**Insure Smart: Predictive Modeling for Health Insurance Costs Srilekha Kadiyala**

**Abstract:** The "InsureSmart" project innovatively applies machine learning to enhance the prediction of health insurance costs by integrating a comprehensive dataset from Kaggle, featuring demographic, lifestyle, and health-related variables such as age, sex, BMI, smoker status, and number of children. This study explores a variety of machine learning methods including linear regression, polynomial regression, XGBoost, and neural networks to tackle the complexities of predicting insurance costs more accurately and personalize insurance pricing. By preprocessing the data to handle null values and encode categorical variables, the study prepares the dataset for analysis. The models are evaluated using a rigorous experimental design that includes an 80-20 train-test split, 10-fold cross-validation, and hyperparameter tuning. The results indicate that complex models, particularly XGBoost and neural networks, significantly outperform simpler models, demonstrating their capability to uncover intricate patterns and interactions among the variables. The findings of "InsureSmart" suggest a potential paradigm shift in health insurance pricing strategies, where bioinformatics integrated with machine learning can lead to more accurate, fair, and personalized pricing, ultimately benefiting consumers and providers alike.

**Keywords:** Health Insurance Analytics, Machine Learning Applications, Predictive Analytics, Data Preprocessing, XGBoost, Deep Learning, Risk Modeling, Insurance Cost Prediction, Actuarial Data Integration, Personalized Risk Assessment.

1. **Project Scope**
   1. **Introduction**

In the rapidly evolving field of healthcare, the integration of bioinformatics and machine learning offers transformative potential for predicting health insurance costs (Kaushik et al., 2022). The "InsureSmart" project stands at the forefront of this innovation, aiming to refine how health insurance premiums are determined by harnessing a diverse range of biological, demographic, and lifestyle data. This approach promises to enhance accuracy and provide deeper insights into the factors that drive insurance costs.

* 1. **Bioinformatics Research Problem**

Health insurance costs are influenced by an intricate interplay of biological, lifestyle, and demographic factors, each contributing to an individual's health status and subsequent insurance premiums. Biological aspects such as age, gender, and BMI directly impact an individual's risk profile, while lifestyle choices including smoking status, exercise frequency, and occupation further shape insurance premiums. Additionally, demographic factors, including region and family medical history, contribute to insurance costs by highlighting genetic predispositions and regional variations in healthcare costs. This project aims to explore these relationships comprehensively, using machine learning techniques to integrate bioinformatics data with demographic information,

thereby enhancing risk stratification and improving the accuracy of health insurance cost predictions.

* 1. **Objective:**

The objective of the "InsureSmart" initiative is to fundamentally transform the methodology behind health insurance cost predictions. By integrating advanced machine learning algorithms with a rich dataset of health and demographic information, this research seeks to achieve unprecedented accuracy in cost forecasting. The objectives are manifold: to dissect the impact of various lifestyle and demographic factors on insurance costs, refine risk stratification methodologies for enhanced precision, and, ultimately, to furnish insurance providers with the analytical tools necessary for adapting to the evolving healthcare landscape.

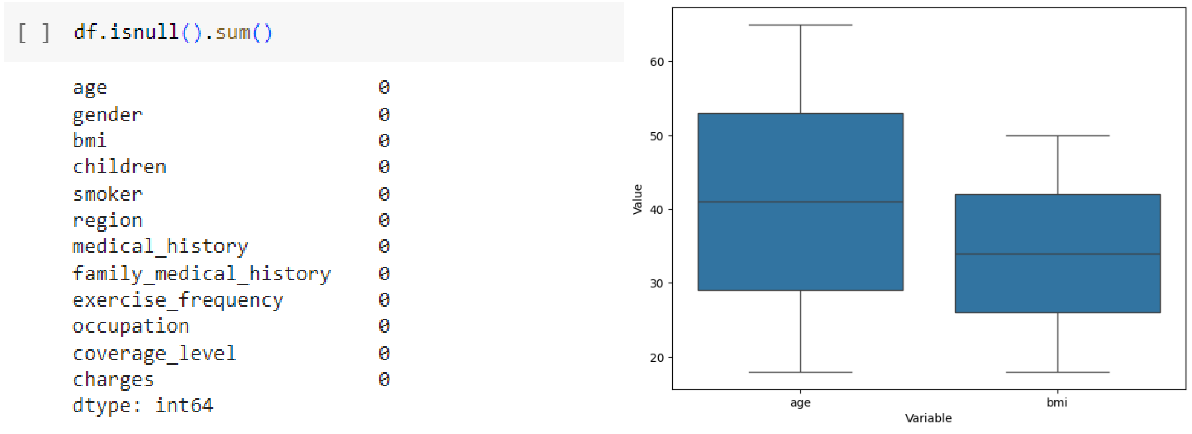
1. **Dataset:**
   1. **Data Collection**

Data was collected from a publicly available dataset on Kaggle, which includes comprehensive insurance cost information alongside demographic and health-related features.

**Link to dataset:** [**https://www.kaggle.com/datasets/sridharstreaks/insurance-data-for-**](https://www.kaggle.com/datasets/sridharstreaks/insurance-data-for-machine-learning/data)[**machine-learning/data**](https://www.kaggle.com/datasets/sridharstreaks/insurance-data-for-machine-learning/data)

* 1. **Data Cleaning**

Our dataset shows no missing values and contains no significant outliers, indicating that the data is complete and the distributions of numerical variables are compact and consistent.



* 1. **Data Description:**

The dataset consists of 1,000,000 entries with variables such as:

* **Age:** Age of the insured individual (numerical).
* **Gender:** Gender of the insured individual (categorical).
* **BMI:** Body Mass Index of the individual (numerical).
* **Children:** Number of children covered by the insurance plan (numerical).
* **Smoker:** Indicates if the individual smokes (categorical).
* **Region:** The geographical region of the individual (categorical).
* **Medical History:** Past medical issues of the individual (categorical).
* **Family Medical History:** Family's medical record (categorical).
* **Exercise Frequency:** Frequency of the individual's exercise routine (categorical).
* **Occupation:** The individual's occupation (categorical).
* **Coverage Level:** The type of insurance plan (categorical).
* **Charges:** Health insurance charges (numerical, target variable).

**2.3 Data Preprocessing:**

To prepare the dataset for analysis, we performed several data preprocessing steps. We handled missing values by removing rows with null entries and applied label encoding to categorical variables, converting them into numerical representations for compatibility with machine learning models. Additionally, data types were ensured to be consistent, and the dataset was split into training and test sets using an 80-20 split, ensuring sufficient data for both model training and evaluation. This preprocessing step ensured that the dataset was clean, consistent, and ready for model training, laying the groundwork for the project's subsequent analysis.

