Medical Insurance Cost Prediction Using Machine Learning

[1]G. MANOJ KUMAR, [2]PEKETI LAHARI

[1]Assistant Professor, Department of MCA, [2]2MCA Final Semester

Master of Computer Applications,

[1]Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India

ABSTRACT:
The increasing cost of healthcare services has made medical insurance a crucial financial tool. Predicting insurance costs accurately helps insurance providers assess risk and allows customers to plan better. This project focuses on building a machine learning model to predict individual medical insurance charges using key features such as age, sex, BMI, number of children, smoking status, and region. The model uses linear regression as a base method to map the relationship between these variables and insurance cost. The data is pre-processed and visualized using exploratory data analysis (EDA), followed by model training and evaluation using metrics like R^2 score. This project demonstrates the application of data science in real-world financial estimation, providing an efficient and scalable solution for insurance cost prediction.

Index Terms: Medical Insurance, Machine Learning, Linear Regression, Cost Prediction, EDA, Streamlit, Health Analytics, Insurance Premium Estimation.

I. INTRODUCTION

Medical insurance is a crucial aspect of modern healthcare, helping individuals manage unexpected medical expenses. As healthcare costs continue to rise, accurately estimating insurance charges becomes essential for both customers and insurance providers. Traditional methods often rely on static tables and manual calculations, which can be inefficient and inaccurate.

Machine learning offers a powerful solution by analyzing past data to predict future costs. In this project, we use a dataset containing demographic and health information such as age, sex, BMI, number of children, smoking status, and region. A Linear Regression model is trained on this data to predict the cost of insurance for new individuals based on these input features.

The final model is deployed through a simple web application built with Streamlit, allowing users to input data and receive instant cost predictions. This system provides transparency, speed, and accuracy—making it a useful tool for both policyholders and insurers.

1.1 Existing System

In the current scenario, insurance companies often rely on traditional and manual methods for estimating medical insurance costs. These methods typically use predefined tables or statistical averages based on demographic groups, which do not account for the unique combinations of individual health and lifestyle factors. As a result, the cost predictions may not be personalized or accurate.

Moreover, existing systems usually involve human intervention, which can be time-consuming and prone to errors. They do not leverage data science or automation to handle large datasets or identify hidden patterns. This can lead to delays in policy approvals and suboptimal premium pricing.

These limitations highlight the need for an intelligent system that can automatically analyze data, learn from it, and provide real-time, personalized insurance cost predictions based on individual input parameters.

1.1.1 Challenges

Developing an accurate medical insurance cost prediction model involves several key challenges:

- **Data Quality and Preprocessing**
Real-world data often contains missing, inconsistent, or non-numeric values that must be cleaned and transformed before training the model. Proper encoding of categorical features like gender, region, and smoking status is essential for the algorithm to learn effectively.
- **Feature Selection and Correlation**
Not all input features contribute equally to insurance costs. Identifying which variables have the most influence—such as age, BMI, or smoking habits—requires careful exploratory data analysis (EDA) and statistical understanding.
- **Model Generalization and Accuracy**
Ensuring that the model performs well not only on training data but also on unseen test data is a common challenge. Overfitting must be avoided while still achieving a high R² score and reliable predictions across various user inputs.
- **User Input Validation in Real-Time Interface**
Designing a web app that accepts inputs from users and converts them into a format the model can use—while handling invalid entries gracefully—is crucial for smooth user interaction.
- **Interpretability and Transparency**
In real-world applications like insurance, it is important for models to be interpretable. Users and companies need to understand *why* a certain cost was predicted, which adds pressure to use simpler models like linear regression over complex black-box algorithms.

1.2 Proposed System

The proposed system uses a machine learning model—specifically Linear Regression—to predict medical insurance costs based on user inputs such as age, gender, BMI, smoking status, number of children, and region. The system processes and encodes the data, trains the model on a historical dataset, and evaluates its performance using the R² score. It is deployed as a web application using Streamlit, allowing users to input their details and receive real-time, accurate cost predictions. This approach provides a faster, more reliable, and personalized alternative to traditional manual estimation methods.

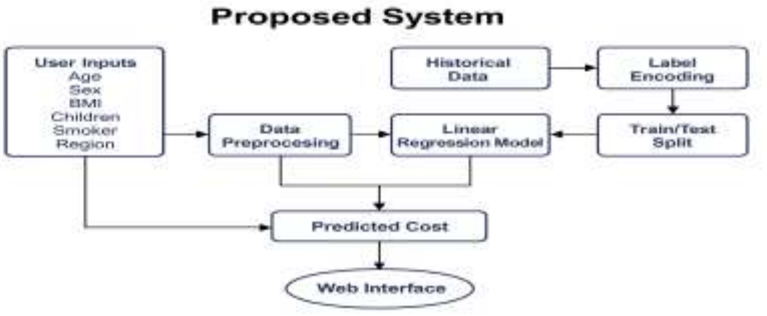


Fig 1: Proposed System

1.2.1 Advantages

- **Accurate Predictions**: Uses machine learning to provide precise and data-driven insurance cost estimates.
- **Real-Time Results:** Predicts insurance charges instantly based on user input through a web interface.
- **User-Friendly Interface:** Built using Streamlit, the system is simple and intuitive for end users.
- **Personalized Output:** Considers individual factors like age, BMI, and smoking status for tailored predictions.
- **Automation:** Eliminates manual calculations, reducing time and human error.
- **Cost-Effective:** Developed using open-source tools, requiring no commercial software licenses.

