahmud, Md Altab Hossin, Md. Razu Ahmed, Sheak Rashed Haider Noori, and Md Nazirul Islam Sarkar. 2018. Machine Learning Based Unified Framework for Diabetes Prediction. In Proceedings of the 2018 International Conference on Big Data Engineering and Technology (BDET 2018). Association for Computing Machinery, New York, NY, USA, 46–50. https://doi-org.uc.idm.oclc.org/10.1145/3297730.3297737

(4) Amanda H. Gonsalves, Fadi Thabtah, Rami Mustafa A. Mohammad, and Gurpreet Singh. 2019. Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis. In Proceedings of the 2019 3rd International Conference on Deep Learning Technologies (ICDLT '19). Association for Computing Machinery, New York, NY, USA, 51–56. https://doi-org.uc.idm.oclc.org/10.1145/3342999.3343015

(5) Anood Manasrah, Aisha Alkayem, Malik Qasaimeh, and Samer Nofal. 2021. Assessment of Machine Learning Security: The Case of Healthcare Data. In International Conference on Data Science, E-learning and Information Systems 2021 (DATA'21). Association for Computing Machinery, New York, NY, USA, 91–98. https://doi-org.uc.idm.oclc.org/10.1145/3460620.3460738

(6) Saritha K and Sajimon Abraham. 2019. Accuracy evaluation of prediction using supervised learning techniques. In Proceedings of the Third International Conference on Advanced Informatics for Computing Research (ICAICR '19). Association for Computing Machinery, New York, NY, USA, Article 26, 1–6. https://doi-org.uc.idm.oclc.org/10.1145/3339311.3339337

(7) Asif Rahman Snigdha, Syeda Nishat Tasnim, Kamran Rafsan Miah, and Tohedul Islam. 2022. Early Prediction of Heart Attack using Machine Learning Algorithms. In Proceedings of the 2nd International Conference on Computing Advancements (ICCA '22). Association for Computing Machinery, New York, NY, USA, 344–348. https://doi-org.uc.idm.oclc.org/10.1145/3542954.3543004

(8) Younas Khan, Usman Qamar, Nazish Yousaf, and Aimal Khan. 2019. Machine Learning Techniques for Heart Disease Datasets: A Survey. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing (ICMLC '19). Association for Computing Machinery, New York, NY, USA, 27–35. https://doi-org.uc.idm.oclc.org/10.1145/3318299.3318343

(9) Haoran Lyu. 2022. A Machine Learning-Based Approach for Cardiovascular Diseases Prediction. In 2022 14th International Conference on Machine Learning and Computing (ICMLC) (ICMLC 2022). Association for Computing Machinery, New York, NY, USA, 59–66. https://doi-org.uc.idm.oclc.org/10.1145/3529836.3529863

(10) Antarleen Pal and Chandra Prakash. 2022. Personalized Knee Angle Prediction Models Using Machine Learning. In Proceedings of the 2022 Fourteenth International Conference on Contemporary Computing (IC3-2022). Association for Computing Machinery, New York, NY, USA, 149–155. https://doi-org.uc.idm.oclc.org/10.1145/3549206.3549233