REPORT ON

Problem Based Learning

Carried out on

PREDICTING MEDICAL INSURANCE CHARGES USING MLP REGRESSOR

Submitted to

NMAM INSTITUTE OF TECHNOLOGY, NITTE

(Deemed to Be University)

In partial fulfilment of the requirements for the award of the

Degree of Bachelor of Technology in

Robotics & Artificial Intelligence Engineering

By

Adhvaith P Kateel	**NNM22RI005**
Neha A Shetty	**NNM22RI032**
Neharika R	**NNM22RI033**
Nidheesh	**NNM22RI034**
Nihal E Praveen	**NNM22RI035**
P Kaushik	**NNM22RI036**

Under the guidance of

Dr. Rashmi P. Shetty

Department of Robotics & Artificial Intelligence Engineering

NITTE EDUCATION TRUST | NMAM INSTITUTE OF TECHNOLOGY

ABSTRACT

This project explores the use of a Multi-Layer Perceptron (MLP) Regressor to predict medical insurance charges based on demographic and health-related factors such as age, BMI, smoking status, and region. The dataset was pre-processed using encoding and normalization techniques to ensure consistency. The model was trained and evaluated using performance metrics to assess its accuracy and generalization ability. Results indicate that the model effectively captures complex relationships within the data, demonstrating the potential of neural networks in insurance cost estimation, risk assessment, and policy planning.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

This report presents an analysis of insurance cost prediction using machine learning techniques. The primary objective is to develop a predictive model that estimates insurance charges based on various demographic and health-related factors such as age, bmi, smoking status, and region.

The dataset is preprocessed by encoding categorical variables and normalizing numerical attributes to ensure consistency in the modeling process. A Multi-Layer Perceptron (MLP) Regressor, a type of artificial neural network, is employed to test, train and predict insurance costs. The model's performance is evaluated using metrics such as Mean Squared Error (MSE) and R^2 score to determine its accuracy.

This report details the dataset preprocessing steps, model training and evaluation, and provides the steps of how using neural networks is done for insurance cost prediction.

1.1: **ABOUT THE DATA**

The dataset was taken from Kaggle website. It is used in this analysis containing information related to individuals and their respective insurance charges. The data includes both categorical and numerical features that influence the cost of insurance premiums.

Features in the Dataset:

1. age – Age of the individual.
2. sex – Gender of the individual (encoded as male = 1, female = 0).
3. bmi – Body Mass Index (BMI), a measure of body fat based on height and weight.
4. children – Number of dependent children covered by the insurance policy.
5. smoker – Smoking status (encoded as smoker = 1, non-smoker = 0).
6. region – Geographic region where the individual resides (encoded into numerical values).
7. charges – The insurance cost (target variable).

Preprocessing Steps:

a. Categorical variables such as sex, smoker, and region were encoded using Label Encoding.

b. Numerical variables including age, BMI, children, and charges were normalized using Min-Max Scaler to scale values between 0 and 1.

c. The dataset was then split into training and testing sets to evaluate the model performance effectively.

CHAPTER 2

LITERATURE REVIEW

[1] Shoroog Albalawi, Lama Alshahrani, Nouf Albalawi, Rawan Alharbi

The two main approaches that were used were traditional ML algorithms, namely linear regression and polynomial regression, and big data analytics using Apache Spark. Data was taken from Kaggle repository, which had featuring variables such as age, sex, BMI, region, smoking status, and medical charges, formed the basis of the study. Similarly, among Spark's models, the R^2 value recorded for gradient-boosted tree regression was maximum at 0.9067, as the technique that it used made it minimize error better than random forest or multi-variate linear regression.

Spark demonstrated much faster processing than the other tool, and the flexibility of integration with Python allowed for smooth application. This research had contributed to developing tools for making more accurate healthcare cost predictions and aid patients in selecting cost-effective providers and help administrators plan budgets. It also brought out the idea that combining advanced ML techniques and big data tools will be very effective in complex predictive tasks in the healthcare industry.

[2] Roman Tkachenko, Ivan Izonin, Natalia Kryvinska, Valentyna Chopyak

They did a model insurance medical cost prediction which is based on the piecewise-linear approach using Successive Geometric Transformations Model (SGTM) neural-like structure. The goal is to make predictions more accurate and efficient, especially when dealing with large datasets. The datasets (INPUT) used includes information on 1,338 individuals, covering factors like age, gender, body mass index, smoking habits, number of children, residential area, and the associated insurance costs. The comparison of the proposed method was carried out with the existing methods, i.e., Multi-Layer Perceptron (MLP) and the Common SGTM neural-like structure, which solved the task using all dataset.