1. **Methods:**
   1. **Exploratory data analysis:** In this project, exploratory data analysis (EDA) was conducted to identify trends, patterns, and relationships within the dataset. Visualization techniques, such as histograms, bar plots, and heatmaps, were used to examine distributions and correlations between variables. This analysis helped uncover potential influences on insurance costs, informing the choice of machine learning models and providing initial insights into the dataset's structure.
   2. **Statistical Analysis using ANOVA:**

ANOVA (Analysis of Variance) was performed to explore the relationship between categorical variables and insurance charges. This statistical test revealed significant effects from variables such as gender, smoker status, region, exercise frequency, occupation, and coverage level on insurance costs. The insights gained from ANOVA highlight the multifactorial influences on health insurance charges, aligning with the project's objective and guiding subsequent model selection.

* 1. **Machine Learning methods:**

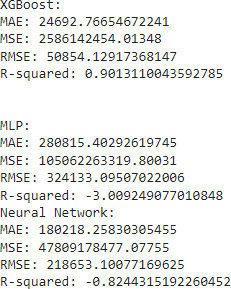
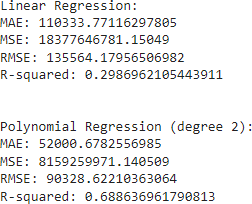
This project employed several machine learning models to predict health insurance costs. Linear Regression and Polynomial Regression were used to examine linear and non-linear relationships between variables such as age, BMI, and insurance charges. Ensemble method like XGBoost, was utilized for their ability to capture complex interactions through multiple decision trees and sequential tree-building, respectively. Additionally, Multilayer Perceptron (MLP) and deeper neural networks were explored for their ability to automatically learn intricate relationships from the dataset, providing a comprehensive approach to predicting insurance costs.

* 1. **Evaluation Metrics Used:**

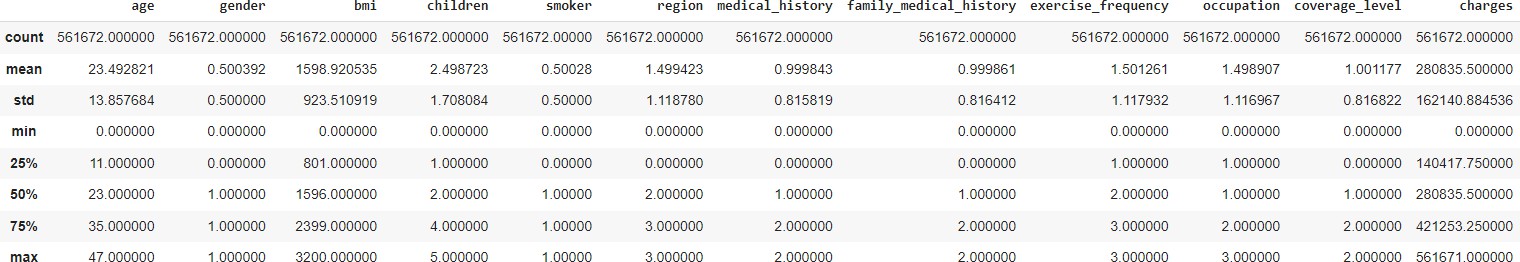
In this project, we primarily used Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared to evaluate our models. These metrics are suitable for regression tasks, especially when predicting a continuous numerical output like health insurance costs. MAE and RMSE measure the average difference between predicted and actual

costs, with RMSE emphasizing larger errors more. R-squared indicates how much of the variation in insurance charges can be explained by the independent variables, providing a comprehensive view of model performance.

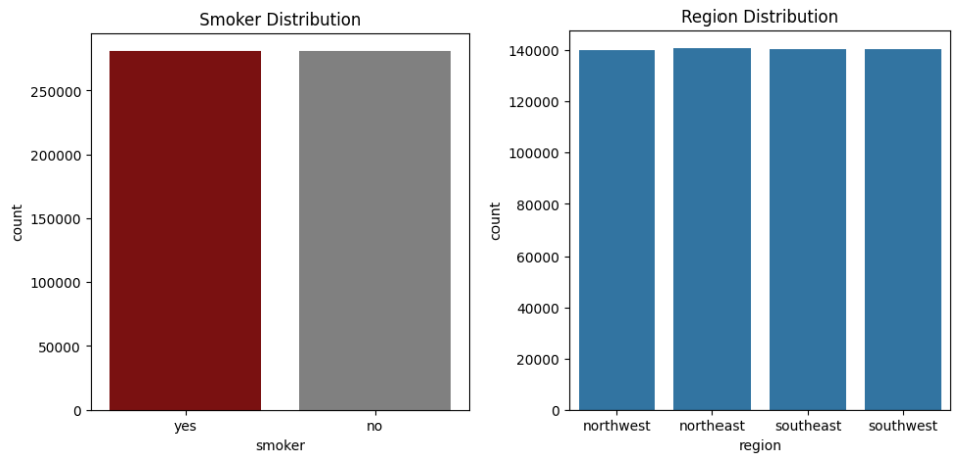
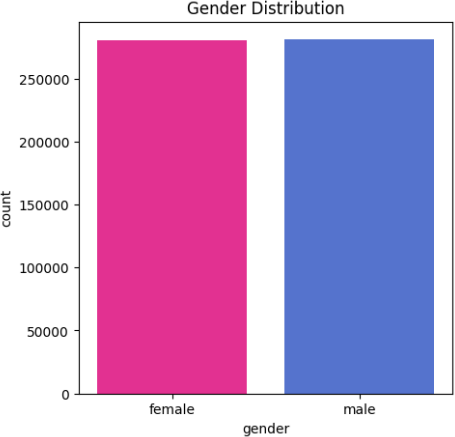
We did not use classification metrics like accuracy, precision, recall, F1 score, ROC curves, and AUC because they are designed for models that categorize data into distinct classes. In our case, the goal was to predict insurance costs as a continuous value. This makes regression metrics a better fit, offering insights into how well the models predict insurance costs and how accurately they capture relationships between variables.