- **Scalable:** Can be extended to handle larger datasets or include more complex models in the future.
- **Interpretable Model:** Linear Regression offers clear insights into how each feature affects the prediction.

II. LITERATURE REVIEW

2.1 Architecture

The architecture of the medical insurance cost prediction system is designed to ensure smooth data flow, accurate predictions, and real-time interaction. It starts with a user-friendly Streamlit web interface where users input personal data such as age, sex, BMI, smoking status, number of children, and region. This input is sent to the data processing module, which formats and encodes the information to match the model’s training data.

The cleaned data is then passed to the trained Linear Regression model, which performs the prediction based on learned relationships from the dataset. The output—the predicted insurance cost—is then displayed back to the user. In the background, an offline training module handles the initial preparation of the model using the insurance.csv dataset, ensuring the system is ready for real-time inference.

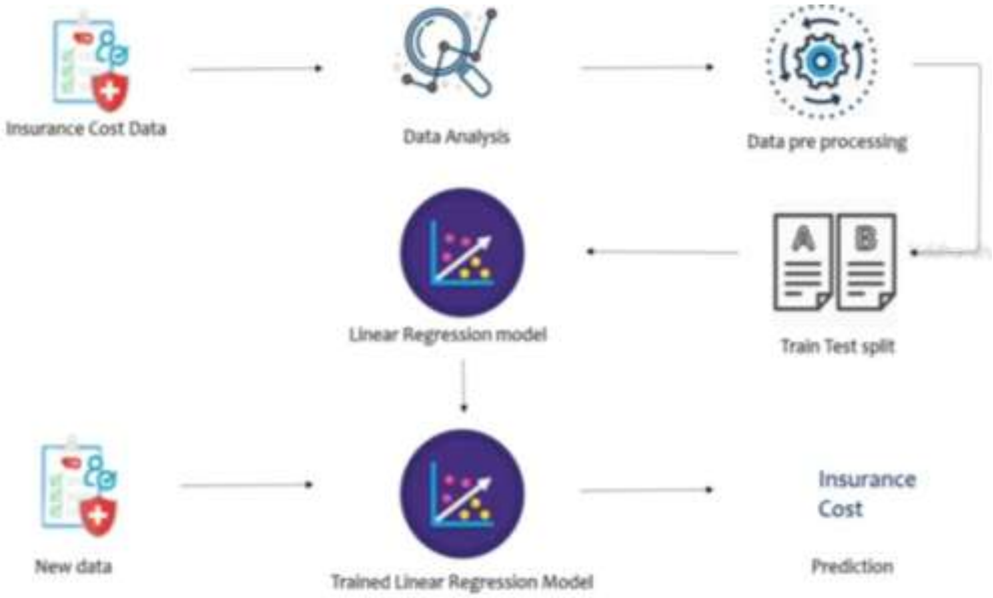


Fig: 2 Architecture

2.2 Algorithm

The algorithm used in this project is Linear Regression, a supervised learning technique commonly used for predicting continuous numeric values. In the context of medical insurance, it estimates the cost of insurance based on input features such as age, BMI, number of children, smoking status, sex, and region. Linear Regression models the relationship between the dependent variable (charges) and one or more independent variables by fitting a linear equation to the data. It is simple, interpretable, and efficient, making it ideal for a web-based prediction system.

The model is trained on historical insurance data using the least squares method, which minimizes the sum of squared differences between the actual and predicted values. After training, it is capable of generalizing to new data inputs, enabling it to provide quick and fairly accurate predictions. Evaluation is done using metrics like R^2 score to assess how well the model fits the data.

Algorithm Used: Linear Regression (Supervised Learning – Regression)
Goal: Predict the insurance cost (charges) based on input features
Input Features:

- Age

- Sex (encoded as 0/1)
- BMI
- Number of Children
- Smoker (0 = Yes, 1 = No)
- Region (encoded from 0 to 3)

Output: Predicted insurance cost (charges)
Training Technique: Ordinary Least Squares (OLS)
Evaluation Metric: R² score (Coefficient of Determination)
Advantages:

- Simple and interpretable
- Efficient with small-to-medium datasets
- Good baseline for regression tasks

2.3 Techniques

- **Data Collection**
  - Collected from the publicly available insurance.csv dataset containing demographic and medical information.
- **Data Preprocessing**
  - Handled categorical features using Label Encoding (sex, smoker, region).
  - Converted all inputs to numerical format to be compatible with machine learning models.
- **Feature Selection**
  - Chose relevant features such as age, BMI, number of children, smoking status, gender, and region**.**
- **Exploratory Data Analysis (EDA)**
  - Used Seaborn and Matplotlib for data visualization.
  - Created correlation heatmaps, distribution plots, and pairplots to understand relationships in data.
- **Train-Test Split**
  - Split data into 90% training and 10% testing using train_test_split() from Scikit-learn.
- **Model Training**
  - Applied Linear Regression using LinearRegression() from Scikit-learn.
  - Trained the model on the processed data.
- **Model Evaluation**
  - Evaluated performance using R² score, which measures prediction accuracy.
- **Web App Development**
  - Built a Streamlit interface for real-time user input and cost prediction.
  - Deployed the trained model in an interactive form.

2.4 Tools

Programming Language

- **Python –** Used for data processing, machine learning, and backend development.

