
# MAJOR PROJECT-2

## Final Report

For

## Medical Insurance Cost Prediction

Submitted By


| Specialization | SAP ID | Name |
|---|---|---|
| B.Tech CSE – AIML (NH) | 500083955 | Milind Sharma |
| B.Tech CSE – AIML (NH) | 500082497 | Rahul Mehta |
| B.Tech CSE – AIML (NH) | 500082578 | Rahul B Nair |
| B.Tech CSE – AIML (NH) | 500082920 | Priyank Singh |


Department of Informatics

School of Computer Science

UNIVERSITY OF PETROLEUM & ENERGY STUDIES,

DEHRADUN- 248007. Uttarakhand

Dr. Priyanka Singh
**Project Guide**

Dr. Anil Kumar
**Cluster Head**

**Final Report**

# **Project Title:** Medical Insurance Cost Prediction

**Content**

# 1. Abstract

This project focuses on the development of a predictive model aimed at estimating the costs associated with individual medical insurance policies. By leveraging a comprehensive dataset that encapsulates variables such as age, sex, Body Mass Index (BMI), the presence of children, smoking status, and geographical region. [3]

We utilize linear regression techniques to project the potential charges one might incur. The initiative seeks to distill these variables into a functional, predictive tool that benefits both insurance providers in setting premiums and individuals in financial planning regarding healthcare expenses. [1]

# 2. Introduction

The challenge of accurately predicting medical insurance costs lies at the heart of both personal financial planning and the insurance industry's pricing strategies. Given the diverse factors that can influence these costs, there is a significant need for predictive models that can offer accurate forecasts based on a wide range of personal attributes. [3]

This study employs machine learning methodologies, with a particular focus on linear regression, to construct a model capable of predicting insurance charges with a high degree of accuracy. The model's simplicity and the interpretability of its results make it an invaluable tool in the ongoing effort to understand and manage healthcare expenditures. [1]

# 3. Literature Review

Extensive research in the field of insurance cost prediction has shown the application of various predictive models, ranging from simple linear regression to more complex algorithms like decision trees, random forests, and neural networks. While complex models may capture nonlinear relationships more effectively, linear regression remains a popular choice due to its straightforward implementation and ease of interpretation. [2]

This project contributes to the body of knowledge by harnessing linear regression, supported by a comprehensive analysis of its effectiveness and limitations in the context of insurance cost prediction, as delineated in existing scholarly works. [3]

## 4. Problem Statement

The project addresses the intricate challenge of forecasting individual medical insurance costs by analyzing a set of personal and demographic factors. The inherent complexity of the insurance cost structure, influenced by a myriad of interrelated variables, necessitates the development of a sophisticated yet accessible predictive model. The goal is to demystify the cost structure and provide accurate estimations that can aid in financial planning and policy formulation. [3]

## 5. Objectives:

The key objectives of the project are as follows: -

1. **Comprehensive Data Exploration**: Initiate a thorough exploration of the dataset to gain insights into the distribution and characteristics of various variables, and identify potential correlations between them. [1]

2. **Rigorous Data Preprocessing**: Undertake meticulous preprocessing steps to ensure the dataset is primed for modeling. This includes addressing missing values, encoding categorical variables for computational processing, and normalizing data to enhance model performance. [1]

3. **Predictive Model Development**: Construct and refine a linear regression model tailored to accurately predict medical insurance costs based on a set of defined variables, ensuring the model captures the underlying patterns and relationships in the data. [1]

4. **Detailed Model Evaluation**: Utilize the R-squared metric, among others, to thoroughly evaluate the model's predictive accuracy on both the training and testing datasets. This step is critical to ascertain the model's reliability and effectiveness in real-world applications. [1]

5. **Practical Application and Prediction**: Demonstrate the model's utility by applying it to predict insurance costs for new, unseen data. This not only tests the model's applicability but also showcases its potential as a practical tool for predicting insurance costs. [2]

## 6. **Methodology:**

The project's methodology is structured into sequential phases, each critical to the development of an accurate and reliable predictive model: -

- **Data Loading and Exploration**: Initial steps involve loading the dataset into a Pandas DataFrame, followed by an exploratory analysis aimed at understanding the dataset's structure, examining the distribution of key variables, and identifying missing values. This phase is crucial for gaining insights and informing subsequent preprocessing steps. [2]