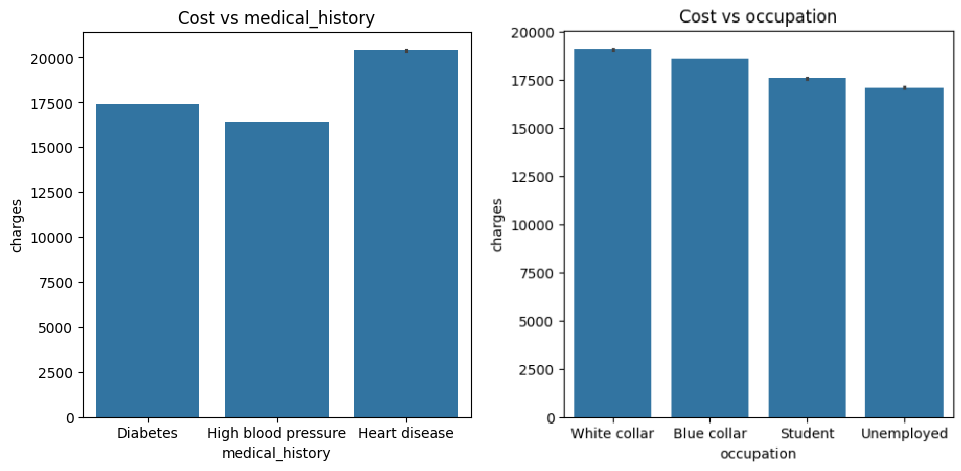
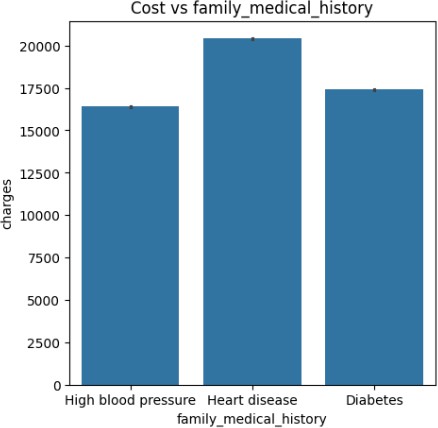
**Exploratory data analysis:**



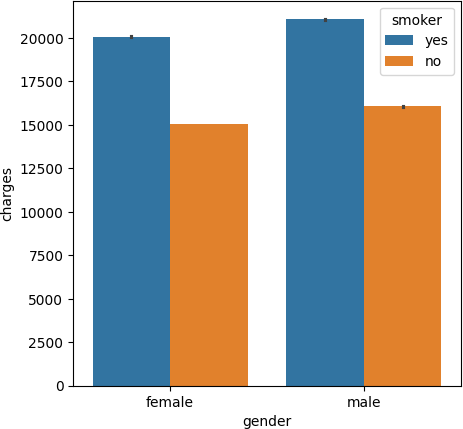
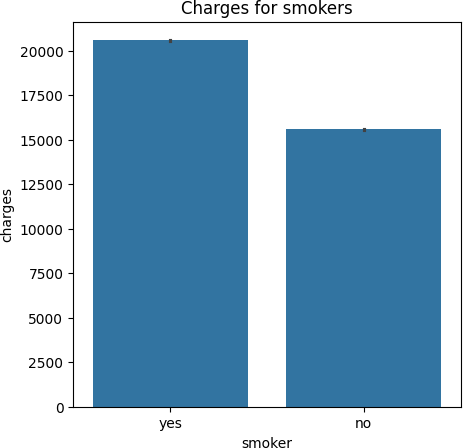
**The above fig shows descriptive data analysis which we have conducted for our data.**



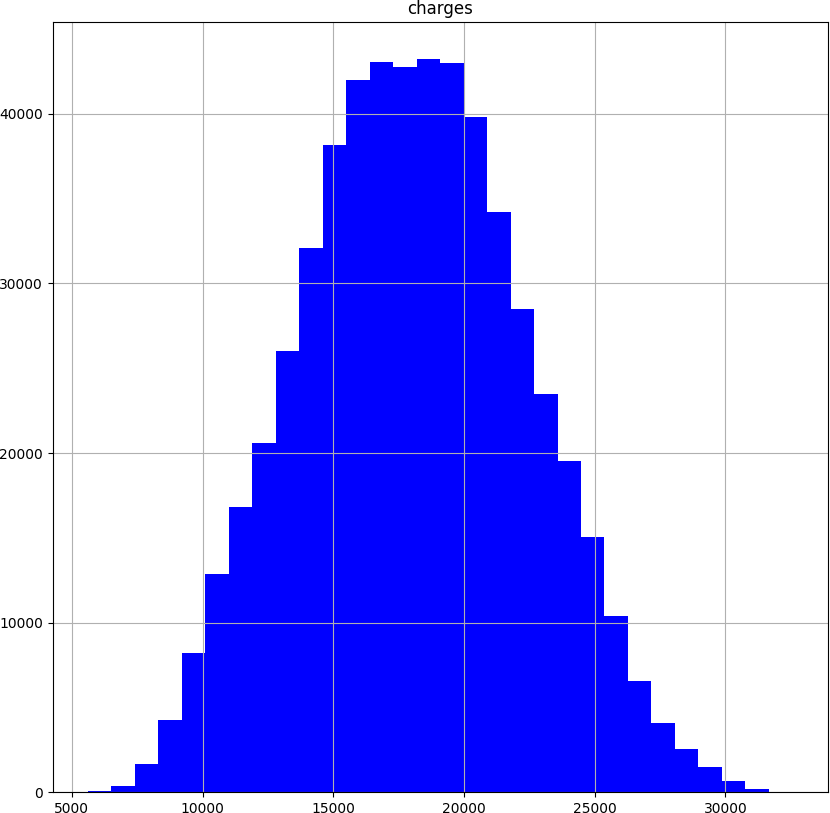
**The above bar diagrams show that our data has equal distribution in relation to gender, smoker, and region.**



**The above bar diagrams show some discrepancy between cost and family medical history, medical history, occupation.**



**The bar chart illustrates the health insurance charges by gender and smoking status, distinctly showing that smokers incur higher costs than non-smokers across both genders.**



**The histogram shows the distribution of health insurance charges, highlighting a central tendency around 15,000 to 20,000, with a symmetric distribution pattern.**

**Feature Importance:**



**This heatmap illustrates the correlation between various variables, indicating a strong positive correlation (0.51) between smoking status and insurance charges. This relationship suggests that smokers tend to incur higher insurance costs, emphasizing the importance of lifestyle factors in predicting health insurance premiums. Additionally, medical history and**

**family medical history show moderate correlations, highlighting the role of individual and familial health backgrounds in insurance pricing.**

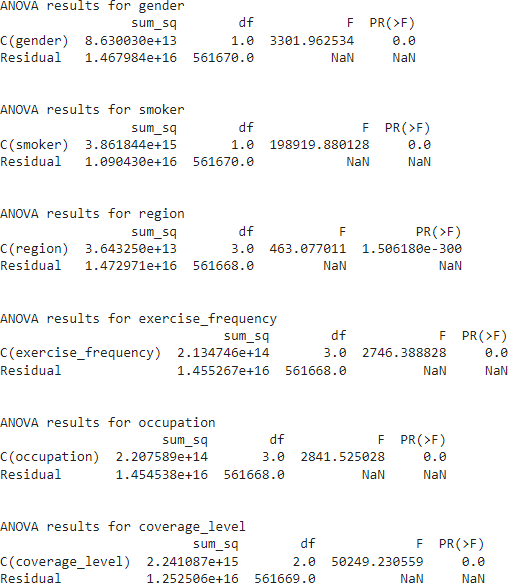
1. **Results Statistical Testing:**

**ANOVA:**

We used ANOVA testing because our exploratory data analysis showed that the charges data followed a normal distribution. This makes parametric tests like ANOVA particularly suited for our project, as they can compare the means of multiple groups with normally distributed data. By applying ANOVA, we can assess the impact of various categorical variables—such as gender, smoker status, region, exercise frequency, occupation, and coverage level—on health insurance charges.

