Shahid Beheshti University

Artificial Neural Networks

M.Sc - Fall 2024

Author:
Zahra Mohammad Beigi

Student Number:
402422144

Assignment 1 - Part 2

October 23, 2024

Contents

1 Introduction

The purpose of this project is to predict individual medical costs using a Multi-Layer Perceptron (MLP) model. The prediction is based on various personal factors provided in the dataset, including age, gender, body mass index (BMI), smoking status, number of dependents, and residential region in the US.

Health insurance charges vary greatly depending on these personal factors, and accurately predicting these charges can be beneficial for insurers and individuals alike. The goal of this assignment is to design and train a custom MLP model to solve this regression task. The model will be optimized through careful tuning of its architecture and hyperparameters to achieve the best predictive performance.

This project presents several challenges, including the need to balance the complexity of the MLP architecture with overfitting and generalization, and the careful selection of hyperparameters to improve the model's performance.

2 Data Analysis

2.1 Dataset Overview

The dataset consists of 1338 entries and 7 columns. The following is a brief description of the columns:

- **age**: The age of the primary beneficiary. (integer).
- **sex**: The gender of the insurance holder, either male or female. (object).
- **bmi**: The body mass index (BMI) of the individual, a measure that correlates weight with height. The ideal range is 18.5 to 24.9.(float).
- **children**: The number of dependents covered by the health insurance policy. (integer).
- **smoker**: Whether the individual is a smoker (yes/no) (object).
- **region**: The geographical region of the policyholder in the US (northeast, southeast, southwest, northwest) (object).
- **charges**: The actual medical costs billed by health insurance. This is the target variable we aim to predict (float).

The dataset contains both categorical features (like sex, smoker, region) and continuous features (like age, bmi, children). The children column will be converted to categorical. All columns have 1338 non-null entries, ensuring no missing values in this dataset.

2.2 Exploratory Data Analysis (EDA)

Summary Statistics

	age	bmi	charges
Count	1338.00	1338.00	1338.00
Mean	39.21	30.66	13270.42
Std	14.05	6.10	12110.01
Min	18.00	15.96	1121.87
25%	27.00	26.30	4740.29
50%	39.00	30.40	9382.03
75%	51.00	34.69	16639.91
Max	64.00	53.13	63770.43

Table 1: Summary statistics for numeric variables

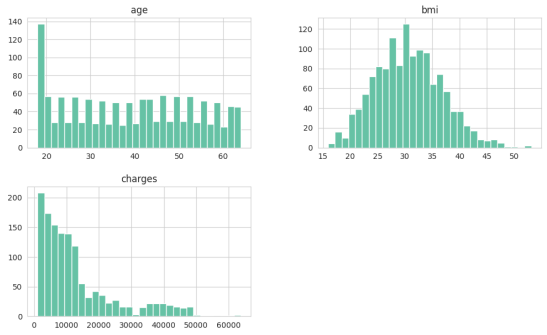
	sex	children	smoker	region
Count	1338	1338	1338	1338
Unique	2	6	2	4
Top	male	0	no	southeast
Freq	676	574	1064	364

Table 2: Summary statistics for categorical variables

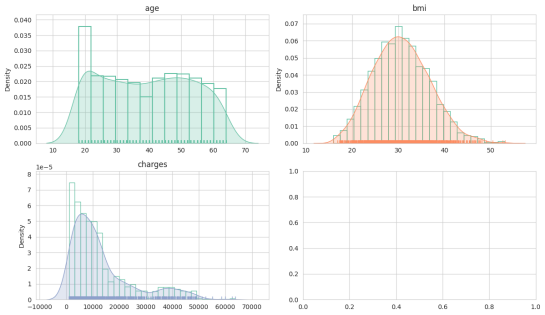
- **Numeric Variables**: The average age is 39.21 years, with a standard deviation of 14.05. The average BMI is 30.66, with a minimum of 15.96 and a maximum of 53.13. The average insurance charges are $13,270.42, ranging from $1,121.87 to $63,770.43.
- **Categorical Variables**: The dataset includes two sexes (with males slightly more frequent), up to six children per individual (most have none), and a majority of non-smokers. The most frequent region in the dataset is the southeast.

Further analysis will delve into relationships between these variables and their impact on insurance charges.

Histograms for Numeric Variables



(a) Histograms of Numeric Variables



(b) Density and KDE Plots of Numeric Variables

Figure 1: Distribution of Numeric Variables (Age, BMI)

The skewness and kurtosis of the numeric variables are as follows: Table 3 presents the skewness and kurtosis values for the numeric variables in the dataset.

Variable	Skewness	Kurtosis
Age	0.056	-1.245
BMI	0.284	-0.055
Charges	1.514	1.596

Table 3: Skewness and Kurtosis for Numeric Variables

The histograms in Figure 1a and the density/KDE plots in Figure 1b provide insights into the distribution of numeric variables, including age, BMI, and charges.