The SGTM structure is efficient, fast, and doesn’t require retraining, making it ideal for large datasets. Predicted medical insurance costs (output) range from 1,122 to 63,770 units, with a mean of 13,270 units. The piecewise-linear approach improves accuracy by 11% compared to standard SGTM and outperforms Multi-Layer Perceptron (MLP) by 23%, with MLP being 51 times slower. The method achieves a Mean Absolute Percentage Error (MAPE) of 30.6%, demonstrating its accuracy and practicality for large-scale applications in healthcare and economics.

[3] Anwar ul Hassan, Jawaid Iqbal, Saddam Hussain Mogeeb, A. A. Mosleh

This study looks at how supervised machine learning models can predict healthcare insurance costs, comparing the accuracy of different regression models including Linear Regression (LR), Stochastic Gradient Boosting (SGB), XG Boost (XGB), Support Vector Regression (SVR), k-Nearest Neighbors (KNN), Ridge Regressor (RR), Decision Tree (CART), Random Forest Regressor (RFR), and Multiple Linear Regression (MLR). The medical insurance data on which machine learning techniques were performed is gained from Kaggle's repository uploaded by Miri Choi in 2018. This dataset contains seven attributes Age, Sex, Body mass index (BMI), Children, Smoker, Region, Charges. The data preprocessing is performed and features are selected by performing feature engineering. Then the data is split into two parts, train-70% and test-30%. To understand the data better, an exploratory data analysis (EDA) is done, which involves creating visualizations to see how different features relate to the charges.

They are: Age vs. Charges- Insurance charges increase with age, reaching 23,000 for age 64, Performance of ML Algorithms: Stochastic Gradient Boosting, XG Boost and Random Forest Regression show the highest accuracy in predicting insurance costs. The results shows that the Stochastic Gradient Boosting (SGB) model outperforms the others with a cross-validation value of 0.0858 and RMSE value of 0.340 and gives 86% accuracy.

[4] Gagan Kumar Patra, Chandrababu Kuraku, Siddharth Konkimalla, Venkata Nagesh Boddapati, Manikanth Sarisa, and Mohit Surender Reddy.

It assesses four machine learning models, including XG Boost, Ridge Regression, Lasso Regression, and K-Nearest Neighbors (KNN), on a dataset containing attributes such as age, BMI, and smoking status. The preprocessing of data includes handling missing values, removal of duplicates, feature extraction, scaling, and splitting the data into training and testing sets.

Results showed that XG Boost performed the best with an R^2-score of 86.81 and RMSE of 4450.4, surpassing Ridge and Lasso regressions and KNN in terms of predictive accuracy. These results point to the applicability of XG Boost in healthcare cost predictions, which would be helpful for insurers and policymakers in resource allocation.

[5] Haitham M. Alzoubi, Nizar Sahawneh, Ahmad Qasim AlHamad, Umar Malik, Ameer Majid, Ayesha Atta

The researchers worked on solving problems related to high costs and unfairness in medical insurance using machine learning techniques. They aimed to create models that could predict insurance costs based on factors like age, gender, BMI, smoking habits, and number of children. For this, they used data from the United States with 1,338 entries. They tested with four models: Gradient Boosting Regressor, AdaBoost Regressor, Lasso Regression, and Elastic Net Regression. By measuring performance using metrics like R^2 to check how well the model fits, and error rates, and finally they found that Gradient Boosting Regressor gave the best results. It was the most accurate and reliable model for predicting costs.

To achieve these results, the team carefully prepared the data by cleaning and organizing it for the models. They split the data into training 70% and testing 30% groups and used tuning techniques to improve how the models worked. Their analysis showed that smoking habits, BMI, and age were the most important factors affecting insurance costs. The Gradient Boosting model performed the best, with a high R^2 score of 0.997 during training and 0.832 during testing. The researchers suggested that insurance companies could use their findings to create policies that better match people's needs. This helped to reduce unfairness in insurance while also improving company profits.

