# Forecasting Medical Insurance Claim Cost with Data Mining Techniques

MSc Research Project
Data Analytics

Aditya Naresh Sahare
Student ID: x21140677

School of Computing
National College of Ireland

Supervisor: Dr. Cristina Hava Muntean

# National College of Ireland
# Project Submission Sheet
# School of Computing

| | |
|---|---|
| **Student Name:** | Aditya Naresh Sahare |
| **Student ID:** | x21140677 |
| **Programme:** | Data Analytics |
| **Year:** | 2022 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Dr. Cristina Hava Muntean |
| **Submission Due Date:** | 15/12/2022 |
| **Project Title:** | Forecasting Medical Insurance Claim Cost with Data Mining Techniques |
| **Word Count:** | 7122 |
| **Page Count:** | 21 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 1st February 2023 |

## PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | □ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Forecasting Medical Insurance Claim Cost with Data Mining Techniques

Aditya Naresh Sahare
x21140677

**Abstract**

In the healthcare sector, data mining has lately become essential for getting vital information. When developing medical facilities, the cost of health insurance is quite significant. To provide better medical treatment, it is essential to forecast health insurance costs, one of the ways to upgrade medical facilities. Estimating the patient-paid component of health insurance premiums is the essay's focus. The inability to charge each customer a premium enough for the risk they represent is a serious problem for the insurance sector. A Kaggle dataset of 15000 records of customer's medical history in USA is used in research. The selected machine learning techniques are Linear Regression (with and without Log Transformed Dependent Variable), Linear Regression with Interaction of Independent variables, Lasso Regression, Elastic Net Regression and Ridge Regression. The performance was found to best for Linear Regression Model with Interaction of Independent variables with R2 score of 0.79 achieved by the model.

# 1 Introduction

Forecasting consumer healthcare spending has become an important technique for enhancing healthcare accountability. The healthcare industry generates a huge quantity of data on patients, illnesses, and diagnoses, but because it has not been adequately examined, it does not deliver the relevance that it possesses in addition to the patient healthcare cost.(Sommers (2020))

The objective of the medical system is to offer high-quality care to as many individuals as possible, but the cost is important. Medical practitioners, including physicians, pharmaceutical companies, and staff members, must be paid for their job. Due to the fact that these expenditures are typically far more than what a single patient can afford, insurance plans are utilized to divide the costs among all patients in the network and pay for the necessary staff and equipment. An insurance system may occasionally be misused or the target of fraud.

Machine learning can help the health insurance sector in a number of ways. A few of the important factors are quicker claim processing, personalized health insurance plans, affordable insurance plans, the identification of insurance fraud, and improvements in the discovery of new medications. (Bhardwaj et al. (2020)). Data mining methods have allowed scientists to concentrate on creating drugs that are both efficient and suitable for patient needs. It is more cost-effective and time-effective to produce medications that could potentially save lives in an emergency.

The aim of this research is to predict the Insurance Premiums for an individual which will help the insurance companies to predict the claims based on the customer's age, sex, weight, BMI (body mass index), number of dependents(Number of dependent persons on the policyholder), if the customer smokes or not, Blood pressure reading, if the customer has diabetes or not if he does regular exercise, job profile, location, hereditary diseases. Early health insurance cost estimation can aid in more thoughtful consideration of the required amount. where a person may be sure that the quantity they are choosing is appropriate.

Based on the work of (Iqbal et al. (2021)), this research is extended with the exploration of Generalized Linear Regression with multiple experiments on target variable (claims of health insurance) and further tuning of the model with Regularisation techniques. The model's performance was then evaluated using key performance metrics such as RMSE (Root Mean Square Error) $R^2$ score, and adjusted $R^2$ score. The proposed machine learning model may be useful for hospitals, patients, insurance companies, and doctors, helping them to carry out their duties more swiftly and efficiently.

## 1.1 Research Question and objective

**Research Question:** *How accurately we can predict Health Insurance premiums based on customers medical background and yearly premium expense with the help of Linear Regression and Regularization techniques to assist healthcare personnel in spending more time delivering appropriate treatment, lowering burnout among medical experts?*

**Objectives**

- Gather the dataset and perform Data Pre-processing for model building
- Implement Linear Regression Model with variations
- Evaluate the results and find out which model is best for predicting Insurance claim

The main goal of this idea is to assist the healthcare business in forecasting the cost of an insurance claim for a specific person and expediting the health insurance procedure. Section 2 of this discusses the Related work on the topic of Insurance Premiums and Linear Regresssion Model. Section 3 provides the brief explanation of Methodology. Section 4 defines the workflow of the Data analysis. Section 5 provides information of Implementation of the Machine Learning Model. Evaluation of all the models are carried out in Section 6. Conclusion and Future work is discussion in Section 7.

# 2 Related Work

(Christobel and Subramanian; 2022) suggested a model to forecast how much someone's health insurance will cost. In order to identify the ideal collection of variables to use when creating prediction models, they have analyzed the effects of variables assessed on several scales. This paper concentrates on regression analysis to predict the insurance amount by contrasting methodologies such as Linear, Ridge, Lasso, and Polynomial Regression. With an accuracy of 88% (Christobel and Subramanian (2022)), polynomial regression offers good accuracy for forecasting insurance payments. The report's drawback is that the assessment measure "Adjusted R2," which is equally significant to R2 score, has not been taken into account.

Bauder et al. (2016) presented a model to identify when doctors submit false claims for medical insurance. In their research, they have offered some helpful data that may be used to evaluate whether and when healthcare providers are acting in ways that are outside the scope of their expertise, which may be an indicator of abuse, fraud, or misunderstanding of billing procedures. The patient billing information and medical history are included in the dataset, which is taken from the US Medicare system. Five cross-validation calculations are performed to evaluate the model's multinomial Naive Bayes approach, which includes calculations for precision, recall, and F-score. The algorithm can predict several different types of doctors with an F-score of over 0.9 (Bauder et al.; 2016). These results show that it is possible to classify doctors into their various fields using only the services they charge for and advanced analytics in a creative way.

In order to reduce the use of human resources and financial losses, Kowshalya and Nandhini (2018) proposed a model to forecast fraudulent insurance claims and computed insurance premium amounts for various consumers based on their personal and financial information. To evaluate the best fit model from the Random Forest, Naive Bayes, and J48 classifiers, they tested with test-train split of the dataset using 50:50 and 66:34 split with 10 cross validations. They contrasted the model's accuracy using both raw and processed data. Under the 66:34 test-train split dataset, Random Forest obtained 99.41% accuracy. The usage of test-train slits of just 50:50 and 66:34, while acceptable for small datasets, is a drawback of this work. Thus, given a dataset of 15,000 rows, we conclude that a 70:30 test-train split is the best option in our scenario which will help to train the model accurately to perform well on test data.

In order to forecast the risk level of life insurance applicants, Mustika et al. (2019) presented the Extreme Gradient tree Boosting (XGBoost) machine learning model. They utilized data from the Prudential Life Insurance Company that was made accessible through the open-source Kaggle repository. It comprises the applicant data necessary for completing a life insurance application. Data that has undergone pre-processing lacks interpretation. They implemented the XGBoost, Decision Tree, and Random Forest models, with XGBoost coming out on top with a 60.7% accuracy rate. The results may have been improved if the author had done more research on learning rate and alpha value. Finding the best value for alpha for the model is feasible with the aid of GridsearchCV.

Morid et al. (2017) conducted a study of the literature on healthcare cost prediction and discovered that cost-on-cost prediction outperforms cost prediction using clinical data and cost data, on par with or better. Furthermore, it was discovered that supervised learning techniques were more accurate predictors. To analyze the data from Utah Health Plans, they used Gradient Boosting, ANN, Ridge regression, Support Vector Machine, ElasticNet, Lasso and Linear regression, and Random Forest model with an accuracy of 92.9%, gradient boosting offers the greatest cost-on-cost prediction models overall, while ANN offers the best performance for patients with higher costs with an accuracy of 94%. This study's combination of a thorough assessment of the literature and an empirical comparison of the various supervised learning techniques mentioned in the literature is one of its key strengths. Their evaluation of cutting-edge supervised learning techniques that hadn't been studied in this paper for cost-on-cost prediction in healthcare is another strength. The primary drawback of this study is the use of only one data set. More studies using various data sets from various organizations and areas might produce more reliable proof of the relative efficacy of various algorithms.

Panay et al. (2019) provided an interpretable regression technique based on the Dempster-Shafer theory (Kaltsounidis and Karali (2020)) (commonly known as the theory of belief function), employing the Evidence Regression model and a discount function depending on the contribution of each dimension.When employing data from electronic medical records from the Japanese Tsuyama Chou Hospital, two separate algorithms—Gaussian Blur and Artificial Neural Networks (ANN)—taught each dimension's relevance throughout the training phase. These approaches' accuracy was 40% and 35%, respectively. Evidence Regression, reported by Petit-Renaud and Denoeux, which predicts the cost of medical treatment with a 44% accuracy, increased the model's performance Petit-Renaud and Denœux (2004). To overcome the problem of predicting the cost of medical treatment on a bigger dataset (10000+ rows), the author advised creating a cost bucket and classify them.

