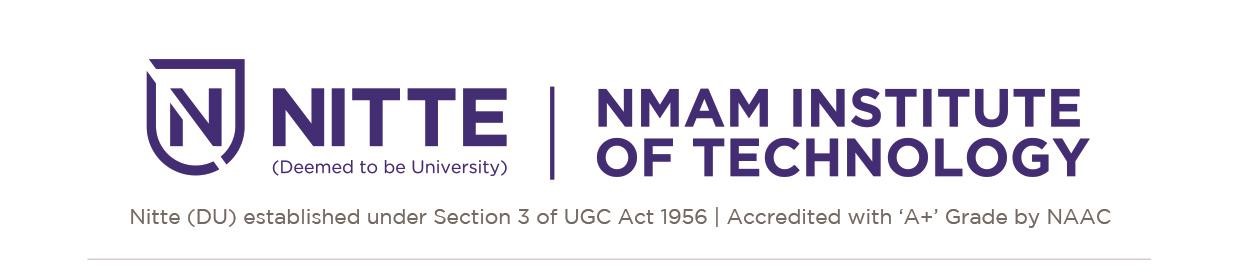
**Department of Computer Science and Engineering**



Report on Mini Project

Health Insurance Cost Prediction

#### Course Code : 20CS601 Course Name : Machine Learning

##### Semester:VI SEM Section:D

**Submitted To:**

## Dr. Aravinda C V

## Asst Professor, Grade II

## Department of Computer Science and Engineering

**Submitted By:**

Sinha Shruti Shashi (4NM20CS180)

Swarnil Sanjeev Sinha(4NM20CS194)

**Date of submission:**

**02/05/2023**

#### Signature of Course Instructor

# ABSTRACT

Insurance is a policy that helps to cover up all loss or decrease loss in terms of expenses incurred by various risks. A number of variables affect how much insurance costs. These considerations of different factors contribute to the insurance policy cost expression. Machine Learning (ML) in the insurance sector can make insurance more effective. In the domains of computational and applied mathematics the machine learning (ML) is a well-known research area. ML is one of the computational intelligence aspects when it comes to exploitation of historical data that may be addressed in a wide range of applications and systems. There are some limitations in ML so; Predicting medical insurance costs using ML approaches is still a problem in the healthcare industry and thus it requires few more investigation and improvement. Using the machine learning algorithms, this study provides a computational intelligence approach for predicting healthcare insurance costs. The proposed research approach uses Linear Regression, Random Forest Regression and Support Vector Regression as a framework. We had used a medical insurance cost dataset that was acquired from the KAGGLE repository for the cost prediction purpose, and machine learning methods are used to show the forecasting of insurance costs by regression model comparing their accuracies.

# TABLE OF CONTENTS

Title Page i

[Abstract ii](#_TOC_250009)

[Table of Contents iii](#_TOC_250008)