Libraries and Frameworks

- Pandas – For loading and manipulating datasets.
- NumPy – For numerical computations and array handling.
- Matplotlib – For plotting basic charts and graphs.
- Seaborn – For advanced data visualization (e.g., correlation heatmaps).
- Scikit-learn – For machine learning (Linear Regression, train-test split, R² score).
- Streamlit – For creating the interactive web application to deploy the model.

Development Environment

- Jupyter Notebook – Used for writing, testing, and visualizing code.
- VS Code / Text Editor – For editing Python and Streamlit files (app.py).

Dataset

- insurance.csv – Public dataset containing features like age, sex, BMI, smoking status, children, region, and charges.

System Requirements

- Python 3.x
- Internet Browser (to run the Streamlit app)
- OS: Windows/Linux/Mac

2.5 Methods

2.5.1. Data Loading

- Used pandas.read_csv() to load the insurance.csv dataset.
- Inspected the dataset using .head(), .info(), and .describe() methods.

2.5.2. Data Preprocessing

- Handled categorical features using **Label Encoding**:
  - sex: male = 0, female = 1
  - smoker: yes = 0, no = 1
  - region: southeast = 0, southwest = 1, northwest = 2, northeast = 3
- Checked for missing values and ensured data consistency.

2.5.3. Exploratory Data Analysis (EDA)

- Used **Seaborn** and **Matplotlib** to generate:
  - Correlation Heatmap
  - Distribution Plots
  - Pair Plots
- Identified relationships between features and insurance charges.

2.5.4. Feature Selection

- Selected relevant features (age, sex, bmi, children, smoker, region) as predictors.
- Target variable: charges

2.5.5. Splitting the Dataset

- Used train_test_split() from Scikit-learn to divide data:
  - 90% for training
  - 10% for testing

2.5.6. Model Training

- Applied **Linear Regression** using LinearRegression() from Scikit-learn.
- Trained the model using .fit() method with training data.

2.5.7. Prediction and Evaluation

- Made predictions using .predict() method.
- Evaluated model using **R^2 Score** to check prediction accuracy.

2.5.8. Web Application Deployment

- Used **Streamlit** to create a web interface.
- Took user inputs, passed them to the model, and displayed predicted cost in real time.

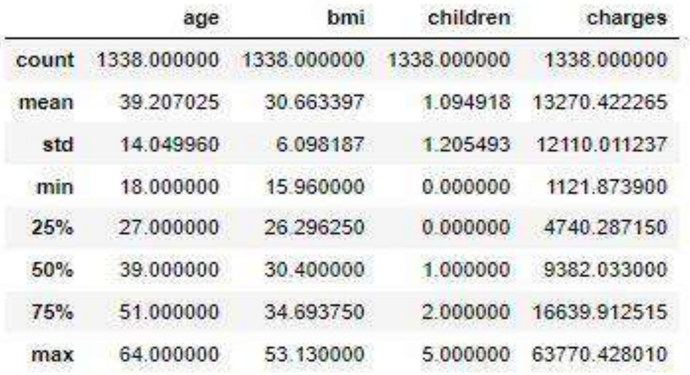
III. METHODOLOGY

3.1 Data Description

The insurance.csv dataset forms the foundation of this medical insurance cost prediction system. It consists of 1338 rows and 7 columns, where each row represents an individual, and each column corresponds to an attribute influencing medical expenses. The features include age (in years), sex (male/female), bmi (Body Mass Index), children (number of dependents), smoker (yes/no), and region (one of four U.S. regions). The dependent variable, charges, represents the insurance premium charged to the individual. The dataset is free from missing values and provides a balanced mix of numerical and categorical data, making it highly suitable for training regression models like Linear Regression. Its comprehensiveness and cleanliness allow for effective preprocessing, visualization, model training, and deployment.

Attribute	Data Description
Age	The age of individual person
Sex	Sex of the person (Male, Female)
BMI	This is Body Mass Index
Children	Total number of children of the person have
Smoker	Whether the person is a smoker or not
Region	Where the person lives. Considering four regions (Southwest, Southeast, Northeast, Northwest)

Table 1: overview of the dataset



	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

Table 2: Statistical measurement

3.2 Data Analysis

There were 1338 rows and 7 columns in our data set. The charges variable, which has a float value, is our aim. Maximum number of individuals in our dataset range in age from 18 to 22.5, and the majority of them are male. Few have more than three children, and the majority of them have a BMI between 29.26 and 31.16. In this dataset, four main regions are taken into account: northeast, northwest, southeast, and southwest. The largest concentration of smokers is in the southeast, where 1064 out of 1338 people smoke. Here are some data visualizations.