- **Data Preprocessing**: This phase focuses on preparing the dataset for modeling. Categorical variables such as sex, smoker status, & region are transformed into numerical formats to facilitate computational analysis. The dataset is then divided into feature sets (X) & the target variable (Y), followed by a split into training & testing sets for model evaluation. [1]

- **Model Building**: A Linear Regression model is developed using the training set. This model is designed to learn the relationships between the input features and the target variable, aiming to accurately predict insurance costs based on these inputs. [3]

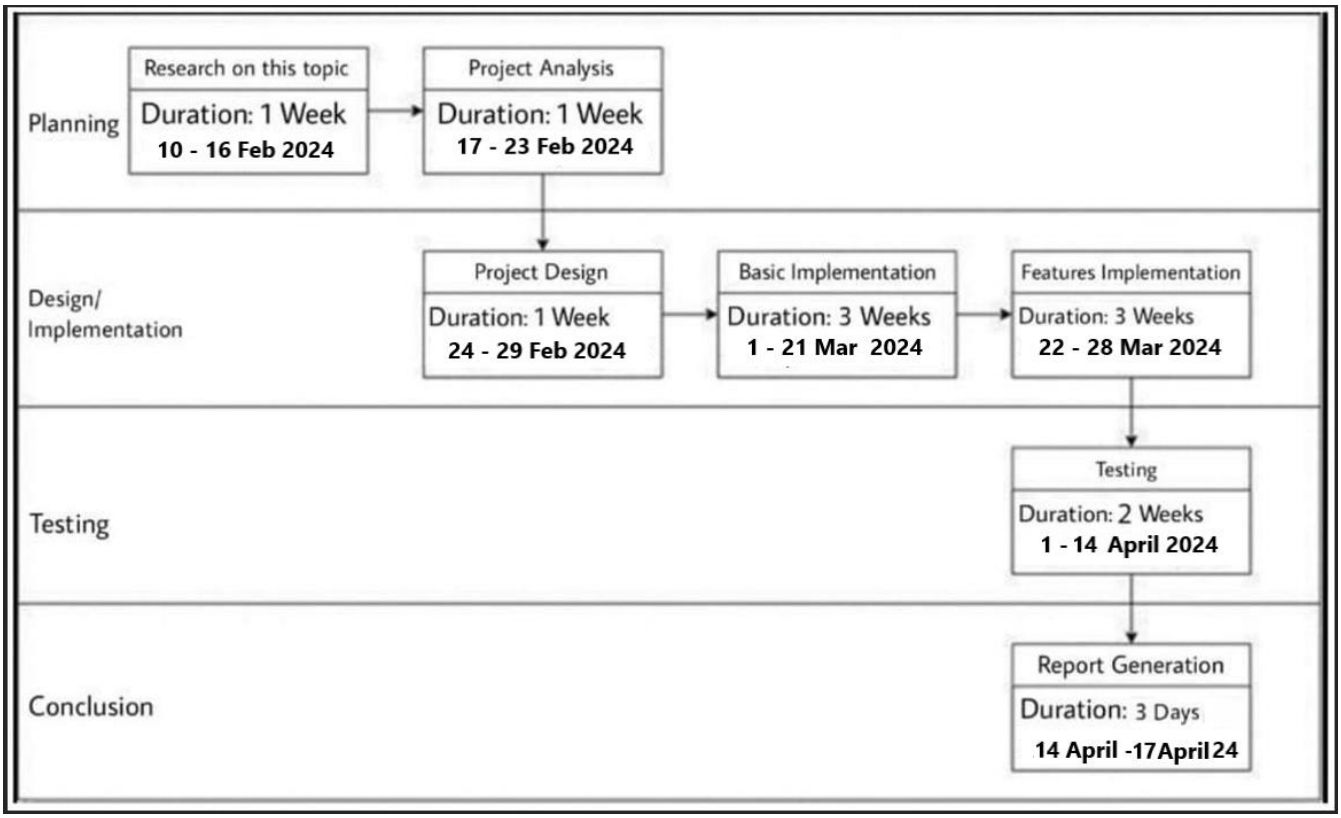
- **Model Evaluation**: The model's performance is critically assessed using the R-squared metric on both the training and testing datasets. This evaluation helps determine the model's accuracy and its ability to generalize to new data. [4]