- **Age**: The age distribution is roughly uniform, with a slight skew towards younger individuals. The rug plot and KDE curve suggest a balanced age distribution with no extreme outliers.
- **BMI**: The BMI histogram shows a right-skewed distribution, with most values concentrated between 20 and 40. The KDE curve confirms that higher BMIs are less frequent, and the rug plot reveals occasional outliers in the upper tail.
- **Charges**: The charges distribution exhibits a significant positive skew, with most individuals having charges below 20,000, while a small number incur much higher charges, as seen in the right tail of the KDE curve.

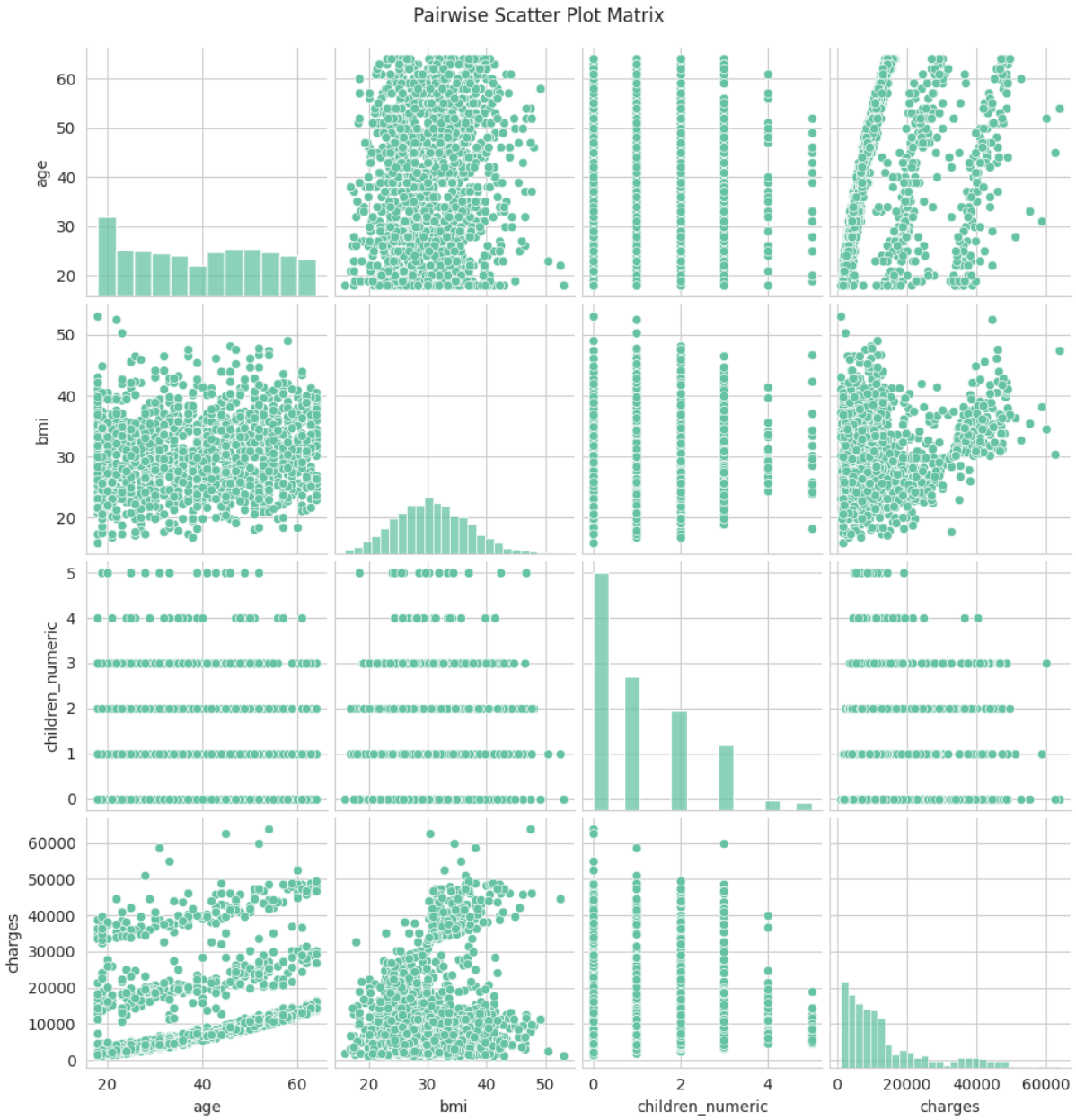
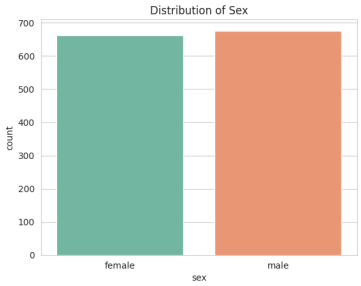


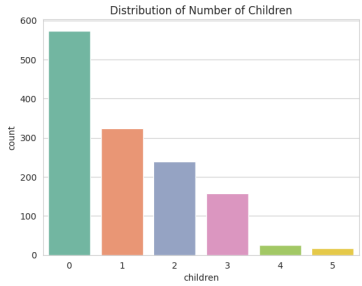
Figure 2: Scatter Plot Matrix of Numerical Variable.