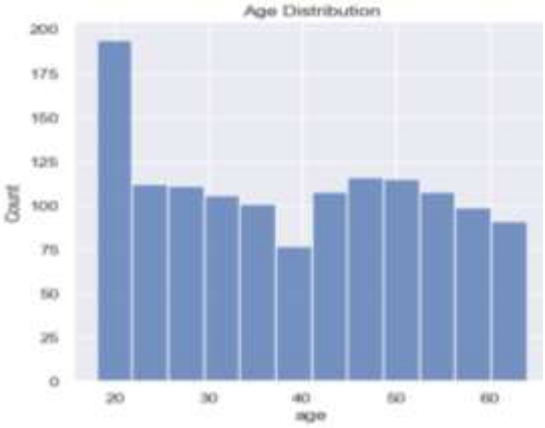


Fig 3: Distribution of age value

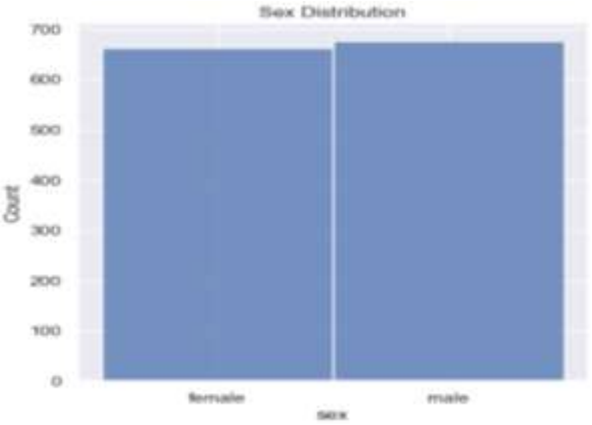


Fig 4: Sex Distribution

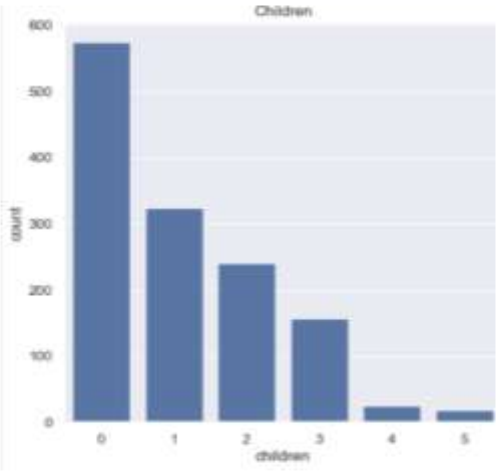


Fig 6: Children counter

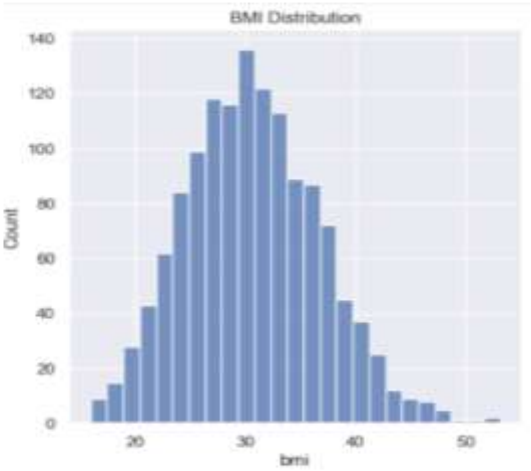


Fig 5: BMI Distribution

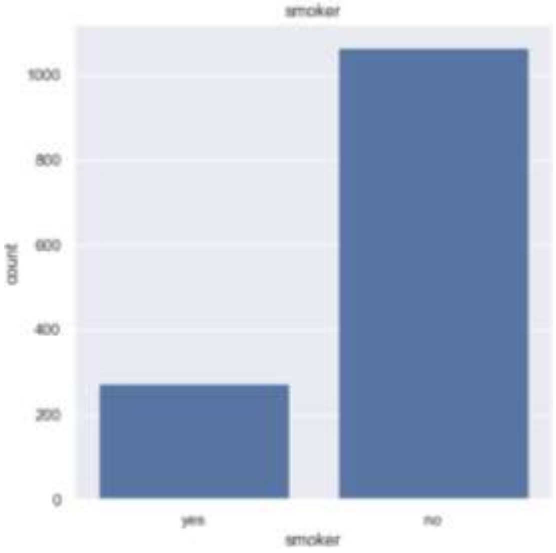


Fig 6: Checking smoker and non-smoker

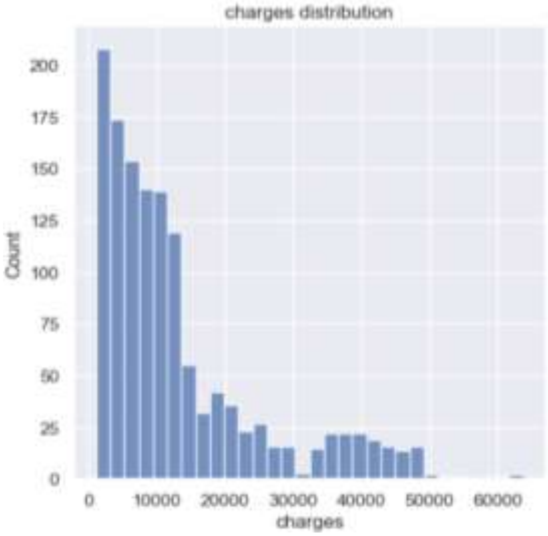


Fig 7: Distribution of charges value

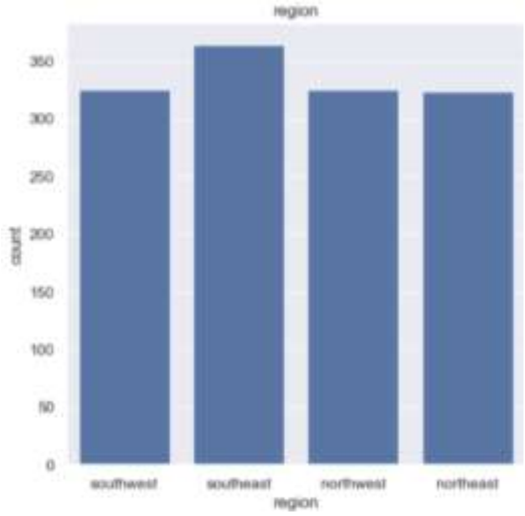
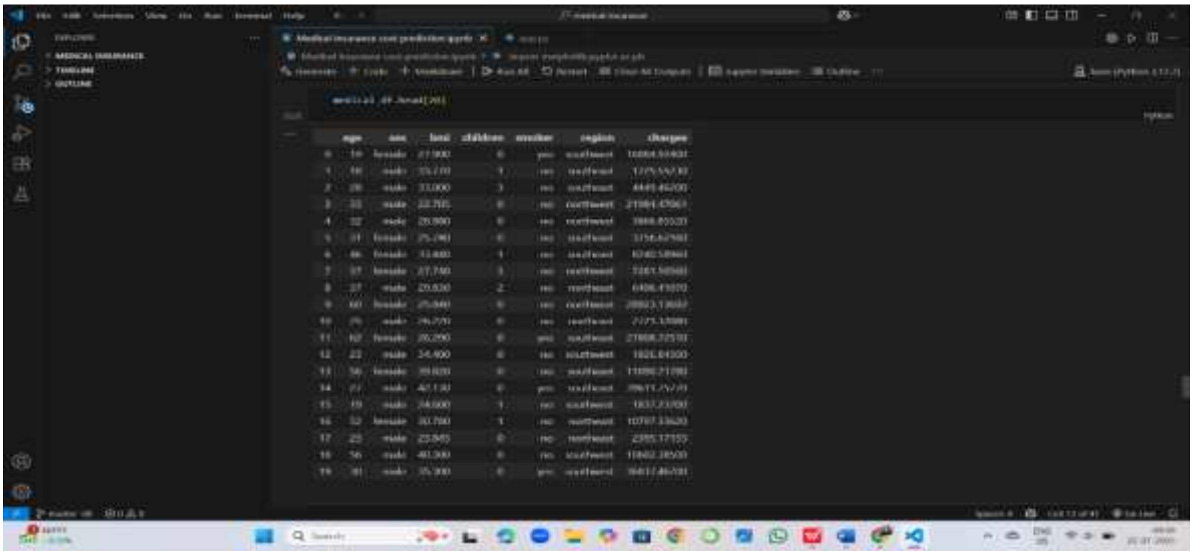


Fig 8: Region

3.3 Data Pre-processing

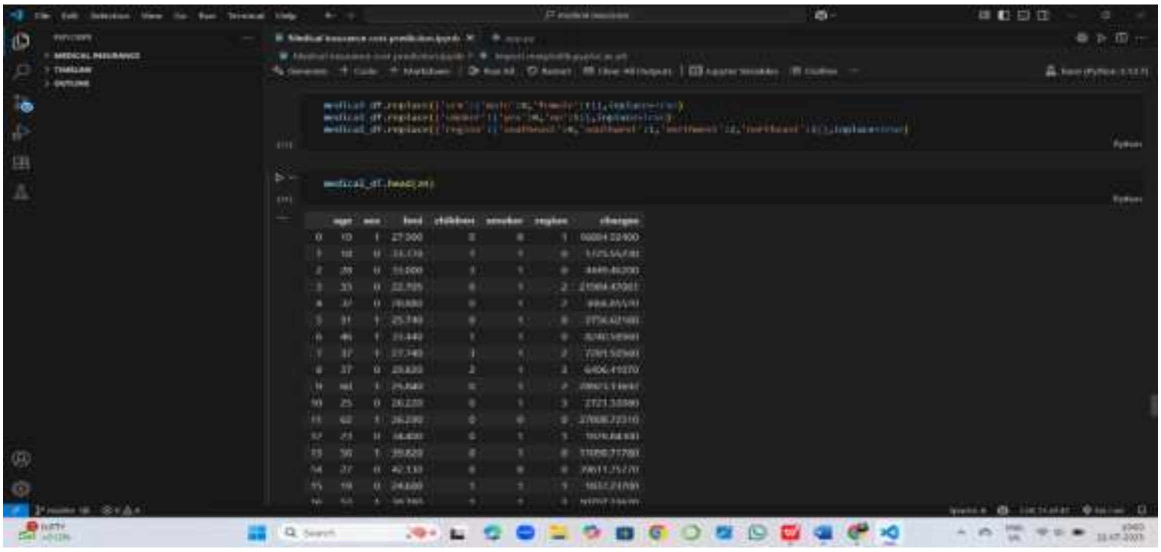
Data preprocessing is a crucial step in preparing the dataset for model training. In this project, preprocessing began with checking for missing values, null entries, or anomalies, but the dataset (insurance.csv) was found to be clean and complete. The next step was encoding categorical variables into numeric format, which is essential for machine learning models that require numerical inputs. The sex column was encoded as 0 for male and 1 for female, smoker as 0 for yes and 1 for no, and the region column was mapped to integers: southeast (0), southwest (1), northwest (2), and northeast (3). No feature scaling was required since the Linear Regression model used here can handle features with varied ranges reasonably well. These preprocessing steps ensured that the input data matched the expected format and structure, making it suitable for training and testing the prediction model effectively.



```
medical_df.head(20)
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692
10	25	male	26.220	0	no	northeast	2721.32080
11	62	female	26.290	0	yes	southeast	27808.72510
12	23	male	34.400	0	no	southwest	1826.84300
13	56	female	39.820	0	no	southeast	11090.71780
14	27	male	42.130	0	yes	southeast	39611.75770
15	19	male	24.600	1	no	southwest	1837.23700
16	52	female	30.780	1	no	northeast	10797.33620
17	23	male	23.845	0	no	northeast	2395.17155
18	56	male	40.300	0	no	southwest	10602.38500
19	30	male	35.300	0	yes	southwest	36837.46700

Fig 9: Before Conversion