- **Prediction**: Finally, the model is employed to make predictions on new data, illustrating its practical use case. This step serves as a proof of concept for the model's application in real-world scenarios. [4]

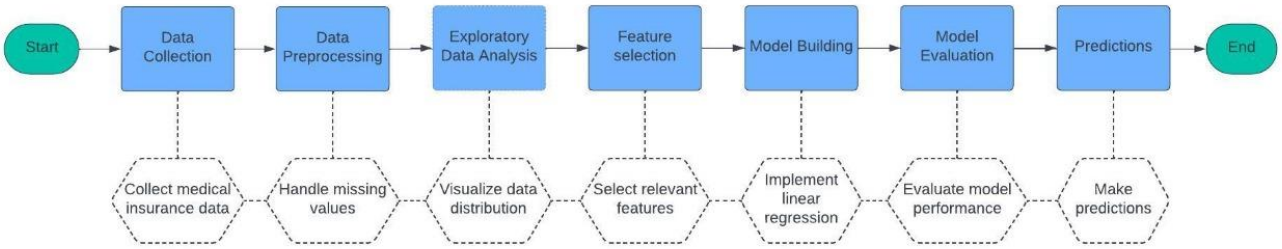
## 7. **References**


1. https://www.scalablepath.com/data-science/data-preprocessing-phase [Data Preprocessing Techniques]
2. https://www.revealbi.io/blog/predictive-analytics-in-healthcare [Predictive Analytics in Healthcare]
3. https://www.coursera.org/in/articles/machine-learning-in-health-care [What Is Machine Learning in Healthcare?]
4. https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/ [R-squared in Regression Analysis in Machine Learning]

# 8. PERT Chart




# 9. Technical Diagram



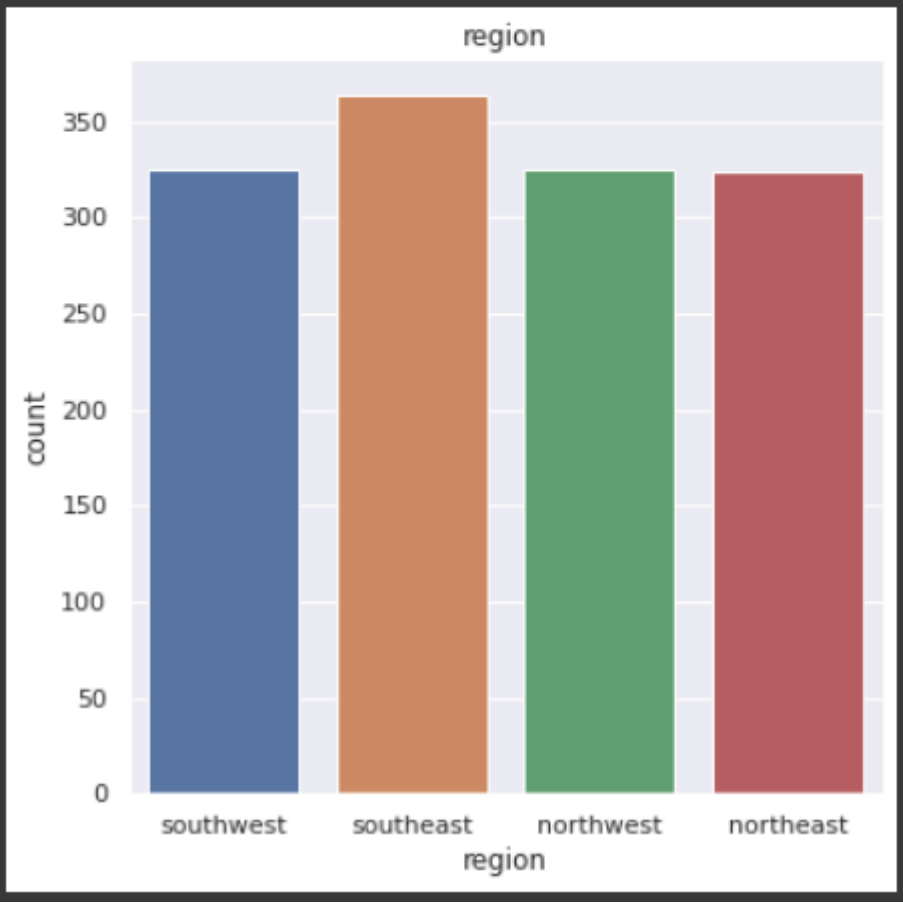
# 10. Result

Data Analysis

```
[ ] # statistical Measures of the dataset
    insurance_dataset.describe()
```