Freyder (2016) examined data from customers of Gateway Health Plan (a healthcare provider in Pennsylvania) before and after a timeline of 6 months worth of costs for those who had an inpatient stay in a hospital (patients who are admitted through a formal doctor's order), allowing them to assess whether they could reduce healthcare costs. An accuracy of 77.6% was obtained by the author using linear regression. They came to the conclusion that patient gender and the average cost prior to an incident together predict more effectively the average cost post-event. The model deviates from the assumptions of linear regression, which is a weakness of this study. Both the normality of the residuals and the homoscedasticity assumption are broken by this model.

(Kaushik et al.; 2022) examined the artificial neural network (ANN) and data mining techniques for predicting insurance premiums in the healthcare industry. Artificial intelligence and machine learning may be used to analyze and assess large data sets in order to simplify the health insurance procedure. AI will handle time-consuming activities, enabling insurance experts to focus on actions that will improve consumers' experiences. The model's performance was then evaluated using key performance metrics such as RMSE (Root Mean Square Error), MSE (Mean Square Error), MAE (Mean Absolute Error), r2, and modified r2. The accuracy of their model was 92.72 percent.(Kaushik et al.; 2022)

(El Bouanani et al.; 2022) Given the high cost and mortality risk associated with the patient's selected class of treatment, (El Bouanani et al.; 2022) suggested a model for anticipating readmissions of patients to intensive unit care (i.e. readmitted or not to the hospital). The study presents a three-stage machine learning-assisted approach that makes use of both support vector machine (SVM) and artificial neural network (ANN) techniques. In order to increase the likelihood of accurately classifying patients and preventing misclassification, generalized logistic regression was used first, followed by generalized sequential pattern (GSP). Then, two best-fit algorithms—the support vector machine and the artificial neural network—were applied to raise the total accuracy to 87% and 85%, respectively.(El Bouanani et al.; 2022)

The framework proposed by (Agarwal and Tripathi; 2022) is built using an IOT and deep learning approach for a transportation networks scenario. According to analysis, ML-based fault forecasting is more accurate and can guarantee that the insurance amount generated is correct. The classic technique's accuracy climbs linearly with the amount of damages, reaching 80The paper by (Goundar et al.; 2020) presents a study on the use of artificial neural networks (ANNs) for predicting health insurance claims. The authors begin by discussing the importance of accurate claim prediction in the health insurance industry, highlighting its role in reducing costs and improving the overall efficiency of the claims process.Next, the authors describe their methodology, which involves using a dataset of health insurance claims and associated features, such as the type of claim and the patient's age and gender. They use this dataset to train an ANN model and evaluate its performance in terms of accuracy and precision.The results of the study show that the ANN model is able to achieve a high level of accuracy and precision in predicting health insurance claims. The authors also compare the performance of the ANN model to other machine learning models, such as decision trees and support vector machines, and find that the ANN model performs significantly better.Overall, this study provides evidence that ANNs can be effectively used for predicting health insurance claims, and can potentially improve the efficiency and cost-effectiveness of the claims process. However, the authors note that further research is needed to evaluate the performance of ANNs in other contexts and to compare their performance to other machine learning models.
After carefully reading through the whole aforementioned literature survey, it is clear that a lot of academics employ Random Forest, Decision tree, Nearest Neighbor (kNN),Linear Regression, AdaBoost, K- XGBoost, and AdaBoost. The interaction of important variables will be the main emphasis of this study, and the use of linear regression in various aspects will be thoroughly investigated with the goal of improving the model's accuracy.

# 3 Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) includes a hierarchical and iterative process model, as well as a framework that can be expanded using a generic-to-specific approach. It begins with six phases and then delves further into general and then specialized jobs stated in Figure 1. The approach is adaptable to accommodate a range of formality levels that may vary in DM projects of different sizes and levels of complexity.(Niaksu (2015))

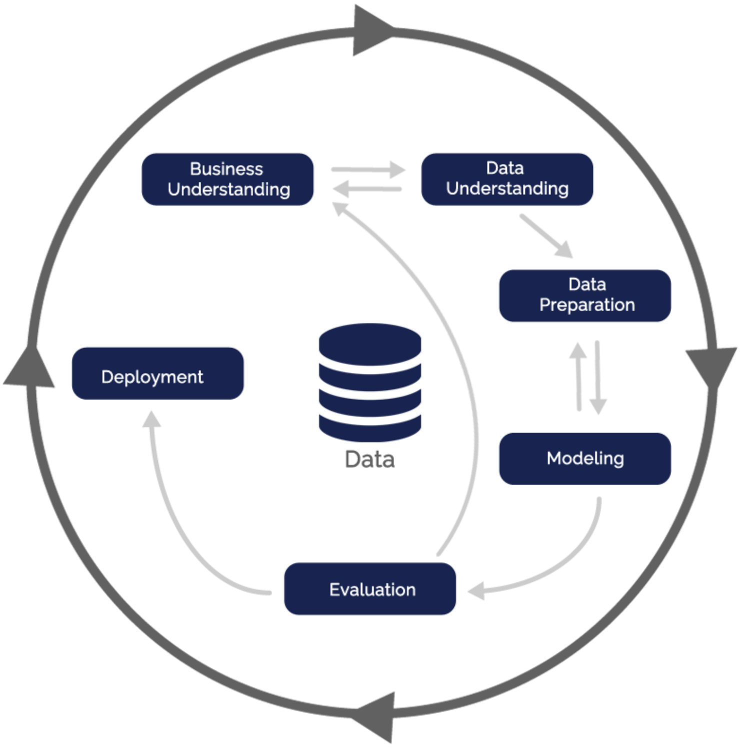


Figure 1: Methodology

[1]

## 3.1 Business Understanding

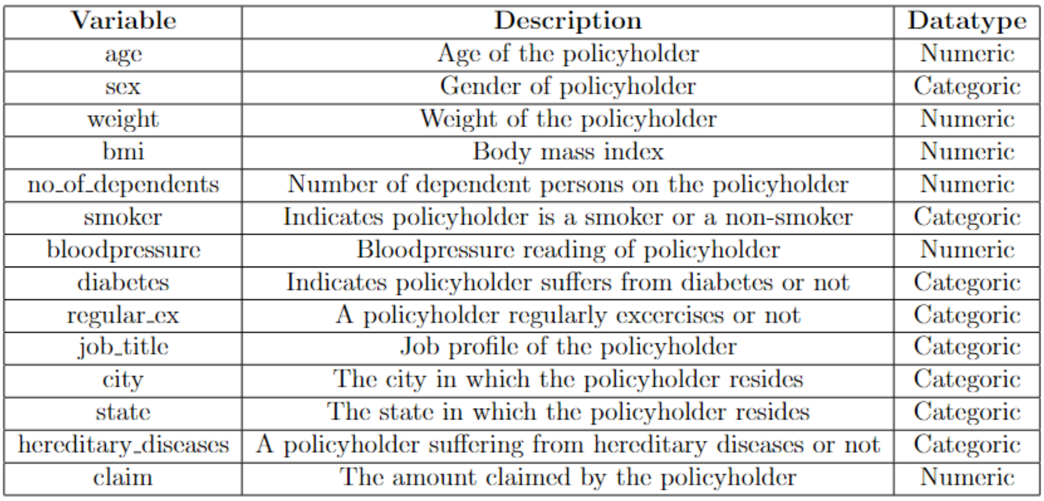
Predicting health insurance premiums using machine learning algorithms is still a topic that needs further research and exploration in the medical care industry. Even as the healthcare sector digitizes more and more, enormous amounts of data will inevitably be created and gathered. The health insurance industry may achieve a number of objectives with the help of AI and machine learning. Personalized health insurance plans, more inexpensive insurance options, the capacity to spot insurance fraud, advances in medication discovery, and quicker claim processing are just a few of the critical elements.

## 3.2 Data Understanding

This stage begins with data collection and getting familiar with the data. The dataset was extracted Kaggle data repository [2]. The data consists of 15000 rows and 14 variables. This data will be divided into 70:30 ratio for training and testing the model. The dataset consists of customer's personal medical records and details taken by the Insurance company for initializing their premium amount.