```
medical_df.replace({'sex':{'male':0,'female':1}},inplace=True)
medical_df.replace({'smoker':{'yes':0,'no':1}},inplace=True)
medical_df.replace({'region':{'southeast':0,'southwest':1,'northwest':2,'northeast':3}},inplace=True)
```

```
medical_df.head(20)
```

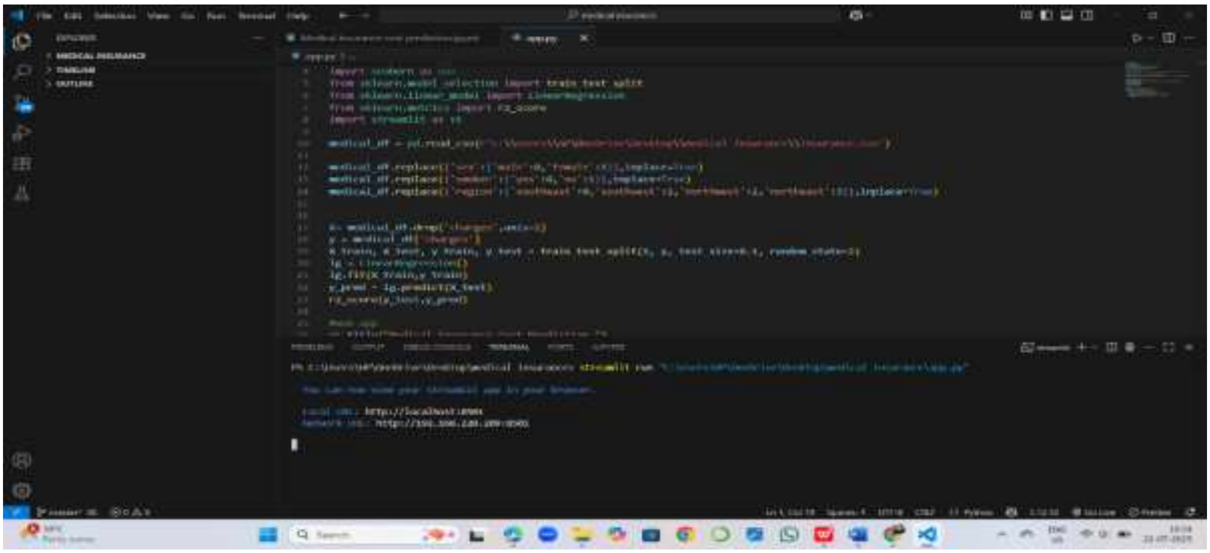
	age	sex	bmi	children	smoker	region	charges
0	19	1	27.900	0	0	1	16884.92400
1	18	0	33.770	1	1	0	1725.55230
2	28	0	33.000	3	1	0	4449.46200
3	33	0	22.705	0	1	2	21984.47061
4	32	0	28.880	0	1	2	3866.85520
5	31	1	25.740	0	1	0	3756.62160
6	46	1	33.440	1	1	0	8240.58960
7	37	1	27.740	3	1	2	7281.50560
8	37	0	29.830	2	1	3	6406.41070
9	60	1	25.840	0	1	2	28923.13692
10	25	0	26.220	0	1	3	2721.32080
11	62	1	26.290	0	0	0	27808.72510
12	23	0	34.400	0	1	1	1826.84300
13	56	1	39.820	0	1	0	11090.71780
14	27	0	42.130	0	0	0	39611.75770
15	19	0	24.600	1	1	1	1837.23700
16	52	1	30.780	1	1	3	10797.33620

Fig 10: After Conversion

3.4 Method of Process

The method of process in this project follows a structured machine learning pipeline. It starts with loading the insurance.csv dataset using Pandas, followed by preprocessing steps such as label encoding for categorical variables (sex, smoker, and region) and checking for null or inconsistent data. Once the data is cleaned and formatted, exploratory data analysis is performed using Seaborn and Matplotlib to visualize distributions and relationships between features. The cleaned data is then split into training (90%) and testing (10%) sets using

Scikit-learn's train_test_split() function. A **Linear Regression** model is trained on the training data using the LinearRegression() class. After training, predictions are made on the test set, and model accuracy is evaluated using the **R² score**. Once the model shows satisfactory performance, it is integrated into a **Streamlit** web app. This app allows users to input their information in real time, and the model predicts the corresponding insurance cost instantly.



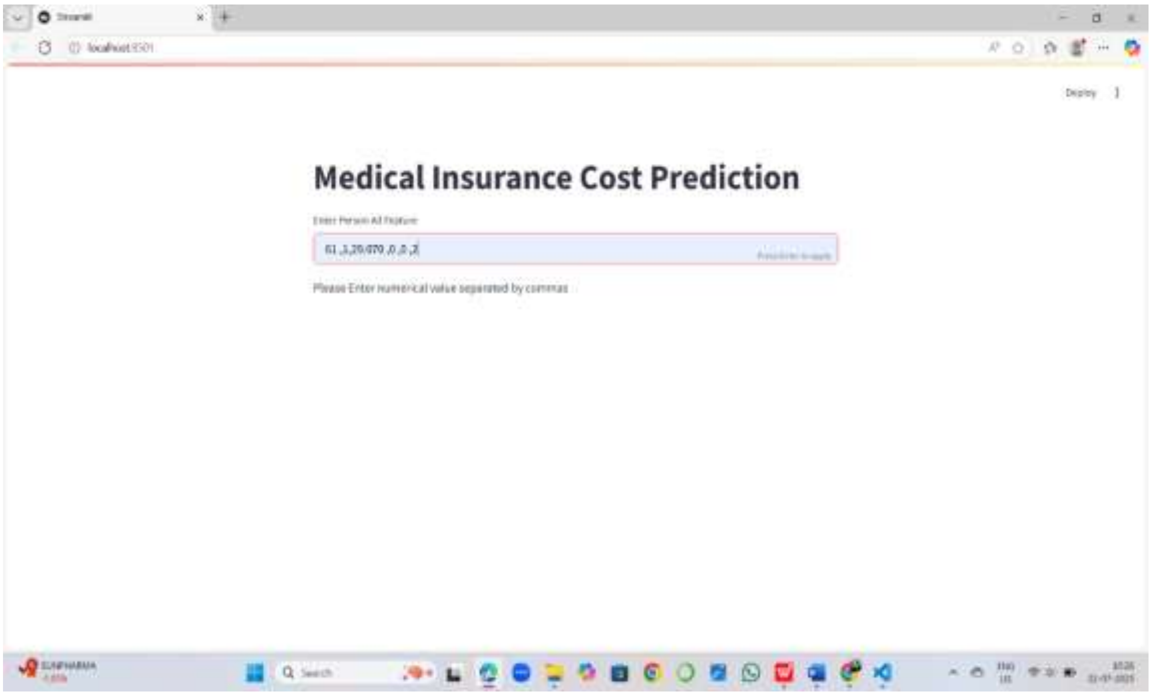
```
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import streamlit as st