| | age | bmi | children | charges |
|---|---|---|---|---|
| count | 1338.000000 | 1338.000000 | 1338.000000 | 1338.000000 |
| mean | 39.207025 | 30.663397 | 1.094918 | 13270.422265 |
| std | 14.049960 | 6.098187 | 1.205493 | 12110.011237 |
| min | 18.000000 | 15.960000 | 0.000000 | 1121.873900 |
| 25% | 27.000000 | 26.296250 | 0.000000 | 4740.287150 |
| 50% | 39.000000 | 30.400000 | 1.000000 | 9382.033000 |
| 75% | 51.000000 | 34.693750 | 2.000000 | 16639.912515 |
| max | 64.000000 | 53.130000 | 5.000000 | 63770.428010 |

```
[ ] # region column
    plt.figure(figsize=(6,6))
    sns.countplot(x='region', data=insurance_dataset)
    plt.title('region')
    plt.show()
```



## Data Pre-Processing

### Encoding the categorical features

```
[ ]  # encoding sex column
     insurance_dataset.replace({'sex':{'male':0,'female':1}}, inplace=True)

     3 # encoding 'smoker' column
     insurance_dataset.replace({'smoker':{'yes':0,'no':1}}, inplace=True)

     # encoding 'region' column
     insurance_dataset.replace({'region':{'southeast':0,'southwest':1,'northeast':2,'northwest':3}}, inplace=True)
```

## Splitting the data into Training data & Testing Data

```
[ ]  X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
```

```
[ ]  print(X.shape, X_train.shape, X_test.shape)

     (1338, 6) (1070, 6) (268, 6)
```

## Model Training

### Linear Regression

```
[ ]  # loading the Linear Regression model
     regressor = LinearRegression()
```

```
[ ]  regressor.fit(X_train, Y_train)

     LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

## Model Evaluation

```
[ ]  # prediction on training data
     training_data_prediction =regressor.predict(X_train)
```

```
[ ]  # R squared value
     r2_train = metrics.r2_score(Y_train, training_data_prediction)
     print('R squared vale : ', r2_train)
```

```
     R squared vale :  0.751505643411174
```

```
[ ]  # prediction on test data
     test_data_prediction =regressor.predict(X_test)
```

```
[ ]  # R squared value
     r2_test = metrics.r2_score(Y_test, test_data_prediction)
     print('R squared vale : ', r2_test)
```

```
     R squared vale :  0.7447273869684077
```

## Building a Predictive System

```
[ ]  input_data = (31,1,25.74,0,1,0)

     # changing input_data to a numpy array
     input_data_as_numpy_array = np.asarray(input_data)

     # reshape the array
     input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

     prediction = regressor.predict(input_data_reshaped)
     print(prediction)

     print('The insurance cost is USD ', prediction[0])
```

```
     [3760.0805765]
     The insurance cost is USD  3760.0805764960587
```

# 11. Acknowledgement:

Presentation inspiration and motivation have always played a key role in the success of any venture.

We express are sincere thanks to Dr. Priyanka Singh, University of petroleum and energy studies, Dehradun.

We pay our deep sense of gratitude to Dr. Priyanka Singh, UPES to encourage us to the highest peak and to provide me the opportunity to prepare the project. We are obliged to our group members for their elevating inspiration, encouraging guidance and kind supervision in completing my project.

We feel to acknowledge our indebtedness and deep sense of gratitude to course coordinator Mr. Teekam Singh whose valuable guidance and kind supervision given to me throughout the course which shaped the present work as its show.

Last, but not least, our parents are also an important inspiration for us. So, with due regards, we express my gratitude to them.