Demographic Distributions

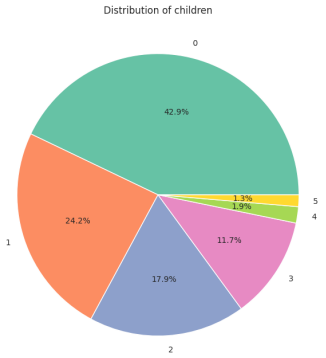
In this section, we analyze the demographic distributions within the dataset, focusing on the variables such as sex, number of children, smoking status, and region.



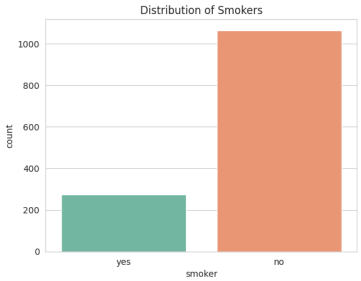
(a) Distribution of Sex



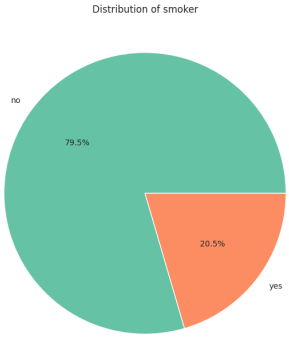
(b) Distribution of Number of Children.



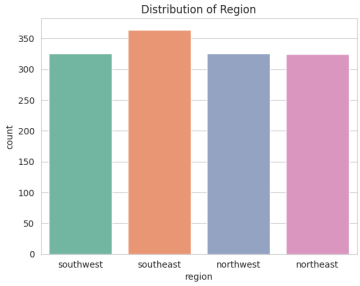
(c) Pie Chart of Number of Children.



(d) Distribution of Smokers



(e) Pie Chart of Smokers.



(f) Distribution of Region.

Figure 3: Demographic Distributions of Categorical Variables.

The dataset reveals several interesting insights regarding the demographic characteristics of the individuals represented. In terms of sex, there is a relatively balanced distribution with **676 males** and **662 females**, indicating that gender is not a significant factor in skewing the data. When examining the number of children, the majority of individuals (**574**) report having **no children**, while the numbers decrease progressively with higher counts—**324** have **one child**, **240** have **two**, **157** have **three**, and only a few have larger families (**25** have **four** children, and **18** have **five**). This suggests that smaller families are more prevalent within this population.

The smoking status presents a stark contrast; **1,064** individuals are **non-smokers**, while only **274** are **smokers**, highlighting a significant majority of non-smokers. This distribution may have important implications for health-related analyses and insurance charge assessments. Lastly, the regional distribution shows that the **Southeast** has the highest representation at **364**, while the **Southwest**, **Northwest**, and **Northeast** regions are relatively balanced, each with counts close to **325** to **324**. Overall, the comparisons among these variables reveal both demographic diversity and concentration within the dataset, laying the groundwork for further exploratory analysis.

Correlation Analysis

The correlation matrix provides insights into the relationships between the numerical features in the dataset. The following table presents the correlation coefficients expressed as percentages:

	Age (%)	**BMI (%)**	**Charges (%)**
Age	100.00	10.93	29.90
BMI	10.93	100.00	19.83
Charges	29.90	19.83	100.00

Table 4: Correlation coefficients between numerical variables expressed as percentages.