medical_df = pd.read_csv([illegible])

medical_df.replace({'sex':{'male':0,'female':1}},inplace=True)
medical_df.replace({'smoker':{'yes':0,'no':1}},inplace=True)
medical_df.replace({'region':{'southeast':0,'southwest':1,'northwest':2,'northeast':3}},inplace=True)


X= medical_df.drop('charges',axis=1)
y = medical_df['charges']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=2)
lg = LinearRegression()
lg.fit(X_train,y_train)
y_pred = lg.predict(X_test)
r2_score(y_test,y_pred)
```

```
PS C:\Users\[illegible]\Desktop\medical insurance> streamlit run "[illegible]\app.py"

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://[illegible]:8501
```

Fig 11: Input interface for aap.py

3.5 Output

The final output of the system is a real-time prediction of medical insurance charges based on user-provided inputs. After the user enters six values—age, sex, BMI, number of children, smoker status, and region—into the Streamlit interface, the input is processed and reshaped to match the format used during model training. The trained Linear Regression model then performs a prediction and returns the estimated cost. This value is displayed immediately on the web interface, allowing users to see how their lifestyle and demographic factors influence insurance pricing. The output is simple, readable, and helps users make informed decisions about insurance planning.



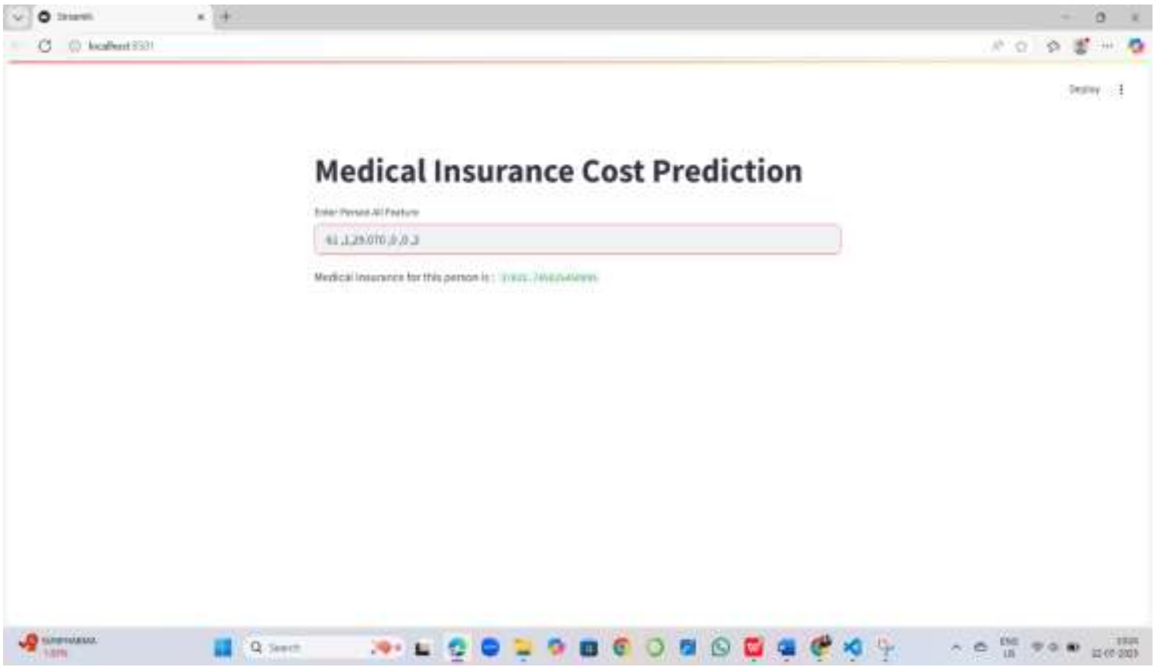
Medical Insurance Cost Prediction

Enter Person All Feature

61 ,1,20.070 ,0 ,0 ,2

Please Enter numerical value separated by commas

Fig 12: Predicting cost



Medical Insurance Cost Prediction

Enter Person All Feature

61 ,1,29.070 ,0 ,0 ,3

Medical insurance for this person is : [illegible]

Fig 13: Final Output

IV. RESULT

The model successfully predicted medical insurance charges with a satisfactory level of accuracy. Using the Linear Regression algorithm, the system achieved a reliable R^2 score, indicating that it explains a significant portion of the variability in the charges. Visual comparisons between actual and predicted values confirmed that the model performs well on test data.

V. DISCUSSION

The project demonstrates how machine learning, specifically Linear Regression, can be effectively applied to predict medical insurance costs. Key features such as smoking status, age, and BMI were found to significantly influence the charges. The model performed well on test data and provided fast, accurate results through a user-friendly web interface.

VI. CONCLUSION

This project successfully demonstrates how machine learning can be used to predict medical insurance costs based on personal and lifestyle factors. By using a Linear Regression model, we achieved accurate and interpretable results, and deployed the system through an easy-to-use Streamlit web application.

VII. FUTURE SCOPE

The system can be enhanced by incorporating advanced algorithms like Random Forest or Gradient Boosting to improve accuracy. Integration with electronic health records (EHRs), adding more diverse datasets, and supporting voice or mobile inputs can also make the system more powerful, accessible, and scalable in real-world healthcare applications.

VIII. ACKNOWLEDGEMENT

G.Manoj kumar working as an assistant professor in masters of computer applications (MCA) in SVPEC Vishakhapatnam Andhra Pradesh completed his Post graduation in Andhra University College of engineering (AUCE) with accredited by NAAC With his area of interest in java, python, database management system.