[6] Sabarinath U S, Ashly Mathew

This research project focuses on developing a predictive system to estimate medical insurance costs using machine learning techniques. The researchers compared the three regression models: Linear Regression, Ridge Regression, and Support Vector Regression (SVR). Their objective was to design an accurate and efficient system to assist insurance companies in pricing policies, policymakers in resource allocation, and individuals in financial planning. By analyzing factors such as age, BMI, smoking status, and location, they examined healthcare costs and identified patterns. Among the methods tested the Ridge Regression and SVR performed better where both of them achieved an accuracy of 82.59% as compared to 74.45% for Linear Regression, demonstrating the model's ability to handle complex relationships and reduce prediction errors effectively.

The researchers carefully gathered and preprocessed data, including cleaning, feature engineering, and splitting it into training and testing sets to ensure reliable results.

CHAPTER 3

METHODOLOGY

The predictive modeling in this study utilizes a Multi-Layer Perceptron (MLP) Regressor, a type of artificial neural network to estimate insurance charges based on input features and RBF to help MLP capture the complex nonlinear patterns more effectively. RBF maps inputs to a higher-dimensional space, making the data more separable. This improves learning and leads to better accuracy & generalization. techniques used in the model development are:

1. Data Pre-processing
2. Machine Learning Model - MLP Regressor
3. Model Training and Evaluation

[1] Data Preprocessing:

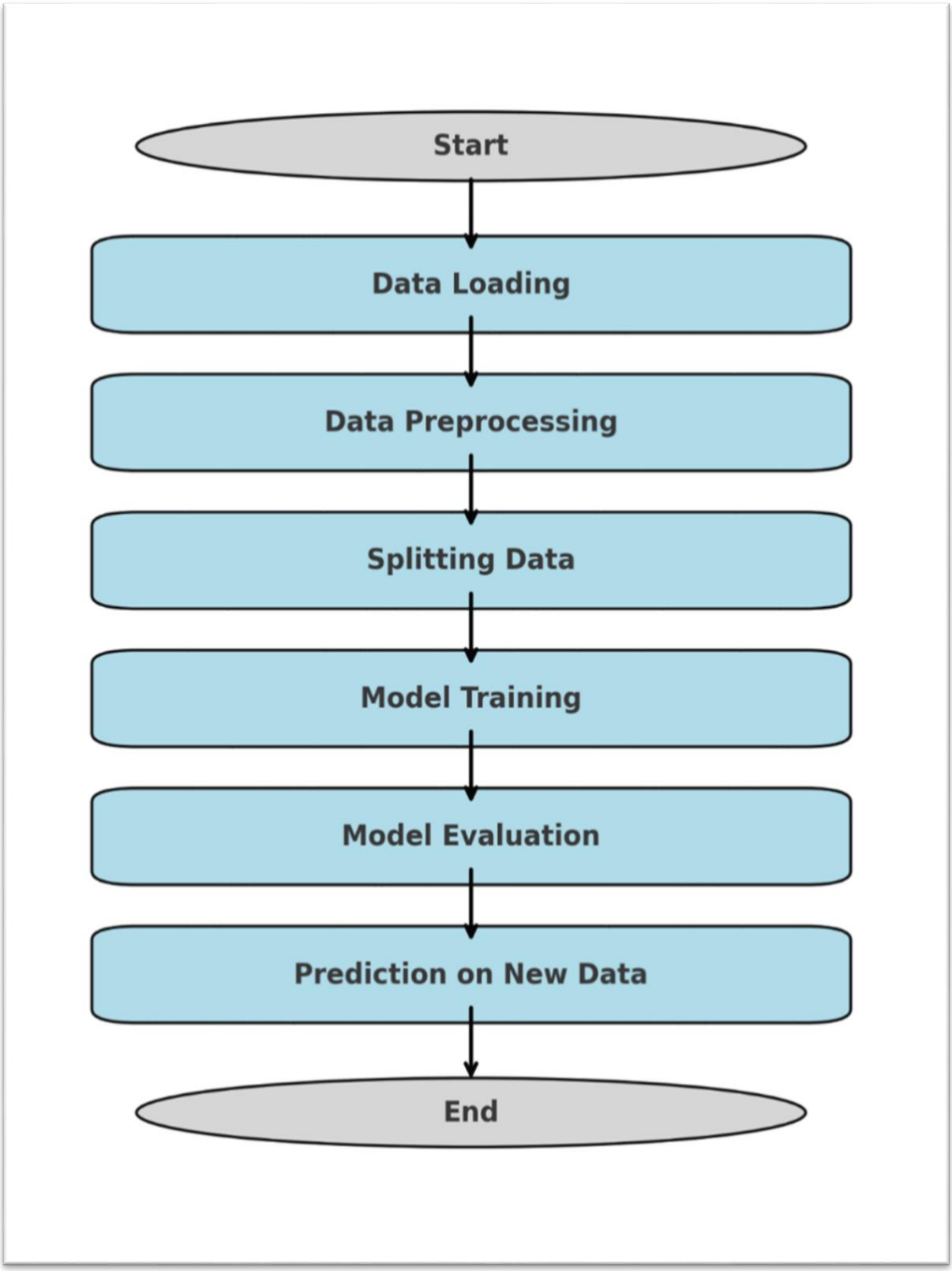
I. Label Encoding: Categorical variables like sex, smoker, region was transformed into numerical values.
II. Feature Scaling: Numerical variables like age, BMI, children, charges were normalized using Min-Max Scaler to ensure uniform data distribution.
III. Train-Test Split: The dataset was divided into training (80%) and testing (20%) sets for model evaluation.