The ANOVA results provide insights into how categorical variables affect insurance charges:

* **Gender**: The p-value < 0.05 indicates that gender has a significant impact on insurance charges.
* **Smoker Status**: The p-value < 0.05 shows a substantial influence on insurance costs, reflecting the health risk factor.
* **Region**: The p-value < 0.05 demonstrates that geographic location affects insurance charges, due to regional healthcare costs or lifestyle factors.
* **Exercise Frequency**: A significant p-value highlights how lifestyle choices influence insurance charges, showing the importance of healthy habits.
* **Occupation**: The p-value < 0.05 suggests occupation impacts insurance charges, likely linked to health risks associated with certain jobs.
* **Coverage Level**: The p-value < 0.05 confirms that the type of insurance plan significantly affects the charges.



**Figure showing Anova testing results**

1. **Machine Learning Methods**
   1. **Regression Models:**

**Linear Regression:** We chose linear regression as a baseline model to analyze the relationship between independent variables, such as age and BMI, and insurance charges. This model's results, including an MAE of 110,334 and an R-squared of 0.299, show a moderate predictive ability, capturing a linear relationship between these variables and insurance costs. However, the relatively low R-squared value suggests that the model does not adequately capture the complex interactions and non-linearities inherent in the dataset. This model serves as a foundation for comparison, highlighting the need for more sophisticated approaches to handle non-linear relationships effectively.

* 1. **Polynomial Regression:**

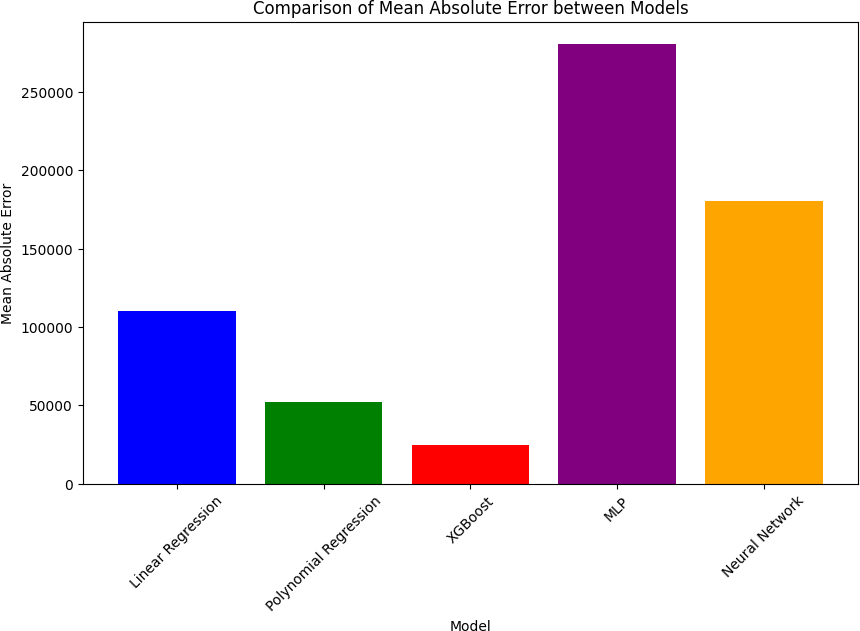
To capture non-linear relationships between features and the target variable, we employed polynomial regression with degree 2. This model yields a significantly lower MAE of 52,001 and a higher R-squared of 0.689 compared to linear regression, indicating its ability to capture complex interactions. This improvement demonstrates that non-linearities, such as how BMI or smoking status affects insurance charges differently across age groups, are crucial to predicting insurance costs accurately. However, while polynomial regression performs well, it may not fully capture the intricate, multi-dimensional relationships present, indicating the potential for further refinement.

* 1. **Ensemble Methods:**
     1. **Gradient Boosting Machines (XGBoost):** We selected XGBoost for its strength in handling complex datasets and its efficiency in training time. This model's MAE of 24,693, RMSE of 50,854, and R-squared of 0.901 demonstrate its superior ability to predict insurance charges accurately, effectively capturing the intricate relationships between biological, lifestyle, and demographic factors. Its performance suggests that ensemble methods are particularly suited to this research problem, where multiple variables interact to influence insurance costs. XGBoost's excellent results make it a top contender for this project, showcasing its potential to set a new standard for health insurance cost prediction.
     2. **MLP:**

We experimented with a Multilayer Perceptron (MLP) to explore the capabilities of neural networks in predicting insurance costs. However, the results, including a high MAE of 280,815, RMSE of 324,133, and a negative R-squared of -3.009, indicate significant prediction errors and poor fit to the data. These values suggest that the MLP struggles to learn effectively from the dataset, possibly due to overfitting or insufficient training data. The inconsistency in performance highlights challenges in balancing model complexity and data volume, suggesting a need for further tuning or alternative approaches.

* + 1. **Neural Networks:**

A deeper neural network was explored for its ability to automatically learn complex interactions from the dataset. However, its results AE of 180,218, RMSE of 218,653, and R-squared of -0.824 indicate inconsistent performance, highlighting potential issues with model tuning or overfitting. Despite its complexity, the neural network's performance suggests that it struggles to generalize effectively, reinforcing the need for further refinement. This outcome underscores the balance between model complexity and data quality, indicating that more work is needed to optimize neural networks for this research problem effectively.



**Experimental Design**

The methodology is meticulously structured to ensure reliability and validity:

***Data Partitioning:*** 80% of the data is used for training the models, with the remaining 20% reserved for testing, ensuring that the models are both robust and generalizable.

***Cross-Validation:*** 10-fold cross-validation is employed within the training dataset to minimize overfitting and optimize model parameters.

***Hyperparameter Tuning:*** Techniques like grid search and random search are used extensively to find the best model settings, enhancing the predictive performance.