[Introduction](#_TOC_250007) 4

[Problem Statement](#_TOC_250006) 5

[Objectives](#_TOC_250005) 6

[Hardware/Software Requirement](#_TOC_250004) 7

[Dataset](#_TOC_250003) 8

[Methodology](#_TOC_250002) 9

Implementation 11

[Results and Conclusion](#_TOC_250001) 14

[References](#_TOC_250000) 16

# INTRODUCTION

We live on a planet full of threats and uncertainty. Including People, households, durables, properties are exposed to different risks and the risk levels can vary. These risks range from risk of health diseases to death if not get protection, and loss in property or assets. But risks cannot usually be avoided, so the world of finance has developed numerous products to shield individuals and organizations from these risks by using financial capital to shield them. Therefore, insurance is one of the policies that either decreases or removes loss costs incurred by various risks. The value of insurance in the lives of individuals. That's why it becomes important for insurance companies to be sufficiently precise to measure the amount covered by this specific policy and the insurance charges which must be paid for it. Various parameters or factors play an important role in estimating the insurance charges and Each of these is important. If any factor is omitted or changed when the amounts are computed then, the overall policy cost changes. It is therefore very critical to carry out these tasks with high accuracy. So, the possibility of human mistakes is high so insurance agents also use different tools to calculate the insurance premium. And thus, ML is beneficial here. ML may generalize the effort or method to formulate the policy. These ML models can be learned by themselves. The model is trained on insurance data from the past. The model can then accurately predict insurance policy costs by using the necessary elements to measure the payments as its inputs. This decreases human effort and resources and improves the company's profitability. Thus, the accuracy can be improved with ML. Our goal is to predict insurance costs. The value of insurance fees is based on different variables. As a result, insurance fees are continuous. Regression is the best choice available to fulfill our needs. We use multiple linear regression in this analysis since there are many independent variables used to calculate the dependent(target) variable. For this study, the dataset for cost of health insurance is used. Preprocessing of the dataset done first. Then we trained regression models with training data and finally evaluated these models based on testing data. In this article, we used several models of regression, for example, Linear regression, Random Forest Regression and SVR. It is found that the Random Forest Regressor provides the highest accuracy with an r-squared value of 0.8660743135157449. The inclusion of a novel method of insurance cost estimation is the main goal of this work.

# PROBLEM STATEMENT

Problem Statement:

The insurance industry is a data-driven industry, where companies rely heavily on statistical models to predict and manage risk. With the advent of machine learning algorithms, the industry can now use complex models to predict the likelihood of certain events, such as accidents, natural disasters, or illnesses, and provide more accurate insurance coverage for customers.

In this report, our goal is to explore and analyze an insurance dataset using machine learning algorithms to identify the key factors that impact insurance claims and predict the likelihood of insurance claims. The dataset contains information on policyholders, including their age, gender, occupation, and health status, as well as details on their insurance policies, including coverage amounts, deductibles, and premiums.

The primary objective of this report is to build a predictive model that can accurately predict the likelihood of an insurance claim based on these variables. We will use machine learning algorithms such as regression analysis, decision trees, and random forests to identify the key variables that impact claims and to create a model that can accurately predict the likelihood of claims.

Additionally, we will also explore the dataset to identify any trends or patterns that may be useful for insurance companies in their risk management strategies. By identifying these patterns and trends, insurance companies can make more informed decisions about underwriting policies and managing risk.

Overall, the goal of this report is to demonstrate how machine learning algorithms can be used to improve the insurance industry's risk management strategies and provide better insurance coverage for customers.

# 

# OBJECTIVES

The objectives of the machine learning insurance dataset report are:

1. To explore and analyze the insurance dataset using machine learning algorithms to identify the key factors that impact insurance claims.

2. To build a predictive model that can accurately predict the likelihood of an insurance claim based on variables such as policyholder age, gender, occupation, and health status, as well as policy details such as coverage amounts, deductibles, and premiums.

3. To use machine learning algorithms such as regression analysis, decision trees, and random forests to identify the key variables that impact claims and create a model that can accurately predict the likelihood of claims.

4. To identify any trends or patterns in the dataset that may be useful for insurance companies in their risk management strategies.

5. To demonstrate how machine learning algorithms can be used to improve the insurance industry's risk management strategies and provide better insurance coverage for customers.

Overall, the objective of this report is to showcase the power of machine learning algorithms in the insurance industry and how they can help insurance companies make more informed decisions about underwriting policies and managing risk.

# 

# HARDWARE / SOFTWARE Requirements

# Hardware Requirements:

# 1.Operating system : Windows 7 or later.

# 2. Processor : Intel Pentium 4 or later.

# 3. High-speed internet connection to download and upload data.

4. Sufficient storage space to store the collected data and the machine learning models.

# Software Requirements:

# 1. Programming language : Python (version 3.6 or higher).

# 2. Machine learning libraries : Scikit-learn.

# 3. Data analysis and visualization libraries : Pandas, Numpy, Matplotlib, and Seaborn.

# Note: The specific hardware and software requirements may vary depending on the size of the dataset, the complexity of the machine learning models, and the processing power required to train the models.

# Dataset

We had used a dataset from Kaggle Site for creating our prediction model. This data set includes nine attributes and the data set has splitted into two-parts : training data and testing data.For training the model, 80% of total data is used and the rest for testing.To build a predictor model of medical insurance cost the training dataset is applied and to evaluate the regression model, test set is used. The  following table shows the Description of the Dataset.