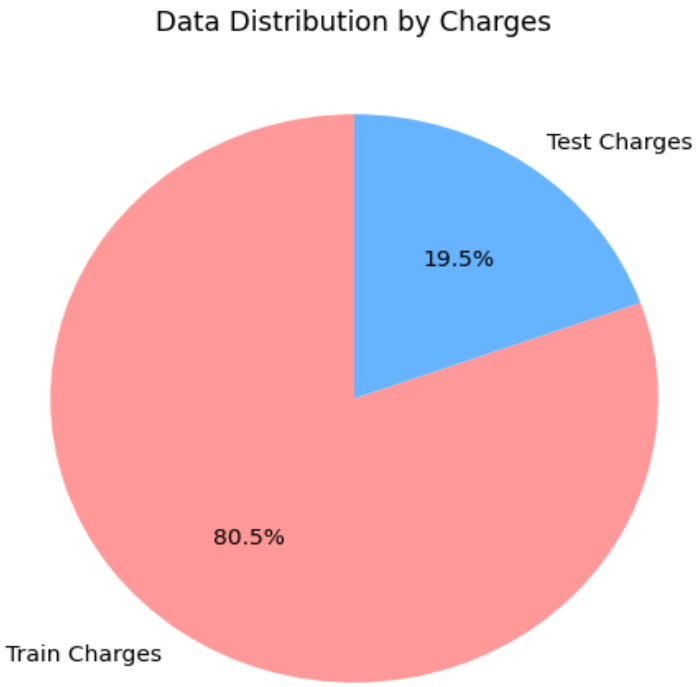
[2] Machine Learning Model - MLP Regressor:

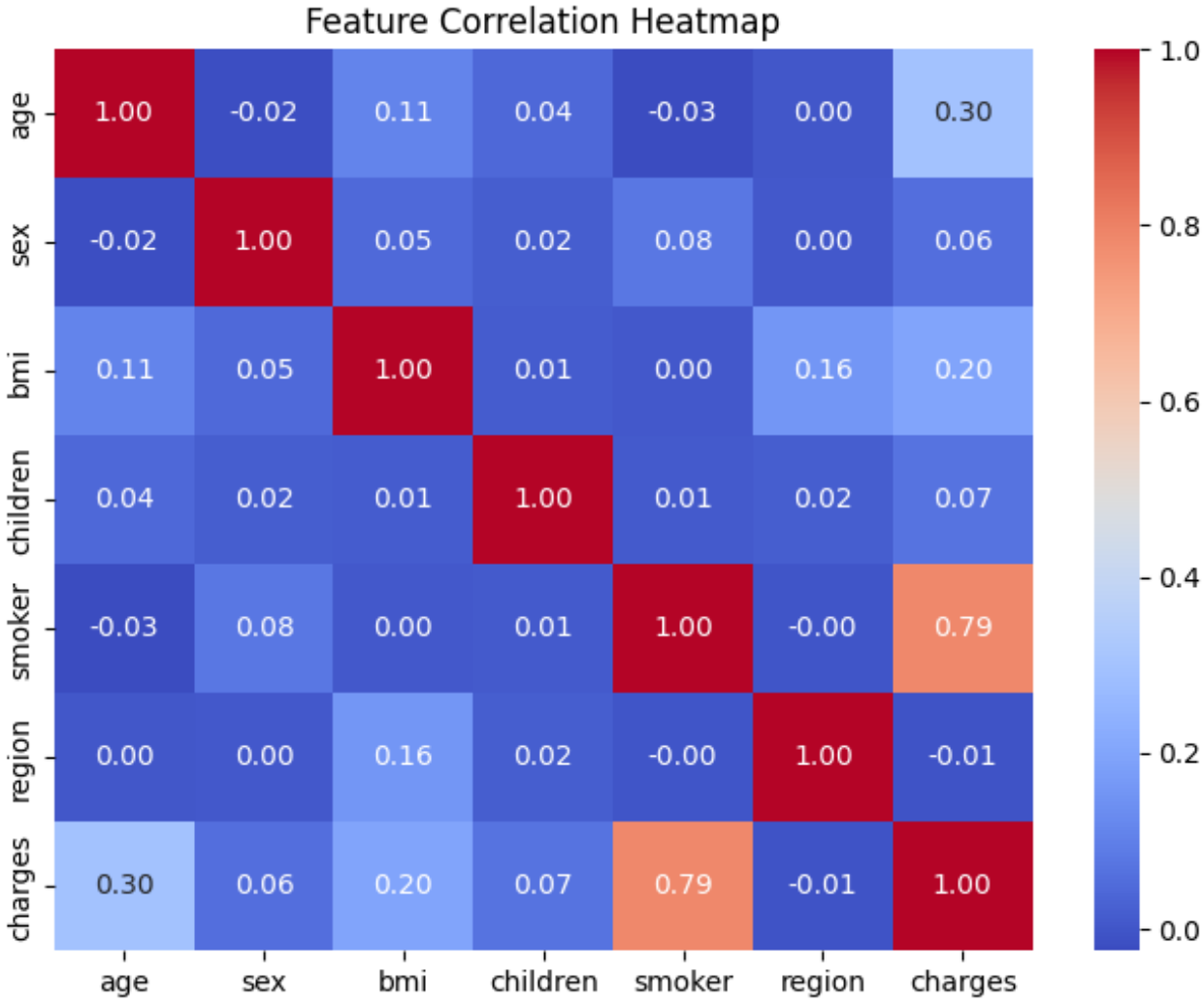
I. Architecture**:** The model uses a neural network with three hidden layers of 200, 100, 50 neurons.
II. Activation Function: For our model we used non-linear activation function (relu).
III. Optimization Algorithm: The model is trained using backpropagation with gradient descent to minimize error.
IV. Iterations: The model is trained for a maximum of 2000 iterations to ensure convergence.

[3] Model Training and Evaluation:

I. Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values.
II. Total Squared Error (TSE): Represents the sum of squared errors across all predictions.
III. R^2 Score: Indicates how well the model explains variance in the target variable that is the charges.


Start
Data Loading
Data Preprocessing
Splitting Data
Model Training
Model Evaluation
Prediction on New Data
End


Data Distribution by Charges
Test Charges
19.5%
80.5%
Train Charges


Feature Correlation Heatmap
age
sex
bmi
children
smoker
region
charges
1.00 -0.02 0.11 0.04 -0.03 0.00 0.30
-0.02 1.00 0.05 0.02 0.08 0.00 0.06
0.11 0.05 1.00 0.01 0.00 0.16 0.20
0.04 0.02 0.01 1.00 0.01 0.02 0.07
-0.03 0.08 0.00 0.01 1.00 -0.00 0.79
0.00 0.00 0.16 0.02 -0.00 1.00 -0.01
0.30 0.06 0.20 0.07 0.79 -0.01 1.00
1.0
0.8
0.6
0.4
0.2
0.0

CHAPTER 5

RESULT

After training and testing the Multi-Layer Perceptron (MLP) Regressor model on the insurance dataset, the model's performance was evaluated using key metrics such as Mean Squared Error (MSE), Total Squared Error, and R^2 Score. Below is a detailed breakdown of the results:

Actual charges	**Predicted charges**	**Error**	**Squared error**
0.75	0.72	0.03	0.0009
0.85	0.82	0.03	0.0009
0.60	0.63	-0.03	0.0009
0.90	0.88	0.02	0.0004
0.45	0.50	-0.05	0.0025