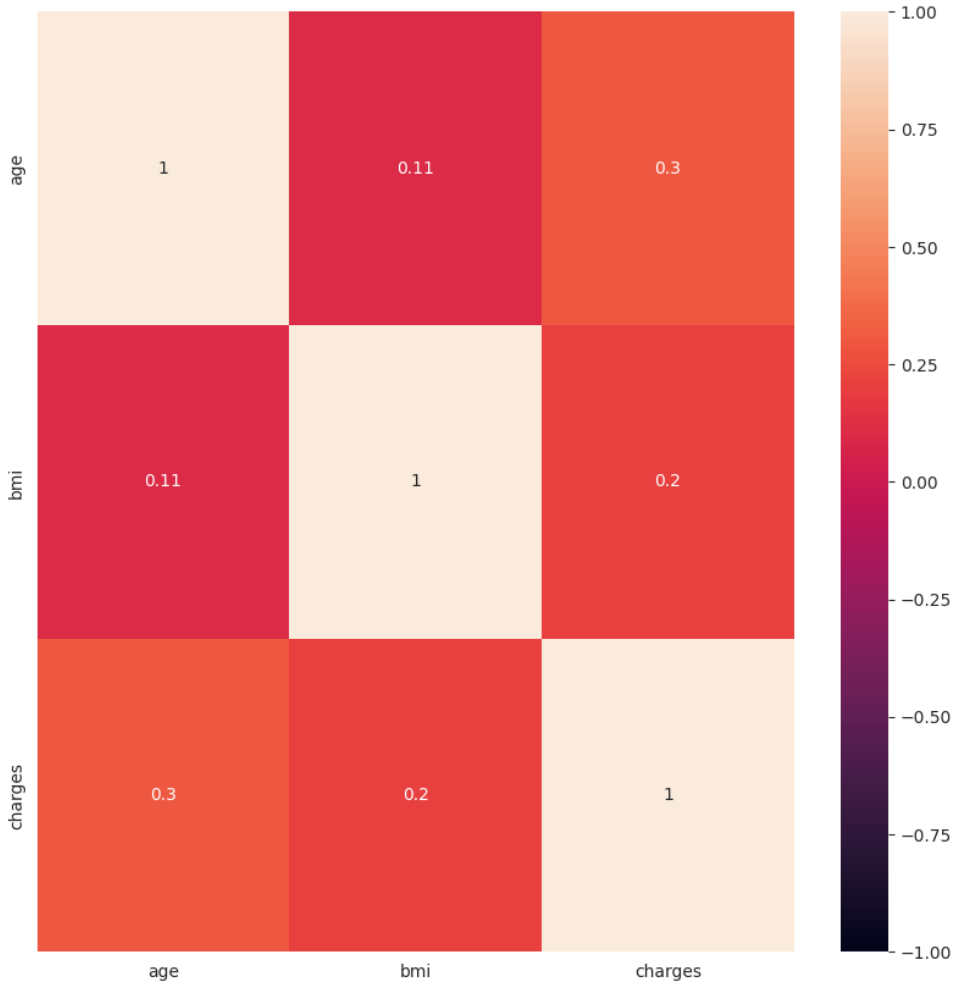


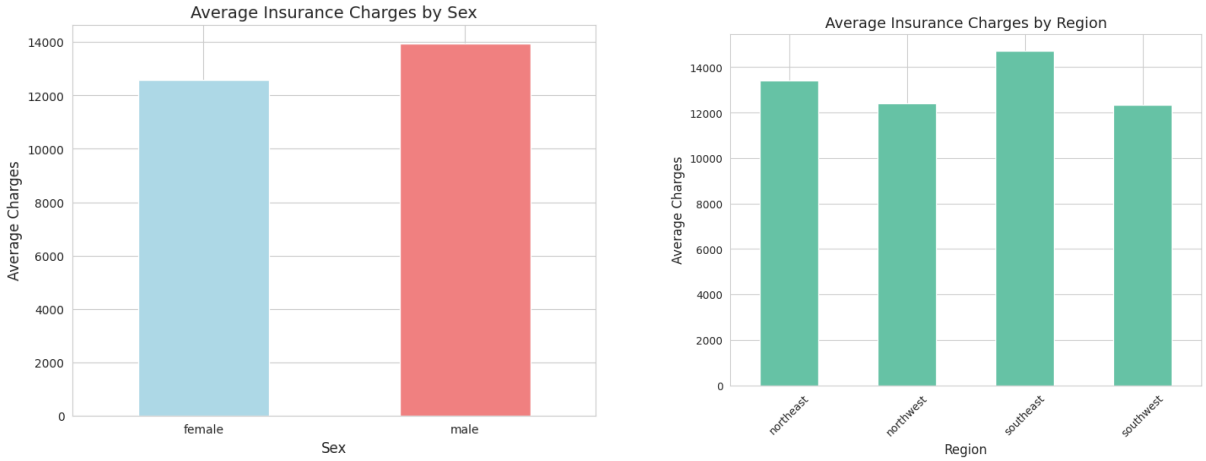
Figure 4: Heatmap of the correlation matrix for numerical variables.

From the correlation matrix, we observe that:

- *Age* shows a moderate positive correlation with *Charges* (29.90%), indicating that as age increases, the charges tend to rise as well. - *BMI* has a low correlation with both *Age* (10.93%) and *Charges* (19.83%), suggesting that BMI is not a strong predictor for these variables.

Analysis of Average Insurance Charges

In this section, we examine the average insurance charges based on various demographic and health-related factors. The findings highlight significant disparities in charges among different groups.