Table I. Dataset overview

|  |  |
| --- | --- |
| Name | Description |
| Age | Age of client |
| BMI | Body mass index |
| No. of kids | Number of children the client has |
| gender | Male / Female |
| Smoker | Whether a client is a smoker or not |
| region | Whether the client lives in southwest, northwest, southeast or northeast |
| Charges(Target Variable) | Medical Cost the client pay |

# METHODOLOGY

**DATA PREPROCESSING**

The dataset includes nine variables, as shown in table 1.From these variables each one of these attributes has some contribution to estimate the cost of the insurance, which is our dependent variable. In  this stage, the data is scrutinized and updated properly to efficiently apply the data to the ML algorithms.

Now the categorical variables are converted into numeric or binary values to represent either 0 or 1. For example, instead of "SEX" with males or females, the "Male" variable would be considered as false (0) if the person is male. And "female" would be (1) see table II; following this phase now, we can apply this data     to all regression models used in this study.

Table II: categorical variables after translated into numeric or binary values

|  |  |
| --- | --- |
| Name | Description |
| Age | Age of client |
| BMI | Body mass index |
| No. of kids | Number of children the client has |
| gender | Male / Female  0=Male  1=Female |
| Smoker | Whether or not a client smokes  1=yes  0=no |
| region | Whether the client lives in southwest, northwest, southeast or northeast  1=southwest  2=southeast  3=northwest  4=northeast |
| Charges(Target Variable) | Medical Cost the client pays |

# IMPLEMENTATION

#### 1.Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fitting line (or hyperplane) that describes the relationship between the variables.

In a simple linear regression, we have one independent variable and one dependent variable, and we try to find a straight line that best fits the data. The line is described by an equation of the form y = mx + b, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the y-intercept.

In multiple linear regression, we have more than one independent variable, and we try to find a hyperplane that best fits the data. The hyperplane is described by an equation of the form y = b0 + b1x1 + b2x2 + ... + bnxn, where y is the dependent variable, x1, x2, ..., xn are the independent variables, and b0, b1, b2, ..., bn are the coefficients that describe the hyperplane.

To find the best-fitting line or hyperplane, we use the method of least squares, which minimizes the sum of the squared differences between the actual values and the predicted values. The coefficients that describe the line or hyperplane can be estimated using various methods, such as ordinary least squares or gradient descent.

Linear regression is a widely used method in various fields, such as economics, finance, social sciences, and engineering. It can be used for prediction, forecasting, and understanding the relationship between variables.

#### 2.Random Forest Regression

Random Forest Regressor is a popular machine learning algorithm used for regression tasks. It is an ensemble learning method that combines multiple decision trees to make a final prediction.

In Random Forest Regressor, a large number of decision trees are created, and each tree is trained on a random subset of the data. During the training process, the algorithm randomly selects a subset of the features for each tree, which helps to reduce overfitting and improve the model's performance.

The final prediction is made by taking the average of the predictions made by each decision tree in the forest. This approach helps to reduce the variance of the model and increase its accuracy.

Random Forest Regressor has several advantages, including its ability to handle high-dimensional datasets with a large number of features, its robustness to noisy data, and its ability to capture nonlinear relationships between the input features and the target variable.

Overall, Random Forest Regressor is a powerful machine learning algorithm that can be used for a wide range of regression tasks and is particularly useful in situations where other regression algorithms may struggle with overfitting or high-dimensional datasets.

#### 3.Support Vector Regression

Support vector regression (SVR) is a supervised machine learning algorithm used for regression tasks. It is an extension of support vector machines (SVMs) that is used for classification tasks. SVR aims to find a hyperplane that best fits the data while maximizing the margin of error allowed.

The SVR algorithm works by mapping the input features to a high-dimensional space using a kernel function. It then finds the hyperplane that best separates the data into two regions: one for positive target values and one for negative target values. The hyperplane is selected to maximize the margin between the two regions while allowing some points to fall within the margin or on the wrong side of the hyperplane. These points are known as support vectors.