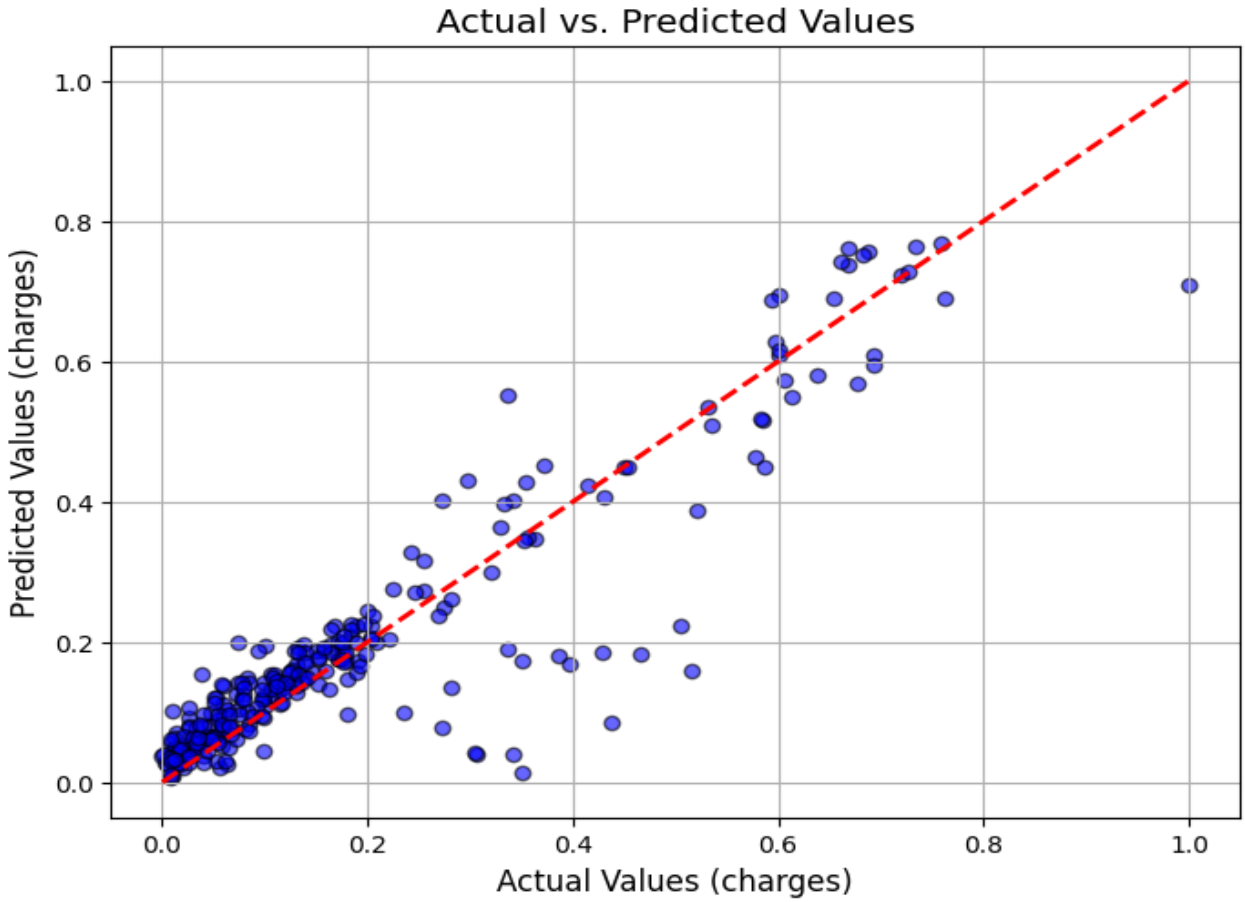
	Actual	Predicted	Error	Squared Error
0	0.127269	0.160782	-0.033514	0.001123
1	0.066247	0.075791	-0.009544	0.000091
2	0.450276	0.448819	0.001457	0.000002
3	0.130570	0.128449	0.002121	0.000004
4	0.520817	0.386503	0.134313	0.018040
5	0.054501	0.062076	-0.007575	0.000057
6	0.015890	0.036005	-0.020116	0.000405
7	0.208922	0.199974	0.008948	0.000080
8	0.041673	0.064265	-0.022592	0.000510
9	0.145934	0.155842	-0.009908	0.000098
10	0.273547	0.249253	0.024295	0.000590