[1] www.otaris.com

[2] `https://www.kaggle.com/datasets/sureshgupta/health-insurance-data-set`

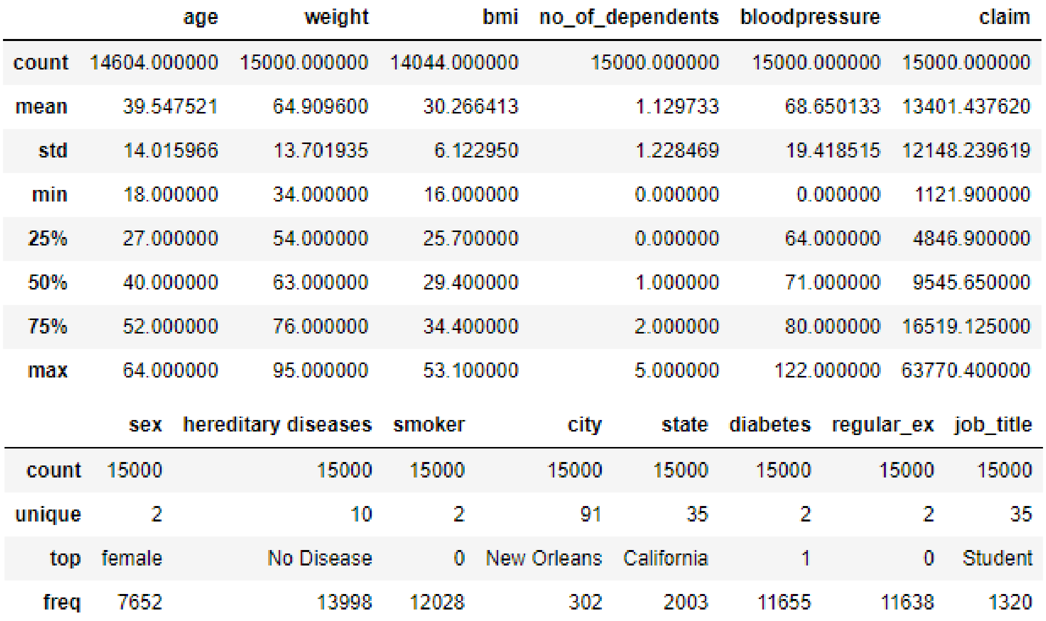


| Variable | Description | Datatype |
|---|---|---|
| age | Age of the policyholder | Numeric |
| sex | Gender of policyholder | Categoric |
| weight | Weight of the policyholder | Numeric |
| bmi | Body mass index | Numeric |
| no_of_dependents | Number of dependent persons on the policyholder | Numeric |
| smoker | Indicates policyholder is a smoker or a non-smoker | Categoric |
| bloodpressure | Bloodpressure reading of policyholder | Numeric |
| diabetes | Indicates policyholder suffers from diabetes or not | Categoric |
| regular_ex | A policyholder regularly excercises or not | Categoric |
| job_title | Job profile of the policyholder | Categoric |
| city | The city in which the policyholder resides | Categoric |
| state | The state in which the policyholder resides | Categoric |
| hereditary_diseases | A policyholder suffering from hereditary diseases or not | Categoric |
| claim | The amount claimed by the policyholder | Numeric |

Figure 2: Data Description

## 3.3 Data Preparation

### 3.3.1 Data Analysis and Pre-Processing

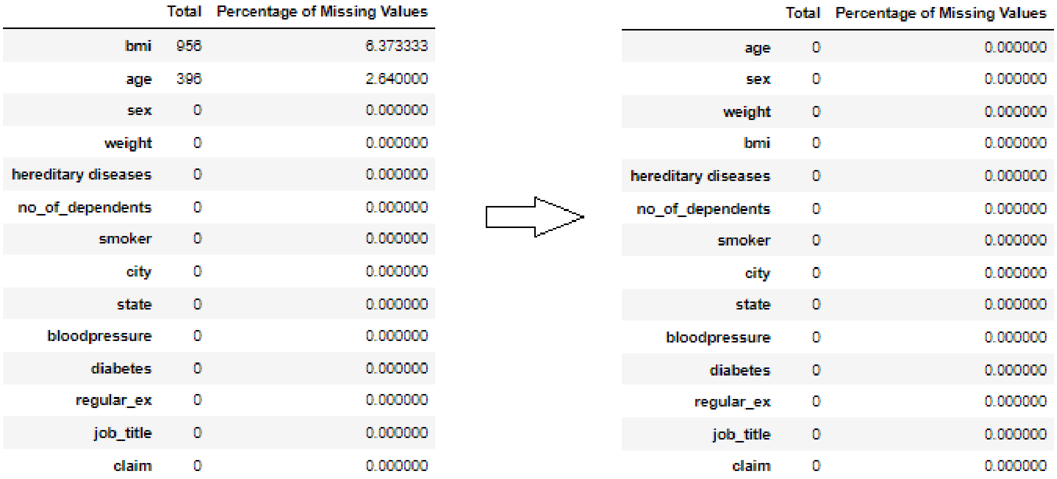


| | age | weight | bmi | no_of_dependents | bloodpressure | claim |
|---|---|---|---|---|---|---|
| count | 14604.000000 | 15000.000000 | 14044.000000 | 15000.000000 | 15000.000000 | 15000.000000 |
| mean | 39.547521 | 64.909600 | 30.266413 | 1.129733 | 68.650133 | 13401.437620 |
| std | 14.015966 | 13.701935 | 6.122950 | 1.228469 | 19.418515 | 12148.239619 |
| min | 18.000000 | 34.000000 | 16.000000 | 0.000000 | 0.000000 | 1121.900000 |
| 25% | 27.000000 | 54.000000 | 25.700000 | 0.000000 | 64.000000 | 4846.900000 |
| 50% | 40.000000 | 63.000000 | 29.400000 | 1.000000 | 71.000000 | 9545.650000 |
| 75% | 52.000000 | 76.000000 | 34.400000 | 2.000000 | 80.000000 | 16519.125000 |
| max | 64.000000 | 95.000000 | 53.100000 | 5.000000 | 122.000000 | 63770.400000 |

| | sex | hereditary diseases | smoker | city | state | diabetes | regular_ex | job_title |
|---|---|---|---|---|---|---|---|---|
| count | 15000 | 15000 | 15000 | 15000 | 15000 | 15000 | 15000 | 15000 |
| unique | 2 | 10 | 2 | 91 | 35 | 2 | 2 | 35 |
| top | female | No Disease | 0 | New Orleans | California | 1 | 0 | Student |
| freq | 7652 | 13998 | 12028 | 302 | 2003 | 11655 | 11638 | 1320 |

Figure 3: Descriptive Statistics of all features

The output in Figure 3 illustrates the summary statistics of all the numeric variables like the mean, median(50%), minimum, and maximum values, along with the standard deviation. Note, the average age of a policyholder claiming the insurance is 39 years. The claim amount is between 1121 to 63770. Here the mean BMI of a policyholder is 30 (the healthy BMI range is between 16 to 24.9) and the average weight is 64.



| | Total | Percentage of Missing Values |
|---|---|---|
| bmi | 956 | 6.373333 |
| age | 396 | 2.640000 |
| sex | 0 | 0.000000 |
| weight | 0 | 0.000000 |
| hereditary_diseases | 0 | 0.000000 |
| no_of_dependents | 0 | 0.000000 |
| smoker | 0 | 0.000000 |
| city | 0 | 0.000000 |
| state | 0 | 0.000000 |
| bloodpressure | 0 | 0.000000 |
| diabetes | 0 | 0.000000 |
| regular_ex | 0 | 0.000000 |
| job_title | 0 | 0.000000 |
| claim | 0 | 0.000000 |

| | Total | Percentage of Missing Values |
|---|---|---|
| age | 0 | 0.000000 |
| sex | 0 | 0.000000 |
| weight | 0 | 0.000000 |
| bmi | 0 | 0.000000 |
| hereditary_diseases | 0 | 0.000000 |
| no_of_dependents | 0 | 0.000000 |
| smoker | 0 | 0.000000 |
| city | 0 | 0.000000 |
| state | 0 | 0.000000 |
| bloodpressure | 0 | 0.000000 |
| diabetes | 0 | 0.000000 |
| regular_ex | 0 | 0.000000 |
| job_title | 0 | 0.000000 |
| claim | 0 | 0.000000 |

Figure 4: Dealing with Missing Values

If we observe the count of all the variables in Figure 3, there is less count for variable age and BMI than other variables. So we can say that there are missing values in these variables. The missing values are handled by replacing them with the mean of the variable 'age' and 'bmi' as can be seen in Figure 4. Also, the minimum blood pressure is zero, which is invalid. Hence, the zeros were replaced by the median of the values in blood pressure.

### 3.3.2 Exploratory Data Analysis

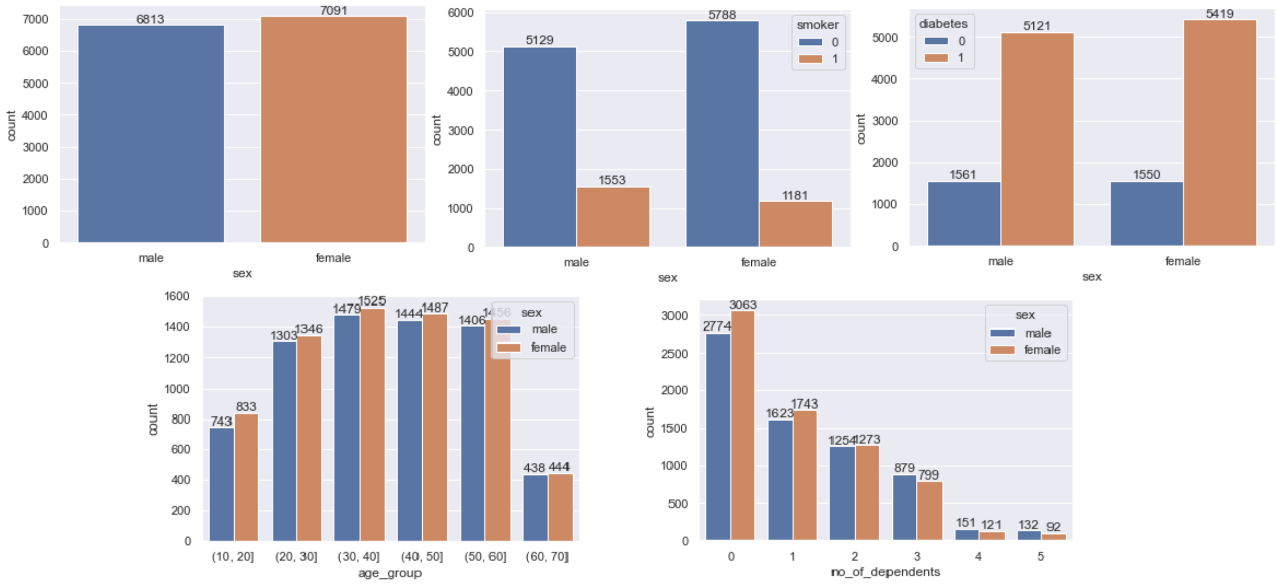


Figure 5: Dealing with Missing Values

A thorough insight into gender-based client distribution is provided by the aforementioned EDA(Exploratory Data Analysis) in Figure 5. Male and female representation in the dataset is roughly equal, as can be observed. The statistics revealed that diabetics and non-smokers were more prevalent than previously thought. Data for the age variable have a normal distribution. Ages 30 to 50 make up the majority of clients. A small percentage of customers have 4 or 5 dependents, whereas the majority of them have none.