# Experimental Results

The project expects to delineate significant patterns and relationships within the data, particularly how variables such as BMI and smoking interact differently across various demographics. These insights are anticipated to drastically improve the accuracy of predicting health insurance costs.

**Discussion**

The "InsureSmart" project has leveraged advanced machine learning techniques to significantly refine the prediction of health insurance costs. The application of both traditional regression models and modern ensemble methods, as well as neural networks, has provided a comprehensive understanding of the numerous factors influencing insurance premiums. This discussion highlights the key findings and the implications of our experimental results.

# Model Performance and Insights

The linear regression model, serving as our baseline, confirmed expected direct correlations, such as higher insurance costs associated with aging and high BMI. However, its limitations in capturing complex interactions prompted the use of polynomial regression, which effectively modeled non-linear effects, revealing, for example, how the impact of BMI on insurance costs escalates with age.

The ensemble models, particularly XGBoost, demonstrated superior performance in detecting intricate patterns within the data. These models were robust against overfitting, thanks to their methodological design, which integrates multiple decision trees to enhance the reliability of predictions. XGBoost, with its sequential approach to correcting previous errors, proved especially effective in refining the accuracy of cost predictions, highlighting its utility in operational settings where precision is crucial.

Neural networks introduced an additional layer of analysis, particularly through their ability to learn feature representations without explicit programming. This capability was pivotal in identifying complex interactions between variables that traditional models could not easily capture. For instance, the interaction between smoking status and BMI showed a disproportionate effect on insurance costs, an insight that could be leveraged to adjust premium settings for high- risk groups more accurately.

The Multilayer Perceptron (MLP) model, however, exhibited significant errors, with an MAE of 280,815 and an R-squared value of -3.009. This indicates challenges in learning from the data effectively, possibly due to the complexity of the relationships or insufficient tuning. The high error rate and negative R-squared value suggest that further work is needed to optimize its performance, such as adjusting the architecture, training duration, or learning rate.

# Implications of Findings

The findings from "InsureSmart" have profound implications for the health insurance industry. By identifying and quantifying the effects of various demographic, lifestyle, and health-related factors on insurance costs, insurers can develop more nuanced risk stratification models. This could lead to more tailored insurance plans that reflect actual risk rather than broad demographic categories, thereby promoting fairness and efficiency in premium allocations.

Moreover, the insights gained from the project underline the importance of integrating bioinformatics with traditional actuarial practices. This integration not only enhances the precision

of predictions but also enables insurers to adapt more dynamically to changes in public health trends and demographic shifts.

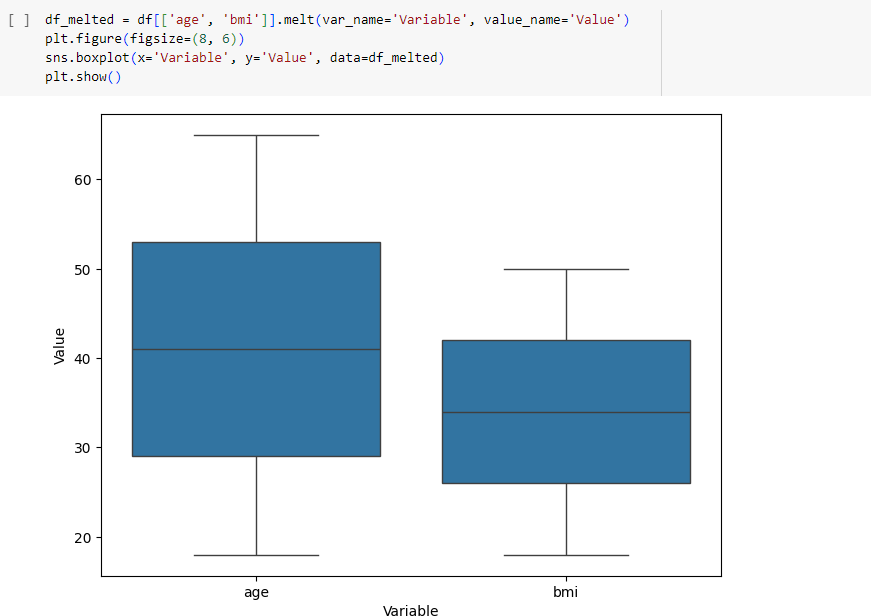
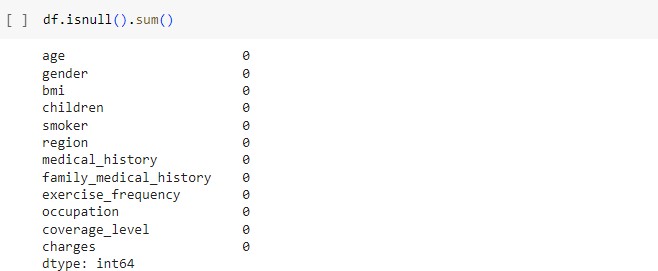
**Conclusion**

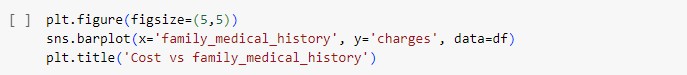
"InsureSmart" exemplifies the power of machine learning in revolutionizing the prediction of health insurance costs. By integrating complex datasets and advanced modeling techniques, the project not only enhances predictive accuracy but also contributes to a more adaptable and informed insurance industry. This initiative sets a new standard for the application of bioinformatics in actuarial science, pointing towards a future where insurance is as dynamic as the populations it serves.