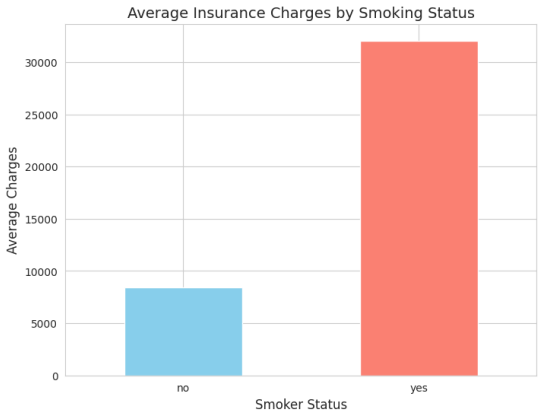
(a) Average Insurance Charges by Sex (b) Average Insurance Charges by Region

Figure 5: Average Insurance Charges by Sex and Region

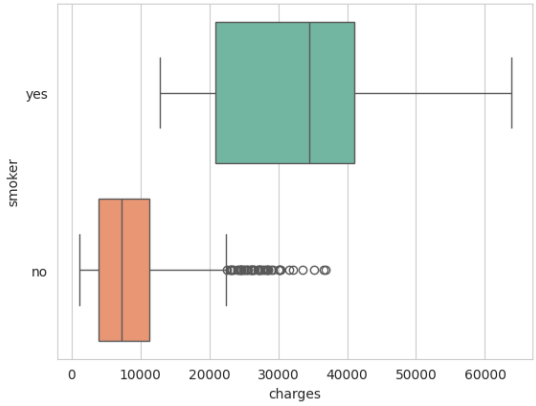
Average Charges by Sex As shown in Figure 5a, the average insurance charges for females are $12,569.58, while for males, it is $13,956.75. This indicates that males tend to incur higher insurance charges compared to females.

Average Charges by Region Figure 5b presents the average insurance charges across different regions. The charges are as follows:
- Northeast: $13,406.38
- Northwest: $12,417.58
- Southeast: $14,735.41
- Southwest: $12,346.94

The Southeast region has the highest average charges, while the Northwest region has the lowest, indicating regional variations in insurance pricing.



(a) Average Insurance Charges by Smoking Status



(b) Boxplot of Insurance Charges by Smoking Status

??

Figure 6: Average Insurance Charges by Smoking Status

Average Charges by Smoking Status The analysis of average charges by smoking status, depicted in Figure 6, reveals a stark contrast. Smokers have an average charge of $32,050.23, significantly higher than the $8,434.27 average charge for non-smokers. This substantial difference underscores the impact of smoking on insurance costs.

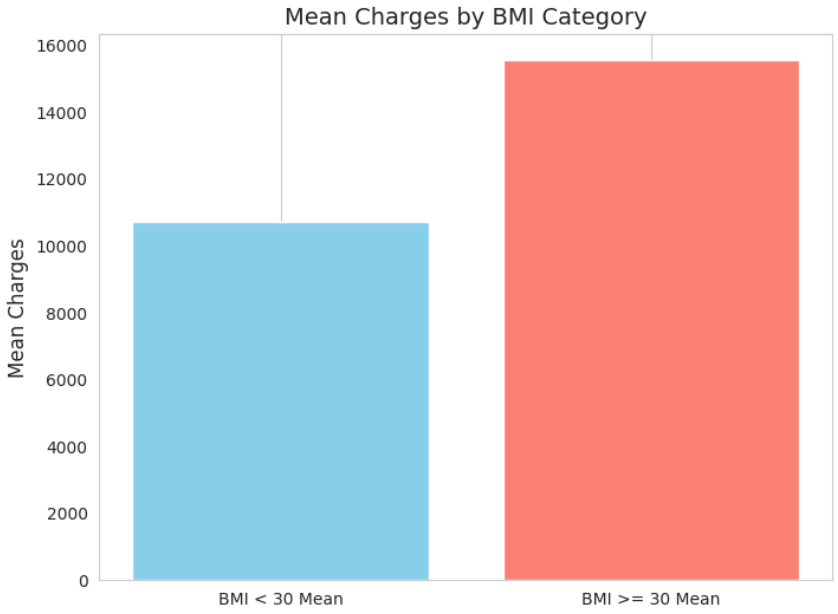


Figure 7: Mean Insurance Charges by BMI Category

Mean Charges by BMI Category The mean charges for individuals with different BMI categories are summarized as follows: (The BMI mean is almost 30)
- Mean Charges for BMI $<$ 30: $10,713.67
- Mean Charges for BMI $>=$ 30: $15,552.34

Figure 7 illustrates these findings, indicating that individuals with a BMI of 30 or higher tend to incur significantly higher average insurance charges compared to those with a BMI below 30.

Missing Data Analysis

Figure 8 presents the missing data matrix for the dataset. The matrix indicates that there are no missing values across any of the variables in the dataset. This is a positive finding, as it suggests that the dataset is complete and ready for analysis without the need for imputation or handling of missing data.