### 3.3.3 Feature Engineering

There is a total of 91 cities in the 'city' variable. A new variable 'Region' is created with the 'city' variable. These 91 cities are divided into 4 regions named 'North-East', 'Southern', 'Mid-West', and 'West'.In Figure 6, it can be observed that for both males and females insurance premium claims are increasing with the increase in age. The distribution of claims between the two categories, 'smoker'(1) and 'non-smoker'(0), are distinct enough to take smokers as a potentially good predictor of the claim amount. The distribution of claims between the two categories, 'smoker'(1) and 'non-smoker'(0), are distinct enough to take smokers as a potentially good predictor of the claim amount. We can see that a 'non-smoker' has a median claim amount of around 10000 while a 'smoker' has a median claim of 40000.

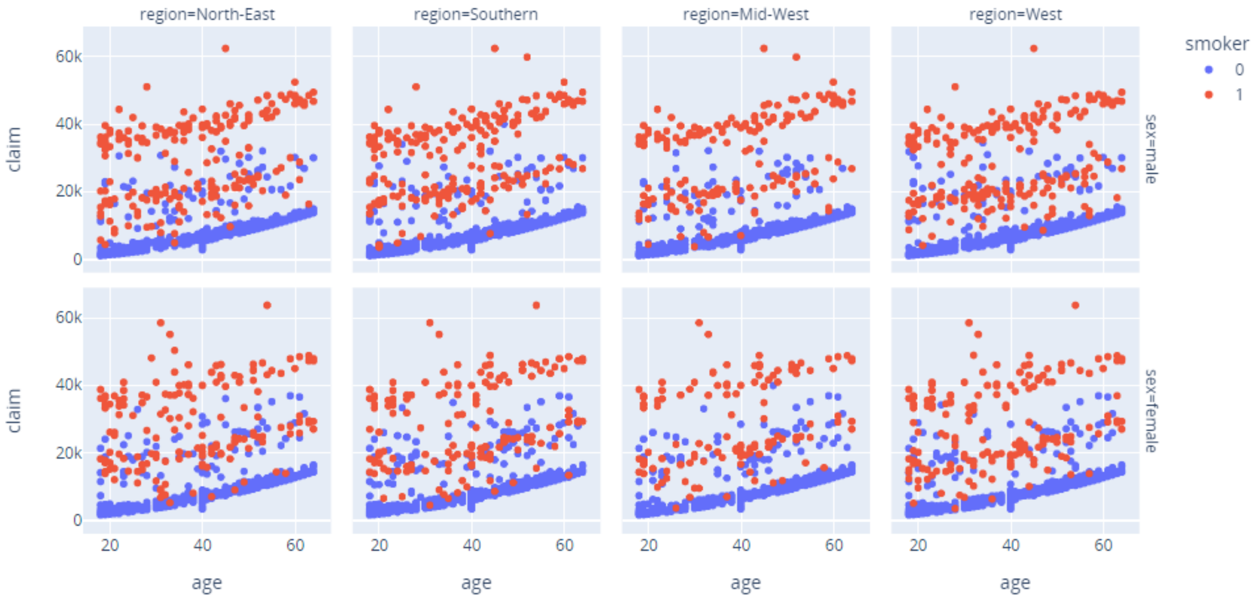


Figure 6: Analysis of new Feature "Region" vs claim, age and smoker

In Figure 7, it can be observed that the insurance claim is significantly increasing with an increase in BMI of the customers. The healthy range of BMI is considered to be between 18.5 to 24.9 (Klein et al. (2007)). When a person's BMI is above 30 and they smoke, their insurance rates are often over $40,000, according to observations. Customers who smoke but have a healthy BMI pay less in premiums.

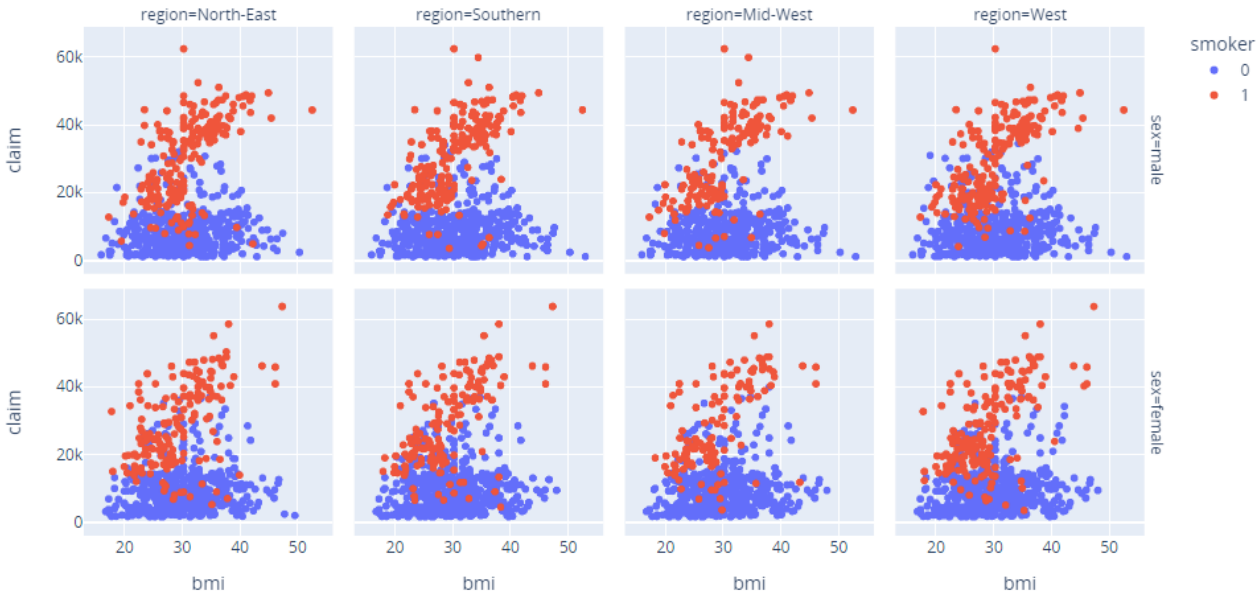


Figure 7: Analysis of new Feature ”Region” vs claim, bmi and smoker

## 3.4 Modelling

**Linear Regression:** A statistical method called linear regression is used to simulate the linear connection between a dependent variable and one or more independent variables. It is based on the linear equation, which states the following relationship between the dependent variable (Y) and the independent variables (X):

$$Y = b_0 + b_1 * X_1 + b_2 * X_2 + ... + b_n * X_n \tag{1}$$

Where b0 is the intercept, and b1, b2, ..., bn are the regression coefficients.

By reducing the total of squared deviations between the observed and predicted values of the dependent variable, linear regression calculates the values of the regression coefficients. The ensuing regression equation is known as the least squares regression equation, and the approach used is known as the least squares technique.

**Regularization:**

In order to minimize the adjusted loss function and avoid over-fitting or under-fitting, regularization refers to methods for calibrating machine learning models. Regularization allows us to properly adapt our machine learning model to a specific test set, hence lowering the mistakes in the test set.

**Lasso and Ridge Regression:** Lasso and Ridge regression are both types of regularized linear regression, which are used to address the problem of multicollinearity in linear regression. Multicollinearity occurs when the independent variables in a regression model are highly correlated and can lead to unstable and unreliable estimates of the regression coefficients.

Lasso and Ridge regression both address multicollinearity by adding a penalty term to the cost function that is used to train the model. This penalty term, called the regularization term, is designed to penalize models with large coefficients, which can help to reduce the impact of multicollinearity and improve the stability of the model.

The Lasso and Ridge regression equations are similar to the linear regression equation but include the regularization term in addition to the terms for the independent variables. The Lasso regression equation is of the form:

$$Y = b0 + b1 * X1 + b2 * X2 + ... + bn * Xn + lambda * \sum \mid bi \mid \tag{2}$$

Where lambda is the regularization parameter, which controls the strength of the regularization, and the summation is over all the coefficients of the independent variables.

**Elastic Net Regression:** Elastic Net regression is a type of regularized regression method that combines the penalties of both L1 and L2 regularization. It is called "Elastic Net" because it can be seen as a combination of both L1 and L2 regularization, which are typically represented by a "net"-like structure.

The regularization term for Elastic Net regression is defined as follows:

$$\lambda * (\alpha * L1\_norm + (1 - \alpha) * L2\_norm) \tag{3}$$

where $\lambda$ is the regularization parameter,  is a mixing parameter that determines the balance between L1 and L2 regularization, and L1_norm and L2_norm are the L1 and L2 norms of the model weights, respectively.