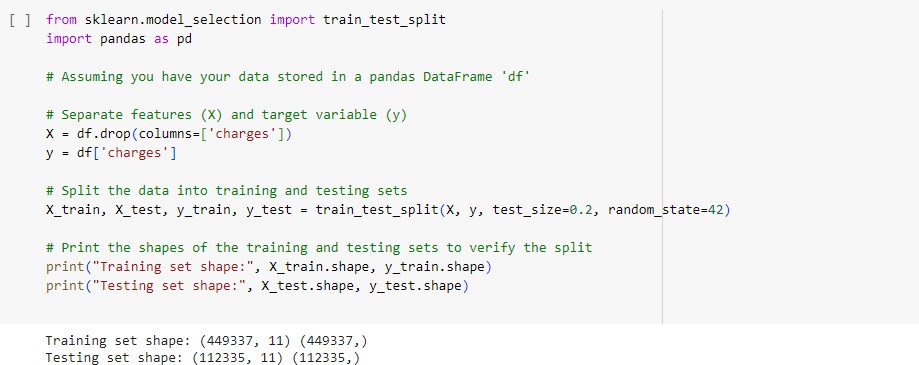
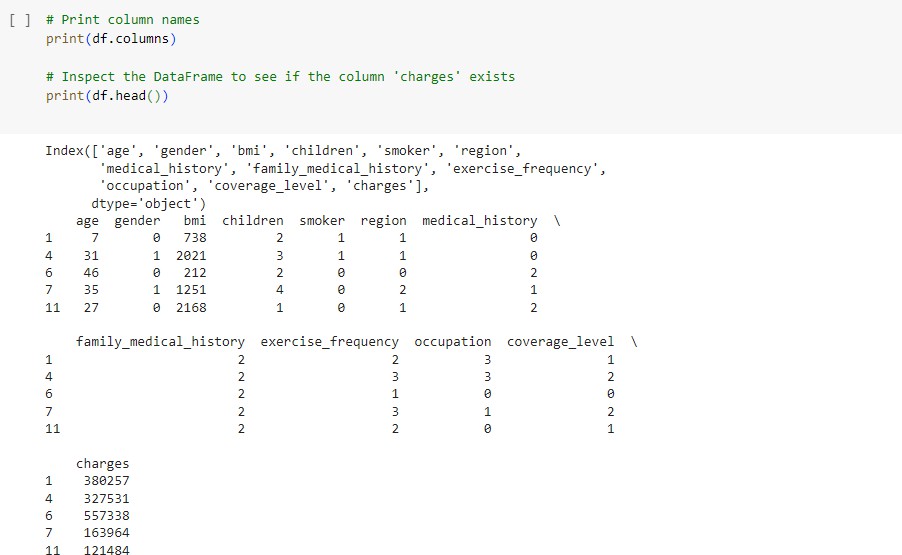
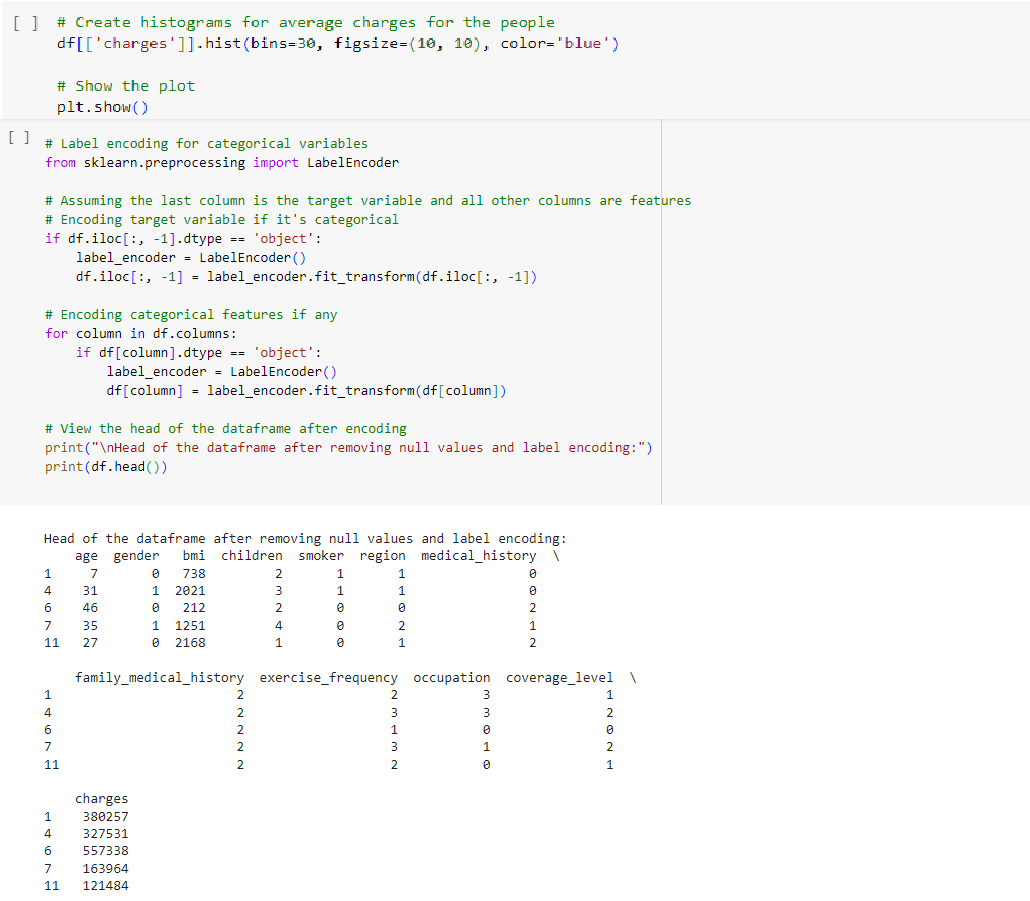
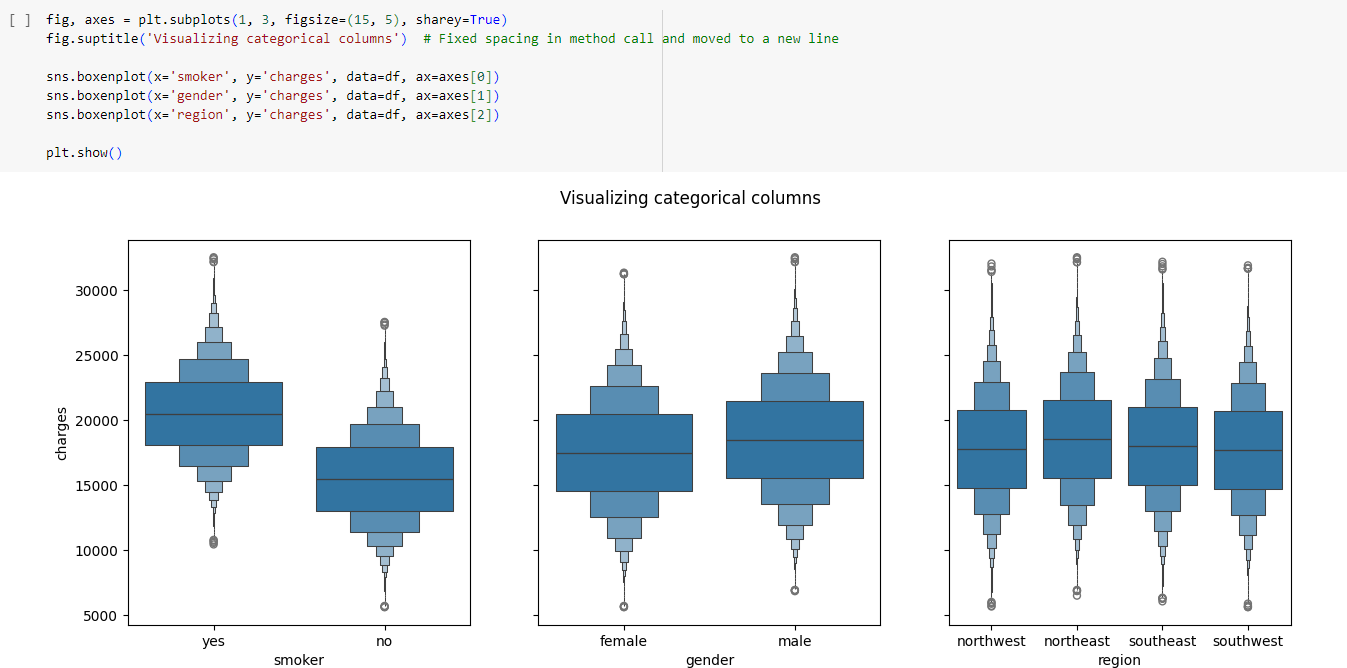