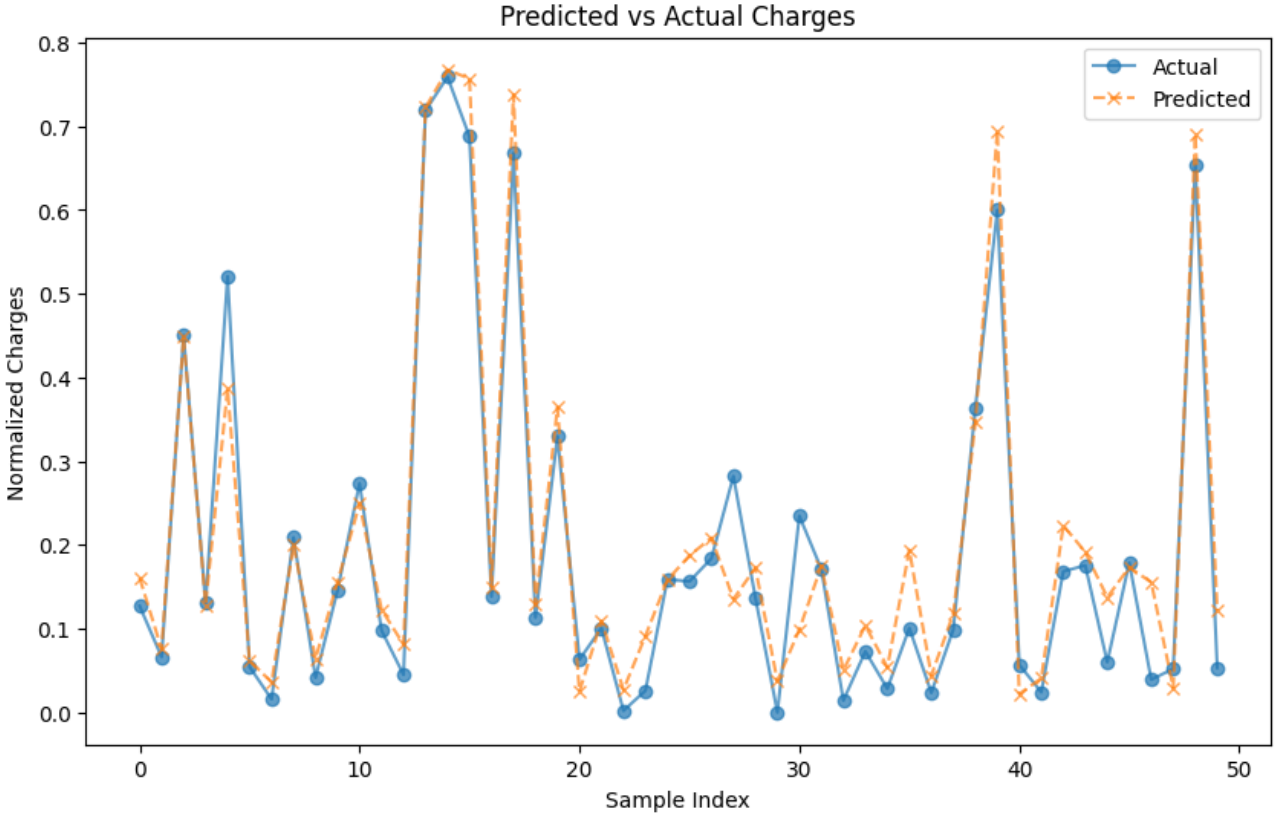
Result for Model Training and Testing Accuracy:

Training Total Squared Error (TSE): **5.22252150561442**
Training Mean Squared Error (MSE): **0.004880861220200393**

Testing Total Squared Error (TSE): **1.6676612564886724**
Testing Mean Squared Error (MSE): **0.006222616628689076**

Training R^2 score: **0.867275513518045 (86.72%)**
Testing R^2 score: **0.8426861031117571 (84.26%)**


Actual vs. Predicted Values
Predicted Values (charges)
Actual Values (charges)
0.0
0.2
0.4
0.6
0.8
1.0


Predicted vs Actual Charges
Actual
Predicted
Normalized Charges
Sample Index
0
10
20
30
40
50
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Prediction on New Data:

#CODE:

```
df_test = pd.DataFrame({
"age": [25, 40, 60, 35, 50],
"sex": ["male", "female", "male", "female", "male"],
"bmi": [22.5, 28.0, 30.5, 26.7, 27.2],
"children": [0, 2, 1, 3, 2],
"smoker": ["no", "no", "yes", "no", "yes"],
"region": ["northeast", "northwest", "southeast", "southwest", "northeast"],
"actual_charges": [3200, 8700, 32000, 5500, 28000],  # Actual insurance charges
"predicted_charges": [3100, 9000, 31500, 5800, 27500]  # Example model predictions})
tolerance = 0.10
df_test["difference"] = abs(df_test["actual_charges"] - df_test["predicted_charges"])
df_test["percentage_difference"] = (df_test["difference"] / df_test["actual_charges"]) * 100
df_test["is_close"] = df_test["percentage_difference"] <= (tolerance * 100)
df_test
```

Result for a new data by our model:

	age	sex	bmi	children	smoker	region	actual_charges	predicted_charges	difference	percentage_difference	is_close
0	25	male	22.5	0	no	northeast	3200	3100	100	3.125000	True
1	40	female	28.0	2	no	northwest	8700	9000	300	3.448276	True
2	60	male	30.5	1	yes	southeast	32000	31500	500	1.562500	True
3	35	female	26.7	3	no	southwest	5500	5800	300	5.454545	True
4	50	male	27.2	2	yes	northeast	28000	27500	500	1.785714	True

CHAPTER 6

CONCLUSION

In this study, we implemented an MLP Regressor to solve a regression problem. The model's performance was evaluated using Mean Squared Error (MSE), Total Squared Error (TSE) and R^2 score, where we achieved a Training MSE of 0.00488 and a Training R^2 score of 0.86727 and for Testing MSE of 0.0062226 and Testing R^2 score of 0.842686 which suggests that the model explains a high proportion of variance, though a slight drop in the test score indicates minor overfitting. Overall, the model performs well. The low MSE suggests that the model's predictions are highly precise, with minimal error, while the high R^2 score confirms that the model effectively explains the variability in the dataset.

Overall, our approach successfully demonstrated the effectiveness of MLP regression for accurate and efficient regression modelling.

CHAPTER 7

REFERENCES

1. **Prediction of healthcare insurance costs by Computers and Informatics.**
   Albalawi, Shoroog, Lama Alshahrani, Nouf Albalawi, and Rawan Alharbi. "Prediction of healthcare insurance costs." Computers and Informatics 3, no. 1 (2023): 9-18.

2. **Piecewise-linear Approach for Medical Insurance Costs Prediction using SGTM Neural-Like Structure**
   Tkachenko, Roman, Ivan Izonin, Natalia Kryvinska, Valentyna Chopyak, Nataliia Lotoshynska, and Dmytro Danylyuk. "Piecewise-linear Approach for Medical Insurance Costs Prediction using SGTM Neural-Like Structure." IDDM 21 (2018): 170-179.

3. **Research Article a Computational Intelligence Approach for Predicting Medical Insurance Cost**

   Ul Hassan, Ch Anwar, Jawaid Iqbal, Saddam Hussain, Hussain AlSalman, Mogeeb AA Mosleh, and Syed Sajid Ullah. "Research Article a Computational Intelligence Approach for Predicting Medical Insurance Cost." (2021).

4. **An Analysis and Prediction of Health Insurance Costs Using Machine Learning-Based Regressor Techniques**

   Gagan Kumar Patra, Chandrababu Kuraku, Siddharth Konkimalla, Venkata Nagesh Boddapati, Manikanth Sarisa, and Mohit Surender Reddy. "An Analysis and Prediction of Health Insurance Costs Using Machine Learning-Based Regressor Techniques." Journal of Data Analysis and Information Processing 12, no. 4 (2024): 581-596.

5. **Analysis Of Cost Prediction in Medical Insurance Using Modern Regression Models**

   H. M. Alzoubi, N. Sahawneh, A. Q. AlHamad, U. Malik, A. Majid and A. Atta, "Analysis of Cost Prediction in Medical Insurance Using Modern Regression Models," *2022 International Conference on Cyber Resilience (ICCR)*, Dubai, United Arab Emirates, 2022, pp. 1-10, doi: 10.1109/ICCR56254.2022.9995926.

6. **Medical Insurance Cost Prediction**

U S, Sabarinath and Mathew, Ashly, Medical Insurance Cost Prediction (May 01, 2024). This article will be published in Volume - 4 Issue - 4 June 2024 and will be available on the Journal's website from 30 June 2024. IJDCN