In other words, Elastic Net regression is a linear regression model with a regularization term that penalizes both the L1 and L2 norms of the model weights. This can help to prevent overfitting and improve the generalization performance of the model.

## 3.5 Evaluation

In this final stage, the performance of a model in forecasting Medical Insurance Claim cost is compared with other models based on the evaluation metrics RMSE, R-Squared, and Adjusted R-Squared.

- RMSE(Root Mean Square Error): Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. It indicates the absolute fit of the model to the data.

- R-Squared: A statistical fit indicator known as R-Squared quantifies how much variance in a dependent variable is explained by one or more independent variables in a regression model.

- Adjusted R-Squared: By taking into account the influence of extra independent factors that have the propensity to distort the outcomes of R-squared measurements, adjusted R-squared, a modified form of R-squared, increases accuracy and dependability.

The metrics used for the analysis of the machine learning model give the complete picture of their performance over the train and test dataset. The difference between $R^2$ and Adjusted $R^2$ is that both explain how much variance in the dependent variable is explained by the independent variable but Adjusted $R^2$ determines whether any addition of the independent variable decrease the accuracy of the model or not. If there is less difference in $R^2$ and Adjusted $R^2$ then it states that all the independent variables/predictors are significant. Root mean square error is useful in the scenario of forecasting the results based on historical data. It gives the standard deviation of residuals, the measure which explains how far the predicted values are from the actual values. P-value gives the significance of the regression model. Durbin Watson explains if there is any auto-correlation between the predictor/independent variables.

### 3.6 Deployment

The project's ultimate goal is not to build modeling; rather, this final stage is putting its research and analysis into a written format that can be easily read. Despite the fact that modeling is meant to provide additional detail to the data, this knowledge still has to be arranged and presented in such a manner that customers can utilize it. The likelihood cost of claims for the insurance industry can only be effectively reduced by disclosing the forecast to the decision maker.

## 4 Design Specification

Following data pre-processing, machine learning techniques will be implemented. The implementation of multiple linear regression will take many factors into account. with or without the interaction of significant factors, deleting insignificant variables, and dependent variables that have been log-transformed. The performance will be improved by hyperparameter optimization. The following procedure will be used by these algorithms in Figure 8.

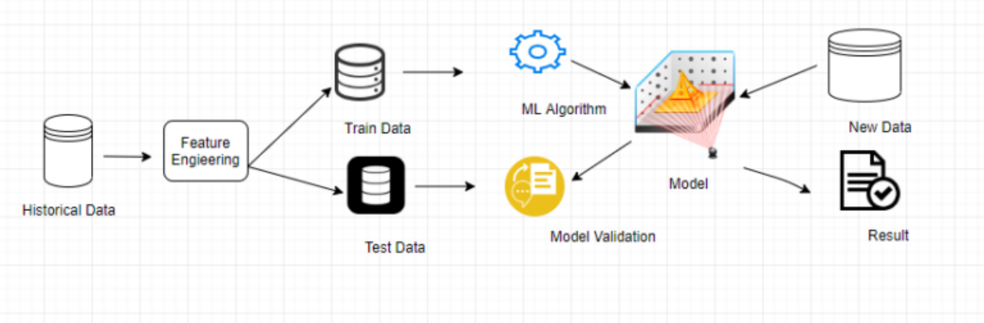


Figure 8: Workflow diagram

## 5 Implementation

All the data analysis is done in Python language on Jupyter notebook. The Data preparation is done with the application of the mean, and median of the variables. The models are implemented with the Ordinary Least Square method. The ordinary least squares (OLS) method is a linear regression analysis method that estimates the parameters of a linear regression model by minimizing the sum of the squares of the differences between the observed responses in the dataset and the responses predicted by the linear model. The goal of OLS is to find the line that best fits the data.(Whatley (2022)) The output gives are $R^2$ score of the model which depicts the value, what percentage variance of the target variable is explained by the model. The linear Regression model is explored in different scenarios. In order to determine if a hyper-parameter adjustment will improve the model's accuracy, regularization techniques were investigated.

# 6 Evaluation

This section uses machine learning techniques to evaluate the results of the model built. All the processes of evaluation are conducted in Python to run the Machine learning algorithms. The metrics used to evaluate the regression models are $R^2$ score, Adjusted $R^2$ score, Root Mean Square Error, P-value, and Durbin Watson test.

When using a multiple linear regression model, the dependent variable must be numerical and have a minimum of two independent variables. Let's say that the function of the dependent variable translates to other variables and some random noise.

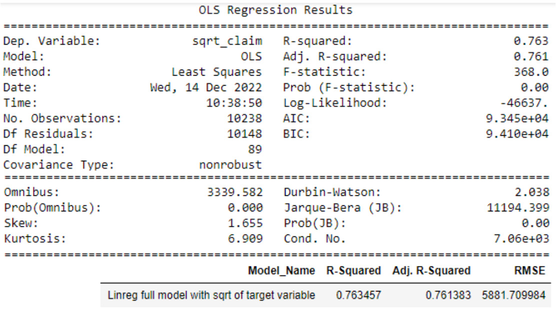
The model creates data using the format y $= \beta_0 + \beta_1 x1 + \beta_2 x2 + ... + \beta_k \mathrm{x_k} + \epsilon$ where the dependent variable is y, the independent variables are $\beta_0 + \beta_1 x1 + \beta_2 x2 + ... + \beta_k \mathrm{x_k}$, is random noise (error) and $\beta_i$ is the contribution value of the independent variables. When all of the independent variables are zero, the dependent variable's value is represented by the y-intercept.

## 6.1 Experiment 1: Multiple Linear Regression - Full Model - with Square Root Transformed Dependent Variable

In this section, a full model with linear regression has been built using OLS (Ordinary Least Square) technique. The full model indicates that all the independent variables have been considered that are present in the dataset.

**Independent Variable:** age, weight, bmi, no_of_dependents, blood pressure, sex, hereditary diseases, smoker, state, diabetes, regular_ex, job_title, region

**Dependent Variable:** claim



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             sqrt_claim   R-squared:                       0.763
Model:                            OLS   Adj. R-squared:                  0.761
Method:                 Least Squares   F-statistic:                     368.0
Date:                Wed, 14 Dec 2022   Prob (F-statistic):               0.00
Time:                        10:38:50   Log-Likelihood:                -46637.
No. Observations:               10238   AIC:                         9.345e+04
Df Residuals:                   10148   BIC:                         9.410e+04
Df Model:                          89
Covariance Type:            nonrobust
==============================================================================
Omnibus:                     3339.582   Durbin-Watson:                   2.038
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            11194.399
Skew:                           1.655   Prob(JB):                         0.00
Kurtosis:                       6.909   Cond. No.                     7.06e+03
==============================================================================
                                 Model_Name  R-Squared  Adj. R-Squared         RMSE
 Linreg full model with sqrt of target variable   0.763457        0.761383  5881.709984
```

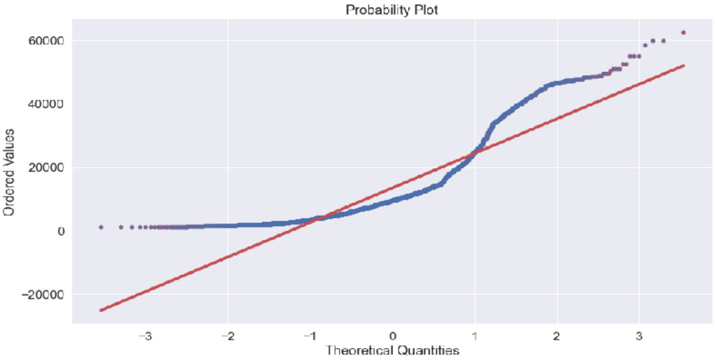
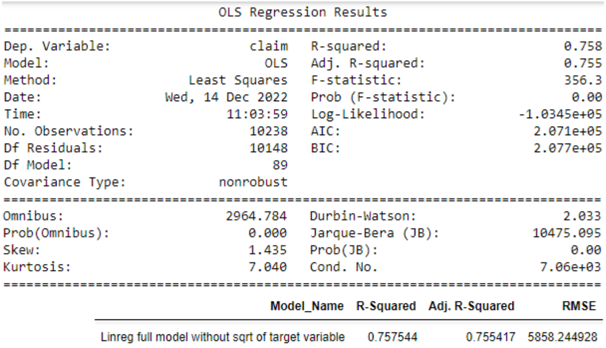


Figure 9: Results of Experiment 1

From the results in Figure 9, This model explains 76.3% of the variation in the dependent variable 'claim'. The Durbin-Watson test statistics is 2.029 and indicates that there is no autocorrelation. Condition Number 7.06e+03 suggests that there is severe collinearity. The Q-Q plot depicts the predicted claim point in blue color and shows how close it is to the actual points ie. the best-fit model line.

## 6.2 Experiment 2: Multiple Linear Regression - Full Model - without Square Root Transformed Dependent Variable

In this section, any kind of transformation on the dependent variable is not considered, the dependent variable ’claim’ is used as it is.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  claim   R-squared:                       0.758
Model:                            OLS   Adj. R-squared:                  0.755
Method:                 Least Squares   F-statistic:                     356.3
Date:                Wed, 14 Dec 2022   Prob (F-statistic):               0.00
Time:                        11:03:59   Log-Likelihood:            -1.0345e+05
No. Observations:               10238   AIC:                         2.071e+05
Df Residuals:                   10148   BIC:                         2.077e+05
Df Model:                          89
Covariance Type:            nonrobust
==============================================================================
Omnibus:                     2964.784   Durbin-Watson:                   2.033
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10475.095
Skew:                           1.435   Prob(JB):                         0.00
Kurtosis:                       7.040   Cond. No.                     7.06e+03
==============================================================================
```

| Model_Name | R-Squared | Adj. R-Squared | RMSE |
|---|---|---|---|
| Linreg full model without sqrt of target variable | 0.757544 | 0.755417 | 5858.244928 |