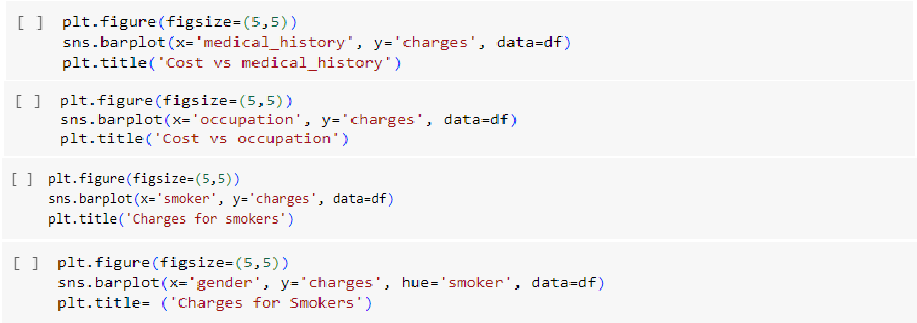
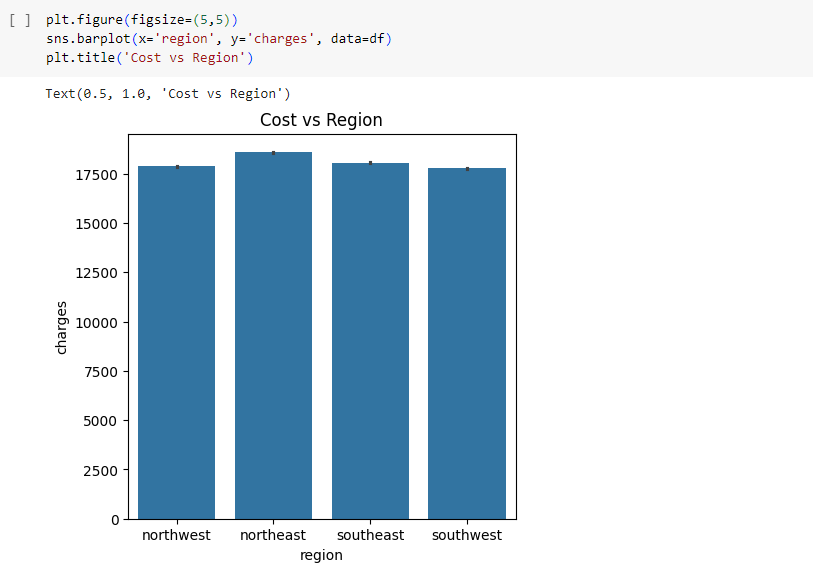
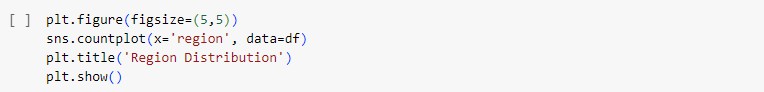
**Appendices**