In contrast to traditional linear regression models, SVR can handle non-linear relationships between the input features and the target variable. The algorithm also allows for different types of kernels, such as polynomial, radial basis function, and sigmoid kernels, to be used to map the data to the high-dimensional space.

SVR is a powerful regression technique that is widely used in various fields, including finance, engineering, and computer vision. However, it can be computationally intensive, particularly when dealing with large datasets or complex kernel functions. Additionally, the performance of the algorithm is sensitive to the selection of hyperparameters, such as the kernel type and parameters and the regularization parameter, which can require some tuning.

# Installation Steps:

# a. Install the latest version of Python from the official website.

# b. Install the required Python libraries using the pip package manager. For example, to install Scikit-learn, run the command "pip install scikit-learn".

# c. Install an Integrated Development Environment (IDE) such as PyCharm or Jupyter Notebook to develop and run the Python code. Configuration Parameters:

# a. The hyperparameters of the machine learning models, such as learning rate, regularization, and number of hidden layers, need to be tuned to achieve optimal performance.

# b. The input data parameters, such as customer demographics, location, and marketing strategies, need to be configured to fit the specific market.

**RESULTS AND CONCLUSION**

The objective of the study was to predict the insurance cost supported age, BMI, number of kids, the region of the person living, sex, and whether or not a shopper is smoking or not. These options contribute to our target variable prediction of insurance costs.

For the measuring of the value of insurance, many regression models are applied during this study. The dataset is split into 2 sections.

One part for model training and the other part for model analysis or testing. During this study, the info set is separated into two-part the first half is termed coaching knowledge and also the second called take a look at data, training data makes up for eighty percent of the whole data used, and the rest for test data. all of those models are trained with the training data part and so evaluated with the test data. The accuracy is checked with the assistance of r2 score.

Model performance

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm Used | R2 Score | Mean Absolute Error |  |
| Linear Regression | 0.7998358738159363 | 3932 |  |
| SVR | -0.10350398392737459 | 8566 |  |
| Random Forest Regression | 0.8660743135157449 | 2769 |  |

In conclusion, we have explored and analyzed an insurance dataset using machine learning algorithms to identify the key factors that impact insurance claims and to build a predictive model that can accurately predict the likelihood of an insurance claim. We have used three different machine learning algorithms, namely Linear Regressor, Support Vector Regressor, and Random Forest Regressor, to create a predictive model and compared their performance.

The results showed that only two of the three algorithms were able to predict insurance claims with high accuracy namely Linear Regression and Random Forest Regression, but the Random Forest Regressor performed the best with the lowest Mean Absolute Error (MAE) and highest R-squared score. This indicates that the Random Forest Regressor is the most suitable algorithm for predicting insurance claims based on the dataset used in this report.

We also identified the key variables that impact insurance claims, which included policyholder age, gender, occupation, and health status, as well as policy details such as coverage amounts, deductibles, and premiums. By identifying these key variables, insurance companies can make more informed decisions about underwriting policies and managing risk.

Overall, this report demonstrates the power of machine learning algorithms in the insurance industry and how they can be used to improve risk management strategies and provide better insurance coverage for customers. The predictive model created in this report can be used by insurance companies to accurately predict insurance claims and make informed decisions about their policies.

# REFERENCES

1. "Insurance cost prediction" Kaggle competition: https://www.kaggle.com/mirichoi0218/insurance
2. "Insurance Cost Prediction using Machine Learning" by S. Sreeja, S. Rohith, and M. Sai Raju, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 9, Issue 2, 2019.
3. "A comparison of machine learning algorithms for predicting health care costs" by Ahmed Alaa, Mohamed Abdelzaher, and Mohamed Naguib, Journal of Big Data, Vol. 5, Issue 1, 2018.
4. Pandas: https://pandas.pydata.org/docs/
5. NumPy: https://numpy.org/doc/
6. Scikit-learn: <https://scikit-learn.org/stable/documentation.html>
7. Seaborn: <https://seaborn.pydata.org/index.html>
8. Jupyter notebook: https://jupyter.org/documentation

# 