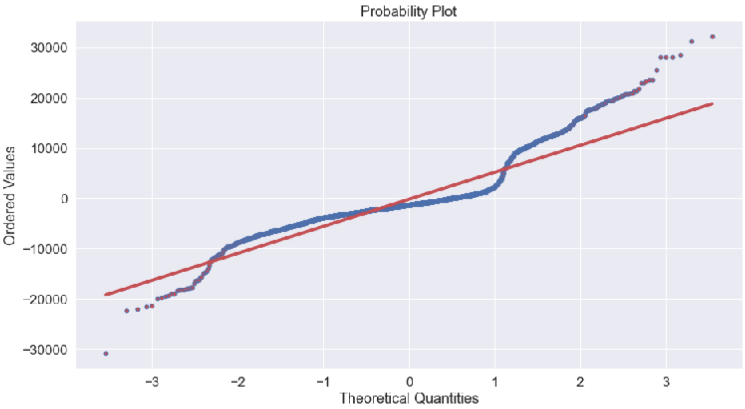


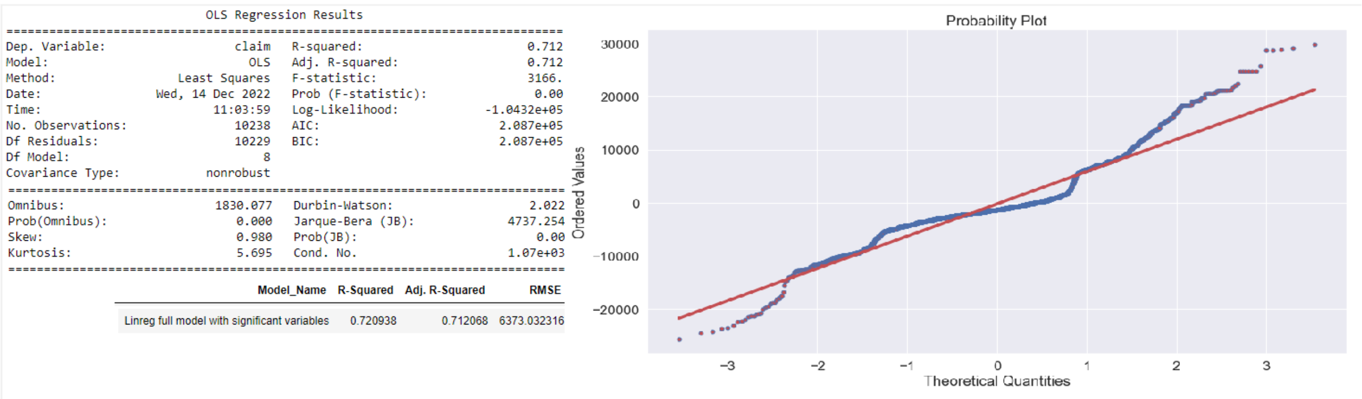
Figure 10: Results of Experiment 2

This model explains 75.8% of the variation in the dependent variable ’claim’. The Durbin-Watson test statistics is 2.033 and indicates that there is no autocorrelation. Condition Number 7.06e+03 suggests that there is severe collinearity. The Q-Q plot depicts the predicted claim point in blue color and shows how close it is to the actual points ie. the best-fit model line. On comparing the above models in Experiment 1 and 2, it is seen that the R-squared and the Adjusted R-squared value for the model considering square root transformation of the variable ’claim’ is lower than the other model. And, the RMSE value of the model without considering the square root transformation is considerably lower. So, we continue with variable ’claim’ as it is, instead of opting for log transformation

## 6.3 Experiment 3: Linear Regression with Significant Variable

After comparing the P-value in Experiment 2, it was found that ’sex’, ’job_title’, ’region’, ’state’, and ’hereditary diseases’ features are insignificant to predict the dependent variable ’claim’ as they have P-value greater than 0.05.

Occam’s razor is a principle to explain the phenomena by the simplest hypothesis possible. The last model where the insignificant variables are removed is performing very close to the other models in spite of having a lesser number of variables. Using Occam’s razor principle, the model is accepted in which we consider the model with significant variables.(Blumer et al. (1987))
This model in explains 71.2% of the variation in the dependent variable ’claim’. The Durbin-Watson test statistics is 2.022 and indicates that there is no autocorrelation. Condition Number 1.07e+03 suggests that there is severe collinearity. Since the accuracy is low for this Experiment than Experiment 2 and 3. This model is rejected.



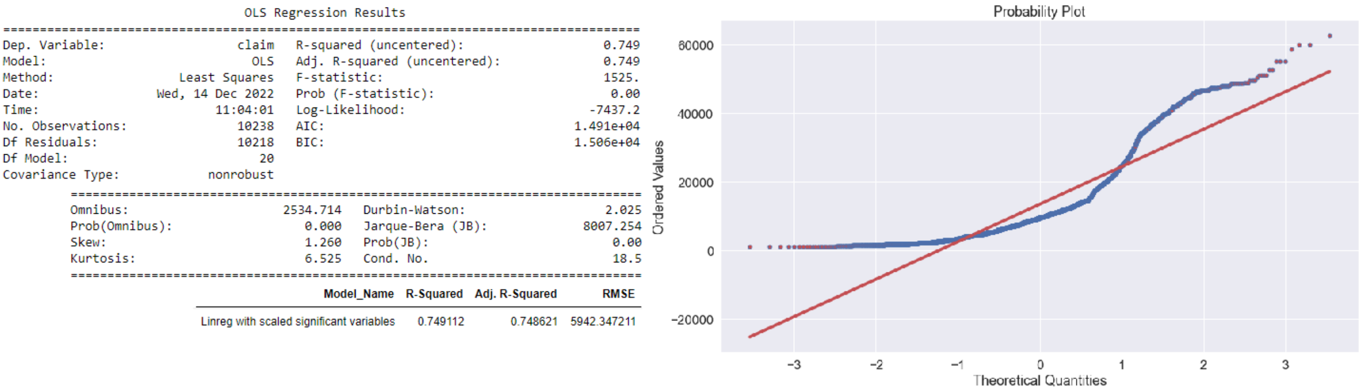
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  claim   R-squared:                       0.712
Model:                            OLS   Adj. R-squared:                  0.712
Method:                 Least Squares   F-statistic:                     3166.
Date:                Wed, 14 Dec 2022   Prob (F-statistic):               0.00
Time:                        11:03:59   Log-Likelihood:            -1.0432e+05
No. Observations:               10238   AIC:                         2.087e+05
Df Residuals:                   10229   BIC:                         2.087e+05
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
Omnibus:                     1830.077   Durbin-Watson:                   2.022
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4737.254
Skew:                           0.980   Prob(JB):                         0.00
Kurtosis:                       5.695   Cond. No.                     1.07e+03
==============================================================================
```

| Model_Name | R-Squared | Adj. R-Squared | RMSE |
|---|---|---|---|
| Linreg full model with significant variables | 0.720938 | 0.712068 | 6373.032316 |

Figure 11: Results of Experiment 3

## 6.4 Experiment 4: Linear Regression with Significant Variable on scaled Data

In this Experiment, the standardization of data is performed as the numerical variables are not normally distributed. After standardization, the insignificant variables from Experiment 3 are removed. The processed data is then fed to the Linear Regression model. The model is tested on unscaled predicted values.