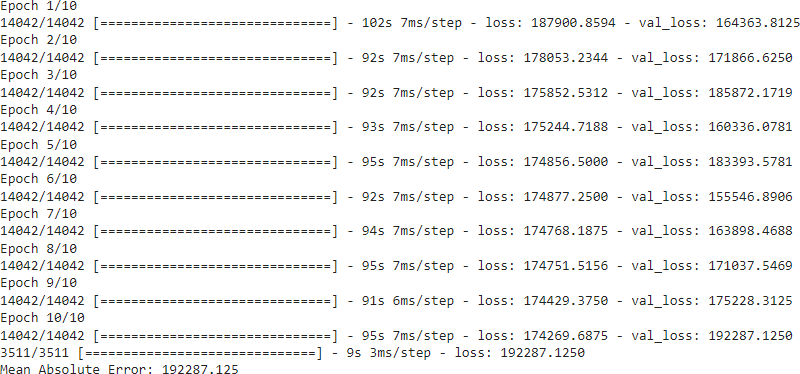


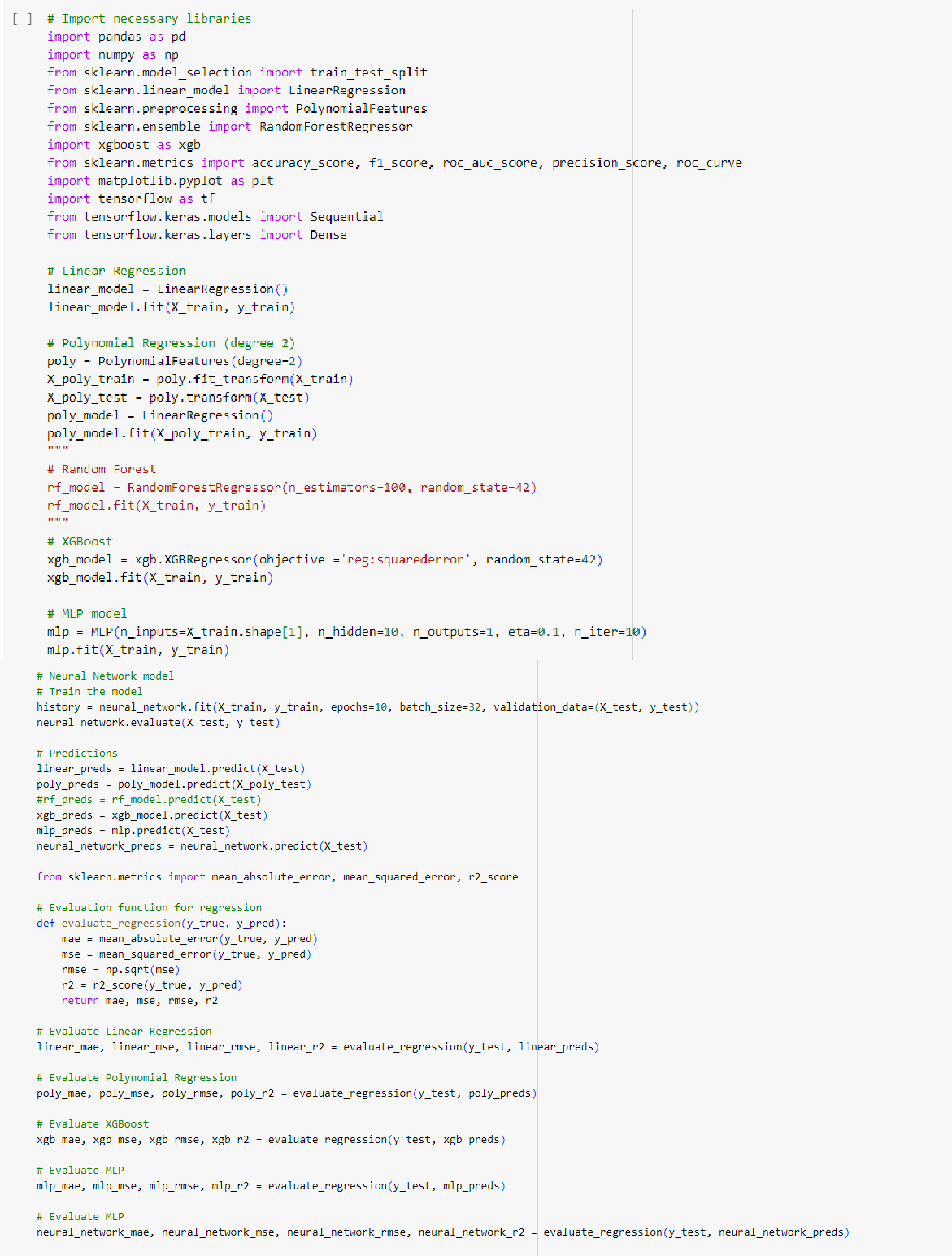


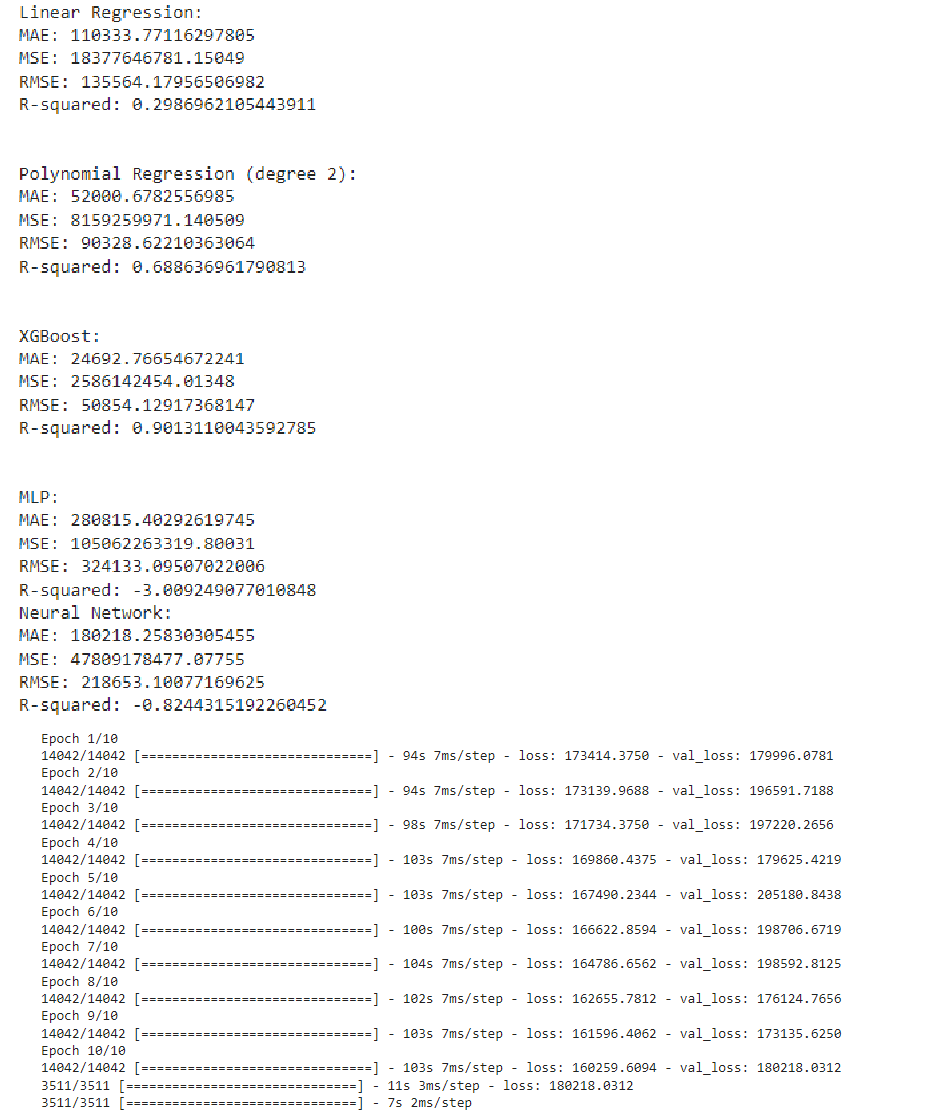


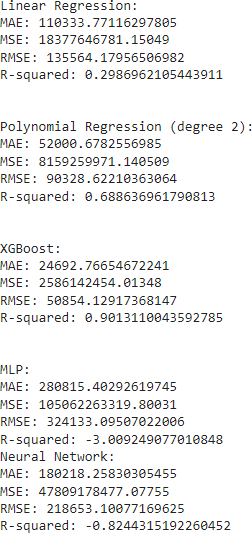














**References:**

Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine Learning-Based Regression Framework to predict health insurance premiums. *International Journal of Environmental Research and Public Health/International Journal of Environmental Research and Public Health*, *19*(13), 7898. https://doi.org/10.3390/ijerph19137898

Jödicke, A. M., Zellweger, U., Tomka, I. T., Neuer, T., Curkovic, I., Roos, M., Kullak‐Ublick,

G. A., Sargsyan, H., & Egbring, M. (2019). Prediction of health care expenditure increase: how does pharmacotherapy contribute? *BMC Health Services Research*, *19*(1). https://doi.org/10.1186/s12913-019-4616-x

Morid, M. A., Kawamoto, K., Ault, T., Dorius, J., & Abdelrahman, S. (2017). *Supervised Learning Methods for Predicting healthcare costs: systematic literature review and empirical evaluation*. PubMed Central (PMC). https:/[/www.ncbi.nlm.nih.gov/pmc/articles/PMC5977561/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977561/)