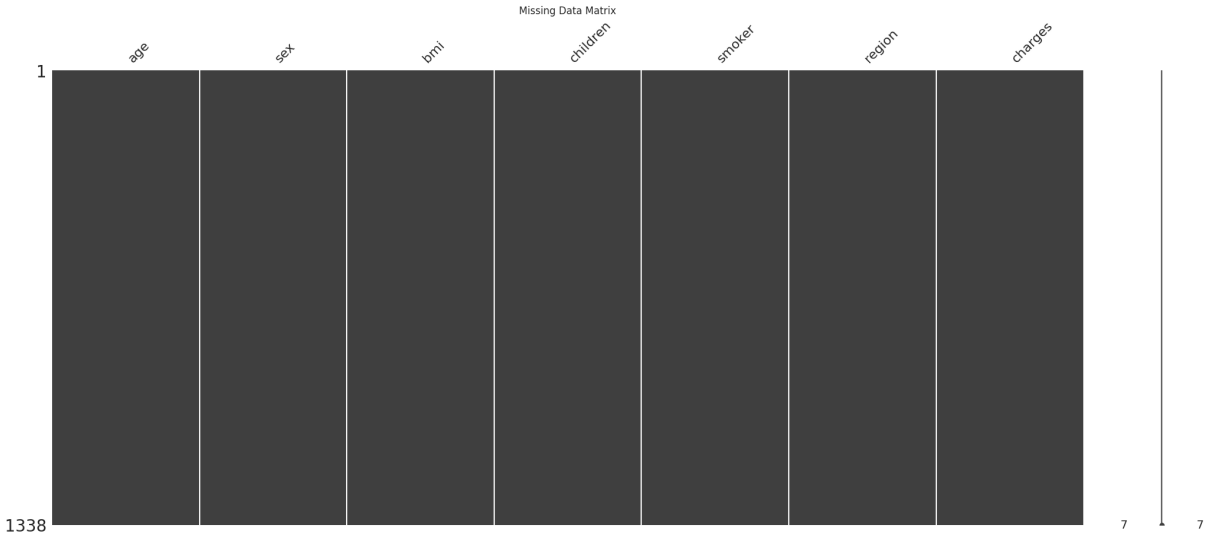


Figure 8: Missing Data Matrix

Outliers Detection

Figure 9 displays the box plots for the numeric variables: age, BMI, and charges. Box plots are useful for visualizing the distribution of the data and identifying potential outliers.

- **Age:** The box plot shows that there are no outliers based on the interquartile range (IQR) method.
- **BMI:** The box plot indicates that there are 9 IQR-based outliers, suggesting some extreme values in the BMI distribution.
- **Charges:** The analysis reveals 139 IQR-based outliers in the charges, indicating a significant number of extreme values in this variable.

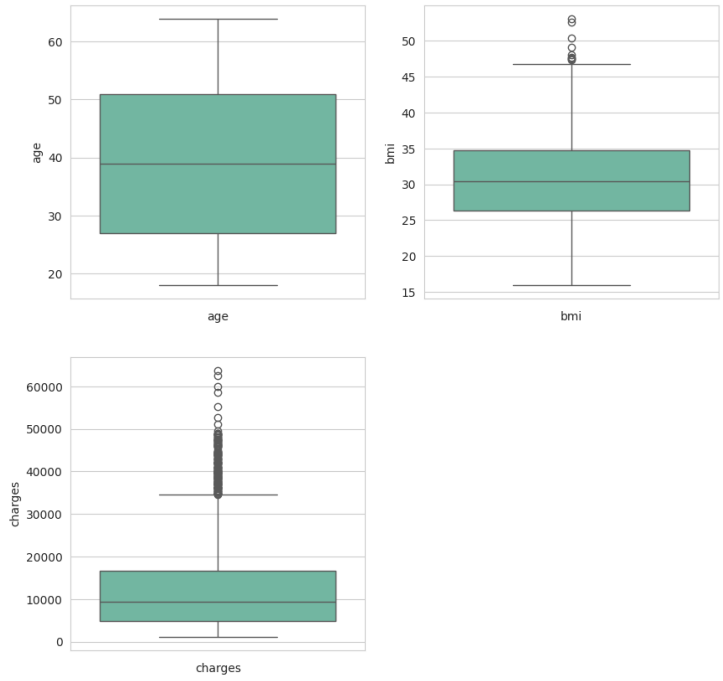


Figure 9: Box Plots for Numeric Variables

3 Model Explanation

3.1 Data Encoding and Preprocessing

The data preprocessing for this model follows a systematic approach to ensure that the input features are in a suitable format for training. The preprocessing steps included:

1. **Data Splitting**:
   - The dataset is divided into features (X) and target variable (y). The target variable in this case is the insurance charges.
   - The dataset is then split into training, validation, and test sets. The training and validation sets constitute 80% of the original data, while the test set represents 20%.
   - Further, the training set is divided into actual training (64% of the original data) and validation (16% of the original data) sets, ensuring robust evaluation.
2. **Encoding Categorical Features**: Categorical features are encoded using one-hot encoding.
3. **Feature Scaling**:
   - Numerical features are scaled using the *StandardScaler*, which standardizes the features by removing the mean and scaling to unit variance.
   - The scaling process is applied separately to the training, validation, and test sets to ensure there is no data leakage. This means that the statistics (mean

and standard deviation) used for scaling are derived solely from the training set, maintaining the integrity of the model evaluation.