```
                            OLS Regression Results
=======================================================================================
Dep. Variable:                  claim   R-squared (uncentered):                   0.749
Model:                            OLS   Adj. R-squared (uncentered):              0.749
Method:                 Least Squares   F-statistic:                              1525.
Date:                Wed, 14 Dec 2022   Prob (F-statistic):                        0.00
Time:                        11:04:01   Log-Likelihood:                         -7437.2
No. Observations:               10238   AIC:                                  1.491e+04
Df Residuals:                   10218   BIC:                                  1.506e+04
Df Model:                          20
Covariance Type:            nonrobust
         ==============================================================================
         Omnibus:                     2534.714   Durbin-Watson:                   2.025
         Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8007.254
         Skew:                           1.260   Prob(JB):                         0.00
         Kurtosis:                       6.525   Cond. No.                         18.5
         ==============================================================================
```

| Model_Name | R-Squared | Adj. R-Squared | RMSE |
|---|---|---|---|
| Linreg with scaled significant variables | 0.749112 | 0.748621 | 5942.347211 |

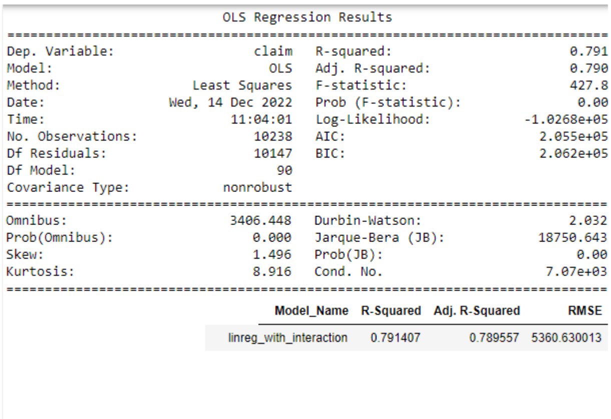
Figure 12: Results of Experiment 4

The model accuracy on scaled data is 74.9% on predicting the insurance premiums for the customers. The Durbin-Watson test statistics is 2.025 and indicates that the is no autocorrelation. The RMSE for this experiment is higher than the dependent variable transformed experiment(Experiment 1 and Experiment 2). The $R^2$ and Adjusted $R^2$ value has slightly declined for scaled data. Hence in further experiments, unscaled data are used.

## 6.5 Experiment 5:Linear Regression with Interaction

In this Experiment in Figure 13, a new variable is introduced with the interaction of BMI and smokers. The interaction of 2 variables provides a new possibility of predicting the dependent variable 'claim'. Any 2 significant variables can be used in this case.

Linear regression with interaction terms can be useful in situations where the relationship between the dependent and independent variables is not well-represented by a simple linear model. By adding interaction terms to the model, you can capture non-linear relationships and potentially improve the model's fit and predictive ability. It helps to improve the fit and predictive ability of your model, especially in situations where the relationship between the dependent and independent variables is more complex than a simple linear relationship.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  claim   R-squared:                       0.791
Model:                            OLS   Adj. R-squared:                  0.790
Method:                 Least Squares   F-statistic:                     427.8
Date:                Wed, 14 Dec 2022   Prob (F-statistic):               0.00
Time:                        11:04:01   Log-Likelihood:            -1.0268e+05
No. Observations:               10238   AIC:                         2.055e+05
Df Residuals:                   10147   BIC:                         2.062e+05
Df Model:                          90
Covariance Type:            nonrobust
==============================================================================
Omnibus:                     3406.448   Durbin-Watson:                   2.032
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            18750.643
Skew:                           1.496   Prob(JB):                         0.00
Kurtosis:                       8.916   Cond. No.                     7.07e+03
==============================================================================
```

| Model_Name | R-Squared | Adj. R-Squared | RMSE |
|---|---|---|---|
| linreg_with_interaction | 0.791407 | 0.789557 | 5360.630013 |

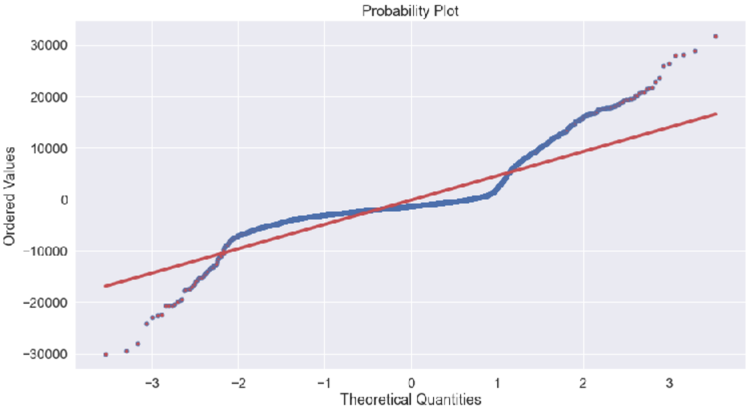


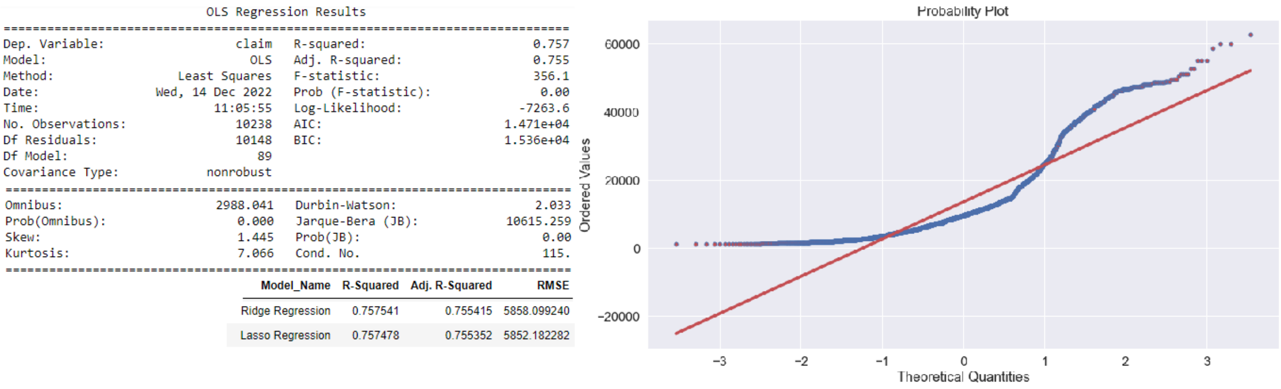
Figure 13: Results of Experiment 5

This model explains 79.1% of the variation in the dependent variable 'claim'. This is the best model so far with RMSE of 5360. This value of RMSE states that it has less residual values. In the Q-Q plot in Figure 13, from quantile -2 to 2, the data points pass from the best line fit. That means all the values between this distribution are more or less correctly predicted. In the range outside -3 and 3 quartile, the data points are drifting apart but that is less in number and can be ignored in terms of a large dataset going across 20000 rows and more. The Durbin-Watson test statistics is 2.032 and indicates that the is no autocorrelation. Condition Number 7070 suggests that there is severe collinearity. The collinearity is likely to increase because of the interaction effect.

## 6.6 Experiment 6:Lasso and Ridge Regression

In this Experiment of Regularization technique in which L1 and L2 Regularization is performed, the dataset used in this scaled dataset ie. normalized data. The alpha value is the penalty term added to the Linear Regression model which allows the less important features to get eliminated and reduces the complexity and multi-collinearity in the model. The value of alpha generally lies between 0 to infinity. The larger the value, the more aggressive the penalization is. In this model, Sum of Squares, RMSE, R2, Adjusted R2 value is generated and the alpha value is picked for which the RMSE is the least. In this case, for both Lasso and Ridge Regression, the best alpha value is 0.0001.

For both Lasso and Ridge Regression, the R2 value is 75% which makes it similar to Linear Regression with a Scaled dataset. The RMSE value depicts the residuals of predicted vs actual claims which are between 5850 to 5860. Hence, the Regularization technique doesn't perform well than Linear regression with interaction on this data set.



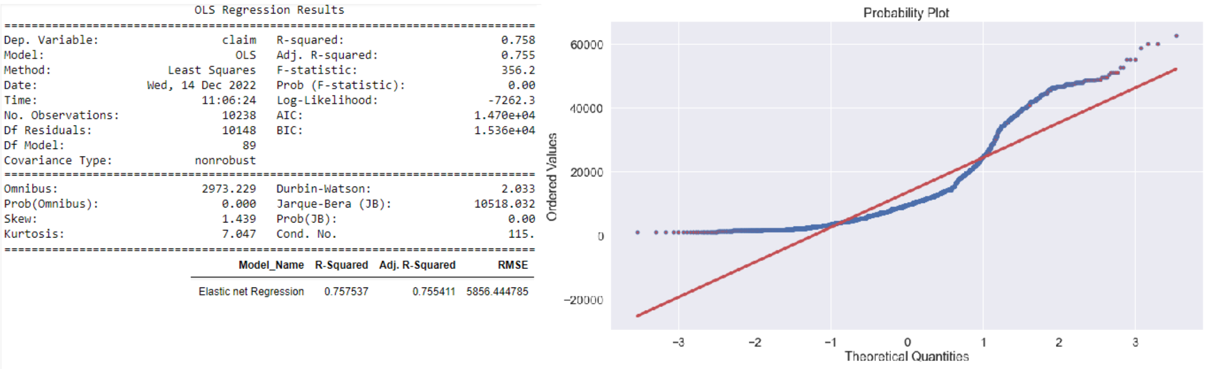
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  claim   R-squared:                       0.757
Model:                            OLS   Adj. R-squared:                  0.755
Method:                 Least Squares   F-statistic:                     356.1
Date:                Wed, 14 Dec 2022   Prob (F-statistic):               0.00
Time:                        11:05:55   Log-Likelihood:                -7263.6
No. Observations:               10238   AIC:                         1.471e+04
Df Residuals:                   10148   BIC:                         1.536e+04
Df Model:                          89
Covariance Type:            nonrobust
==============================================================================
Omnibus:                     2988.041   Durbin-Watson:                   2.033
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10615.259
Skew:                           1.445   Prob(JB):                         0.00
Kurtosis:                       7.066   Cond. No.                         115.
==============================================================================
```

| Model_Name | R-Squared | Adj. R-Squared | RMSE |
|---|---|---|---|
| Ridge Regression | 0.757541 | 0.755415 | 5858.099240 |
| Lasso Regression | 0.757478 | 0.755352 | 5852.182282 |

Figure 14: Results of Experiment 6

## 6.7 Experiment 7: Elastic Net Regression

In this Experiment, the elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. The model is trained using both L1 L2(Lasso and Ridge Regression) which allows learning of sparse model where few entries are zero similar to Lasso and also maintaining the regularization properties similar to ridge regression. The parameter for Elastic Net regression is selected with the help of GridsearchCV. It is supposed to be between 0 to 1. For Ridge Regression the L1_ratio is taken as 1 and for Lasso Regression it is 0. The best parameter for L1_ratio for Elastic Net Regression is 0.2 according to the GridSearchCV technique with 10 cross-validations. The alpha value is set to 0.0001 because it gives the minimum RMSE value after applying the model. The R2 value of this experiment is 75.5% which is similar to Ridge and Lasso Regression. Overall, the regularization technique doesn't improve the model's accuracy.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  claim   R-squared:                       0.758
Model:                            OLS   Adj. R-squared:                  0.755
Method:                 Least Squares   F-statistic:                     356.2
Date:                Wed, 14 Dec 2022   Prob (F-statistic):               0.00
Time:                        11:06:24   Log-Likelihood:                -7262.3
No. Observations:               10238   AIC:                         1.470e+04
Df Residuals:                   10148   BIC:                         1.536e+04
Df Model:                          89
Covariance Type:            nonrobust
==============================================================================
Omnibus:                     2973.229   Durbin-Watson:                   2.033
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10518.032
Skew:                           1.439   Prob(JB):                         0.00
Kurtosis:                       7.047   Cond. No.                         115.
==============================================================================
```

| Model_Name | R-Squared | Adj. R-Squared | RMSE |
|---|---|---|---|
| Elastic net Regression | 0.757537 | 0.755411 | 5856.444785 |