P.Lahari is pursuing her final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Artificial intelligence. P.Lahari has taken up her PG project on MEDICAL INSURANCE COST PREDICTION USING MACHINE LERNING and published the paper in connection to the project under the guidance of Mr G.Manoj Kumar, Assistant Professor, SVPEC.

IX. REFERENCES

1. Development of medical cost prediction model based on statistical machine learning using health insurance claims data
https://www.valueinhealthjournal.com/article/S1098-3015(18)33095-X/fulltext

2. An Efficient Framework for Predict Medical Insurance Costs Using Machine Learning
https://jocc.journals.ekb.eg/article_380150.html

3. Machine learning and prediction in medicine—beyond the peak of inflated expectations
http://pmc.ncbi.nlm.nih.gov/articles/PMC5953825/

4. Consumer credit-risk models via machine-learning algorithms
https://www.sciencedirect.com/science/article/abs/pii/S0378426610002372

5. A review on the application of deep learning in system health management
https://www.sciencedirect.com/science/article/abs/pii/S0888327017306064

6. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine
https://academic.oup.com/database/article/doi/10.1093/database/baaa010/5809229

7. [HTML] Comparing different supervised machine learning algorithms for disease prediction
https://link.springer.com/article/10.1186/s12911-019-1004-8?error=cookies_not_supporte

8. Predicting the future—big data, machine learning, and clinical medicine
https://pmc.ncbi.nlm.nih.gov/articles/PMC5070532/

9. Machine learning for medical diagnosis: history, state of the art and perspective
https://www.sciencedirect.com/science/article/abs/pii/S093336570100077X

10. Big data and machine learning algorithms for health-care delivery
https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(19)30149-4/abstract

11. Implementing machine learning in health care—addressing ethical challenges
https://pmc.ncbi.nlm.nih.gov/articles/PMC5962261/

12. Machine learning techniques for personalised medicine approaches **in** immune-mediated chronic inflammatory diseases: applications and challenges
https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2021.720694/full

13. Empirical asset pricing via machine learning
https://academic.oup.com/rfs/article/33/5/2223/5758276

14. Applications of machine learning in drug discovery and development
https://www.nature.com/articles/s41573-019-0024-5

15. The potential for artificial intelligence in healthcare
https://www.sciencedirect.com/science/article/pii/S2514664524010592

16. Privacy in the age of medical big data
https://www.nature.com/articles/s41591-018-0272-7

17. Multitask learning and benchmarking with clinical time series data
https://www.nature.com/articles/s41597-019-0103-9

18. Machine learning and deep learning
https://link.springer.com/article/10.1007/s12525-021-00475-2

19. [HTML] Can machine-learning improve cardiovascular risk prediction using routine clinical data?
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174944&source=post_page-----9cb5c78ee582----------------------

20. [HTML] Machine learning applications in cancer prognosis and prediction
https://www.sciencedirect.com/science/article/pii/S2001037014000464