- Additionally, the target variable is also scaled using the *StandardScaler* to ensure that it aligns well with the model's outputs.

3.2 Model Design and Implementation

The structure of the first implemented model is described as follows:

- **Input Layer**: The model takes input data with a shape corresponding to the number of features in the dataset.
- **First Hidden Layer**:
  - A fully connected dense layer with 512 units and the `ReLU` (Rectified Linear Unit) activation function is used. The `ReLU` activation helps the network capture complex patterns by introducing non-linearity.
  - A `Dropout` layer with a rate of 0.2 is applied, where 20% of the neurons are randomly set to zero during training to prevent overfitting.
- **Second Hidden Layer**:
  - A dense layer with 256 units and `ReLU` activation is used to further process the learned features.
  - Another `Dropout` layer with a 0.2 rate is applied for regularization.
- **Third Hidden Layer**:
  - A dense layer with 128 units and `ReLU` activation, capturing more abstract representations of the input data.
- **Output Layer**:
  - A dense layer with 1 unit represents the regression output, as this is a continuous target prediction task. No activation function is applied at this stage.
- **Model Compilation**:
  - The model is compiled using the `Adam` optimizer, which adaptively adjusts the learning rate during training.
  - The loss function is `Mean Squared Error (MSE)`, suitable for regression problems.
  - The model's performance is evaluated using `Mean Absolute Error (MAE)`, providing an interpretable measure of the model's prediction error.

Total Parameters: The model consists of 171,009 trainable parameters, distributed across the layers. This parameter count reflects the model's complexity and its ability to capture relationships within the data.

The model was trained over 100 epochs with a batch size of 32, using both the training and validation datasets. Dropout layers are crucial in mitigating overfitting, and the `ReLU` activation functions ensure the network captures non-linear relationships effectively. The results of the first model are detailed in the Results section.

In the following section, we will perform hyperparameter tuning to optimize the model's performance and identify the most effective configuration.

4 Hyperparameter Tuning

To optimize the performance of the model, we explored different configurations of the fully connected layers. Four different layer architectures were tested:

- `[128, 64]`
- `[64, 32, 16]`
- `[256, 128, 64]`
- `[512, 256, 128, 64]`

For each configuration, we trained a model using the same number of epochs (100) and batch size (32), while employing the Adam optimizer. A dropout rate of 20% was applied after each hidden layer to prevent overfitting. The validation mean absolute error (MAE) was used to evaluate the performance of each model.

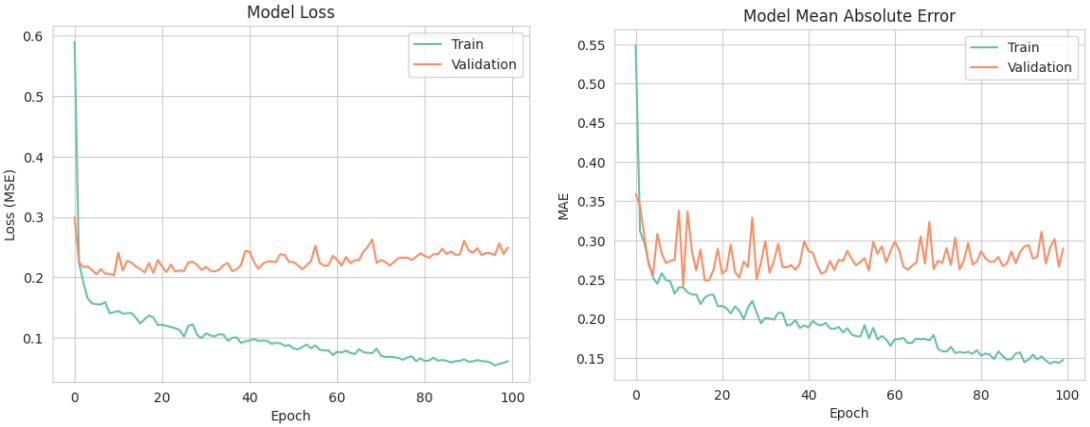
After testing all configurations, the model with the architecture `[64, 32, 16]` achieved the lowest validation MAE of `0.25428396463394165`, making it the optimal architecture.

5 Results

5.1 Model Performance Comparison

Figure 10 illustrates the training and validation loss, as well as the mean absolute error (MAE), for the first model. It can be observed that while the training loss decreases significantly, the validation loss does not follow a similar trend, indicating that the model is overfitting to the training data. This is further supported by the widening gap between the training and validation MAE.