Figure 15: Results of Experiment 7

## 6.8 Discussion

Table1 compares the $R^2$, Adjusted $R^2$, Root Mean Square Error(RMSE) of all the 7 models in the variation of multiple Linear Regression models. Linear Regression with the Interaction of 2 variables turns out to be the best model for the given data of customers

Table 1: Results of all the Models Applied

| Model Name | R-Squared | Adj.R-Squared | RMSE |
|---|---|---|---|
| Linear Reg with Sq. root of target variable | 0.763457 | 0.761383 | 5881.709984 |
| Linear Reg without Sq. root of target variable | 0.757544 | 0.755417 | 5858.244928 |
| Linear Reg with Significant variables | 0.720938 | 0.712068 | 6373.032316 |
| Linear Reg with Scaled Significant variables | 0.749112 | 0.748621 | 5942.347211 |
| Linear Reg with Interaction of 2 variables | 0.791407 | 0.789557 | 5360.630013 |
| Ridge Regression | 0.757478 | 0.755415 | 5858.099240 |
| Lasso Regression | 0.757478 | 0.755352 | 5852.182282 |
| Ridge Regression | 0.757537 | 0.755411 | 5856.444785 |

medical records. The model accuracy is 79.1% on the test data set with an RMSE of 5360.63. For other models, the accuracy is more or less similar to each other. Among these models, the transformation of the target variable has performed well with an accuracy of 76.3%. Overall, the normalization of the dataset has not improved the model. Regularization of the model generally improves the model by penalizing the insignificant variables to reduce the complexity of the model. But in this case, it is similar to the normal Linear Regression model with just 75% accuracy. Hence, The model Linear Regression with Interaction of 2 variables can be used by an Insurance company to predict the claim for future customers.

# 7 Conclusion and Future Work

The intended study will significantly advance the field of healthcare. Everybody has a different type of health insurance. Every healthcare center aspires to provide high-quality care to as many patients as possible. Insurance companies must guarantee that all patients receive adequate care at a fair price in order to pay for all personnel and medical expert expenditures. The application of machine learning algorithms by the insurance company can hasten the process of submitting an insurance claim for treatment. The purpose of this study was to investigate several Linear Regression approaches to help healthcare professionals spend more time providing adequate care and reduce burnout among medical professionals. The best model in this research was Linear Regression with the Interaction of 2 variables with an accuracy of 79.1%.

In addition to helping insurers receive the right proportion for their health care premiums, doing this will strengthen the bonds of trust between them and the insurance companies, encouraging them to keep paying their premiums. These forecasting studies, which focus on healthcare companies that provide insurance would assess and give the aforementioned choice.

- Which insurance policy would be appropriate and desired by various insurer types?
- How much of insurance cost should be based on a particular patient and behavior?
- How smoking cigarettes can affect the cost of insurance?
- How can insurance providers and insurers establish strong, trusting relationships?

As a result, gathering and analyzing insurer qualities is a very valuable informational process for health insurance businesses. Insurance firms need both internal and external data sources that are pertinent to the segmentation of the insurance industry. To create a distinctive personality for precision marketing, a thorough study of consumer data and behavior will be used to foresee and comprehend increasing client demands and offer relevant items at affordable prices.

In future work, this research can be extended in exploring Stochastic Gradient Descent model which is another hyperparameter tuning tool. The dataset consists of biased data with more customers with No hereditary diseases. Hereditary disease is something that stays in the genes and is sometimes passed to future generations. This might help the doctors and Insurance companies to provide the customer with proper checks and treatment opportunities. Other than that, there is more number of students whose insurance premium is less than $2000-3000. This makes it difficult to predict the insurance premiums for those customers whose premium exceeds a higher amount of $40000. The regularization technique might have performed well if the dataset was large and unbiased. This will be a valuable addition to future studies, helping insurance firms gain a better understanding of various individuals' behaviors and attributes and providing more precise forecasts on the cost of health insurance premiums and risk management.

## 7.1 Acknowledgement

I would like to sincerely thank Dr.Cristina Hava Muntean for their input, time, and patience. She provided her experience and expertise. I would also like to thank all the authors I have referred to this study. I have learned a great deal about machine learning with their help. I would also like to thank my fellow colleagues for their help brainstorming over the techniques to apply and get rid of errors.

# References

Agarwal, D. and Tripathi, K. (2022). A framework for structural damage detection system in automobiles for flexible insurance claim using iot and machine learning, *2022 International Mobile and Embedded Technology Conference (MECON)*, IEEE, pp. 5–8.

Bauder, R. A., Khoshgoftaar, T. M., Richter, A. and Herland, M. (2016). Predicting medical provider specialties to detect anomalous insurance claims, *2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI)*, IEEE, pp. 784–790.

Bhardwaj, N., Anand, R. and Gupta, A. D. (2020). Health insurance amount prediction international journal of engineering research & technology, *(IJERT)* **9**(05).

Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M. K. (1987). Occam's razor, *Information processing letters* **24**(6): 377–380.

Christobel, Y. A. and Subramanian, S. (2022). An empirical study of machine learning regression models to predict health insurance cost, *Webology (ISSN: 1735-188X)* **19**(2).

El Bouanani, F., Qaraqe, K. A. et al. (2022). Mathematical modeling and optimal stopping theory-based additional layers for 30-day rate risk prediction of readmission to intensive care units.

Freyder, C. (2016). *Using linear regression and mixed models to predict health care costs after an inpatient event*, PhD thesis, University of Pittsburgh.

Goundar, S., Prakash, S., Sadal, P. and Bhardwaj, A. (2020). Health insurance claim prediction using artificial neural networks, *International Journal of System Dynamics Applications (IJSDA)* **9**(3): 40–57.

Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M. A., Sajid Ullah, S. et al. (2021). A computational intelligence approach for predicting medical insurance cost, *Mathematical Problems in Engineering* **2021**.

Kaltsounidis, A. and Karali, I. (2020). Dempster-shafer theory: $\eta$ow constraint programming can help, *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Springer, pp. 354–367.

Kaushik, K., Bhardwaj, A., Dwivedi, A. D. and Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums, *International Journal of Environmental Research and Public Health* **19**(13): 7898.

Klein, S., Allison, D. B., Heymsfield, S. B., Kelley, D. E., Leibel, R. L., Nonas, C. and Kahn, R. (2007). Waist circumference and cardiometabolic risk: a consensus statement from shaping america's health: Association for weight management and obesity prevention; naaso, the obesity society; the american society for nutrition; and the american diabetes association, *The American journal of clinical nutrition* **85**(5): 1197–1202.

Kowshalya, G. and Nandhini, M. (2018). Predicting fraudulent claims in automobile insurance, *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 1338–1343.

Morid, M. A., Kawamoto, K., Ault, T., Dorius, J. and Abdelrahman, S. (2017). Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation, *AMIA Annual Symposium Proceedings*, Vol. 2017, American Medical Informatics Association, p. 1312.

Mustika, W. F., Murfi, H. and Widyaningsih, Y. (2019). Analysis accuracy of xgboost model for multiclass classification - a case study of applicant level risk prediction for life insurance, *2019 5th International Conference on Science in Information Technology (ICSITech)*, pp. 71–77.

Niaksu, O. (2015). Crisp data mining methodology extension for medical domain, *Baltic Journal of Modern Computing* **3**(2): 92.

Panay, B., Baloian, N., Pino, J. A., Peñafiel, S., Sanson, H. and Bersano, N. (2019). Predicting health care costs using evidence regression, *Multidisciplinary Digital Publishing Institute Proceedings* **31**(1): 74.

Petit-Renaud, S. and Denœux, T. (2004). Nonparametric regression analysis of uncertain and imprecise data using belief functions, *International Journal of Approximate Reasoning* **35**(1): 1–28.

Sommers, B. D. (2020). Health insurance coverage: what comes after the aca? an examination of the major gaps in health insurance coverage and access to care that remain ten years after the affordable care act., *Health Affairs* **39**(3): 502–508.

Whatley, M. (2022). Ordinary least squares regression, *Introduction to Quantitative Analysis for International Educators*, Springer, pp. 91–112.