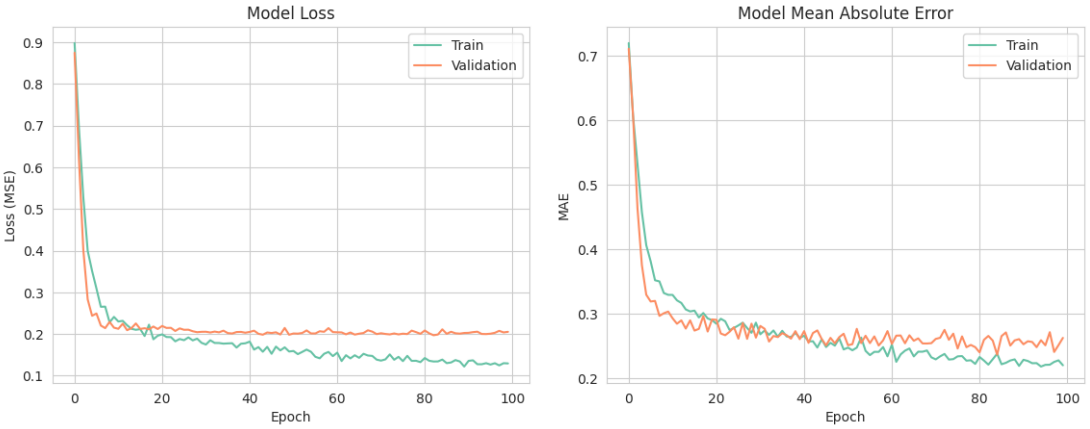
On the other hand, the results of the best-performing model are presented in Figure 11. In this case, both the training and validation loss remain low, and there is no significant divergence between the two, suggesting better generalization. Additionally, the validation MAE closely follows the training MAE, confirming that the best model achieves improved performance with lower error rates and reduced overfitting compared to the initial model.



(a) Training and Validation Loss over Epochs

(b) Training and Validation MAE over Epochs

Figure 10: Results for the First Model



(a) Training and Validation Loss over Epochs

(b) Training and Validation MAE over Epochs

Figure 11: Results for the Best Model

5.2 Residual Analysis and Model Evaluation

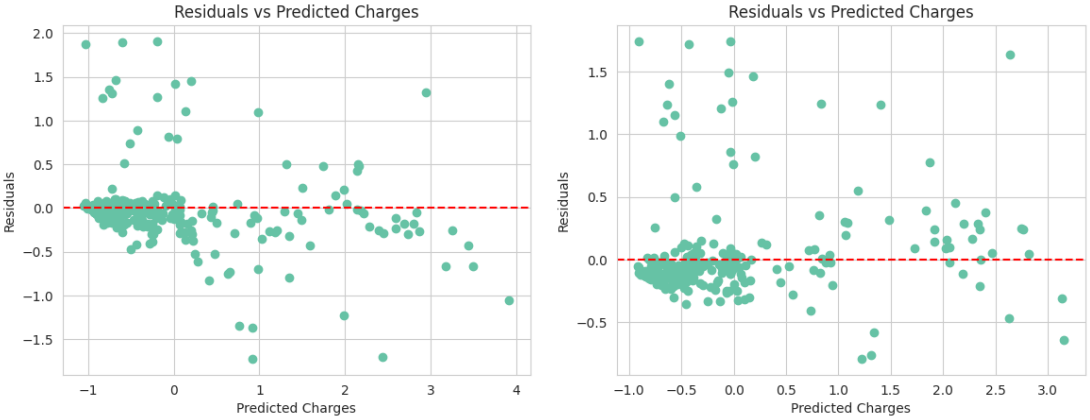
In order to evaluate the performance of both the first and best models, residual plots and error metrics were analyzed.

The residuals for the best model, represented as the difference between the actual and predicted values, are shown in a scatter plot between the predicted charges and the residuals in Figure 12b. Ideally, residuals should be randomly scattered around the horizontal line at zero, which indicates that the model is not systematically over- or under-predicting. As illustrated in the plot, the residuals for the best model are more evenly distributed around zero, suggesting that this model provides relatively unbiased predictions. In contrast, the residuals for the first model reveal a less ideal pattern. The Figure 12a shows greater variation around the horizontal line, indicating that this model has higher predic-

tion errors.

The Mean Absolute Error (MAE) for the best model is `0.2284`, and the Mean Squared Error (MSE) is `0.1549`, both demonstrating improvement over the first model, indicating higher predictive accuracy for this configuration.

The MAE for the first model is `0.2401`, and the MSE is `0.2002`, both of which are higher compared to the best model, confirming that the first model overfits and exhibits weaker predictive performance, leading to larger errors in test data predictions.



(a) Residuals of the First Model

(b) Residuals of the Best Model

Figure 12: Comparison of Residuals for the Best Model and the First Model, highlighting the performance differences in predictive accuracy.

5.3 Conclusion

The residual analysis and error metrics confirm that the best model outperforms the first model, with lower MAE and MSE values and a more random residual distribution, indicating improved generalization and predictive accuracy.