Medical Insurance Prediction

Introduction/Background

Literature Review

Health insurance costs present financial challenges, highlighting the importance of identifying factors that influence premiums. This project aims to develop a model that predicts medical insurance costs based on demographic and lifestyle. Previous studies have explored this issue, with Tuli et al. using regression models to predict medical expenses based on age, BMI, smoking status, and gender, highlighting AI's role in healthcare prediction [1]. Raza et al. compared ML models like regression, support vector regression, and gradient boosting, finding gradient boosting to be the most accurate for predicting insurance costs [2]. Farhadian et al. used an artificial neural network to analyze how demographic factors and chronic conditions influence medical insurance claims [3]. This study also confirmed the region plays a substantial role in medical premiums. Collectively, these studies provide valuable insights for building accurate prediction models.

Dataset Description

We will use the Medical Insurance Price Prediction dataset [4] from Kaggle, which contains 1338 individual records with 6 features that affect health insurance cost. These features include age, gender (male or female), body mass index (BMI), the number of children/dependents, smoking habits, and geographical region. The dataset includes a “Charges” column, representing each individual’s actual medical insurance premium.

Dataset Link

https://www.kaggle.com/datasets/harishkumardatalab/medical-insurance-price-prediction

Problem Definition

Rising health insurance costs strain individuals and families, making it crucial to understand the factors influencing premiums. While factors like age, gender, and smoking habits are known to impact costs, predicting premiums remains difficult. This project will create a predictive model using demographic and lifestyle data to reveal how factors influence premiums. Accurate predictions could help individuals plan finances and aid insurers in creating fairer, personalized policies.

Methods

Data Preprocessing Methods Used:

1. LabelEncoder: We converted categorical labels like "smoker" and "sex" into binary values (0 and 1), allowing these features to be processed by models that require numerical input. This binary representation is effective for these specific attributes, as they each have only two unique values.
2. KBinsDiscretizer: We converted the number of children column from numbers to different groups (like 0, 1-2, 3+) since insurers are likely to treat large families similarly, reducing the impact of outliers.
3. OneHotEncoder: We used this encoding to expand the “regions” column into multiple binary columns, where each row has a binary vector indicating the specific region. This encoding avoids assuming any inherent order among the regions, which could mislead the model.

ML Algorithms/Models Implemented:

1. Linear Regression (Supervised): We chose Linear Regression as the first model due to the expectation of a linear relationship between the input features (such as age, BMI, smoker status) and the target variable, medical expenses. Linear Regression is an ideal baseline for this scenario because it can be trained quickly due to it involving lighter computations and it captures direct correlations between variables and outputs. The model helps point to features that most strongly impact medical expenses.
2. Random Forest (Supervised): In the context of predicting medical expenses, Random Forest is particularly effective because it can model complex, non-linear relationships between input features such as age, BMI, and smoker status and the target variable which is medical expenses. Random forest also highlights which features affect the predictions the most.
3. K-Means Clustering (Unsupervised): K-means Clustering groups customers with similar characteristics, revealing patterns in medical insurance expenses. Generated clusters can show common characteristics of customers, helping identify outliers in background that would affect the cost of insurance. We also used PCA to select for the top two features that influence variance, to create a good 2D visualization of K-Means.

Results and Discussion

Metrics

Linear Regression Model:

Root Mean Squared Error (RMSE): 6307.012070376673

Mean Absolute Error (MAE): 4152.790049675925

R-squared (R2): 0.7408251682308504

Random Forest Model:

Root Mean Squared Error (RMSE): 2752.215319621075

Mean Absolute Error (MAE): 1271.6996583094422

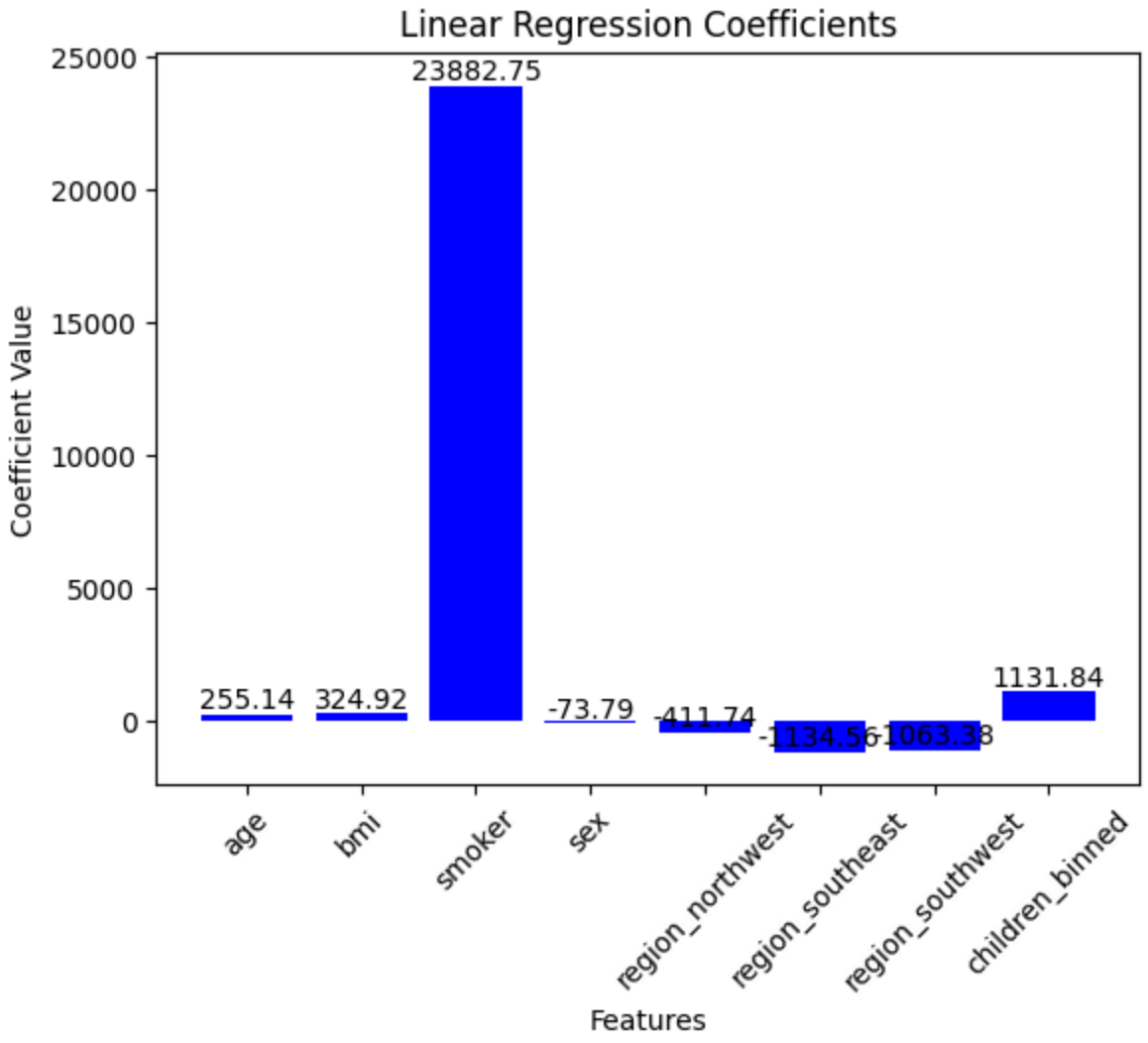
R-squared (R2): 0.9506473682147603

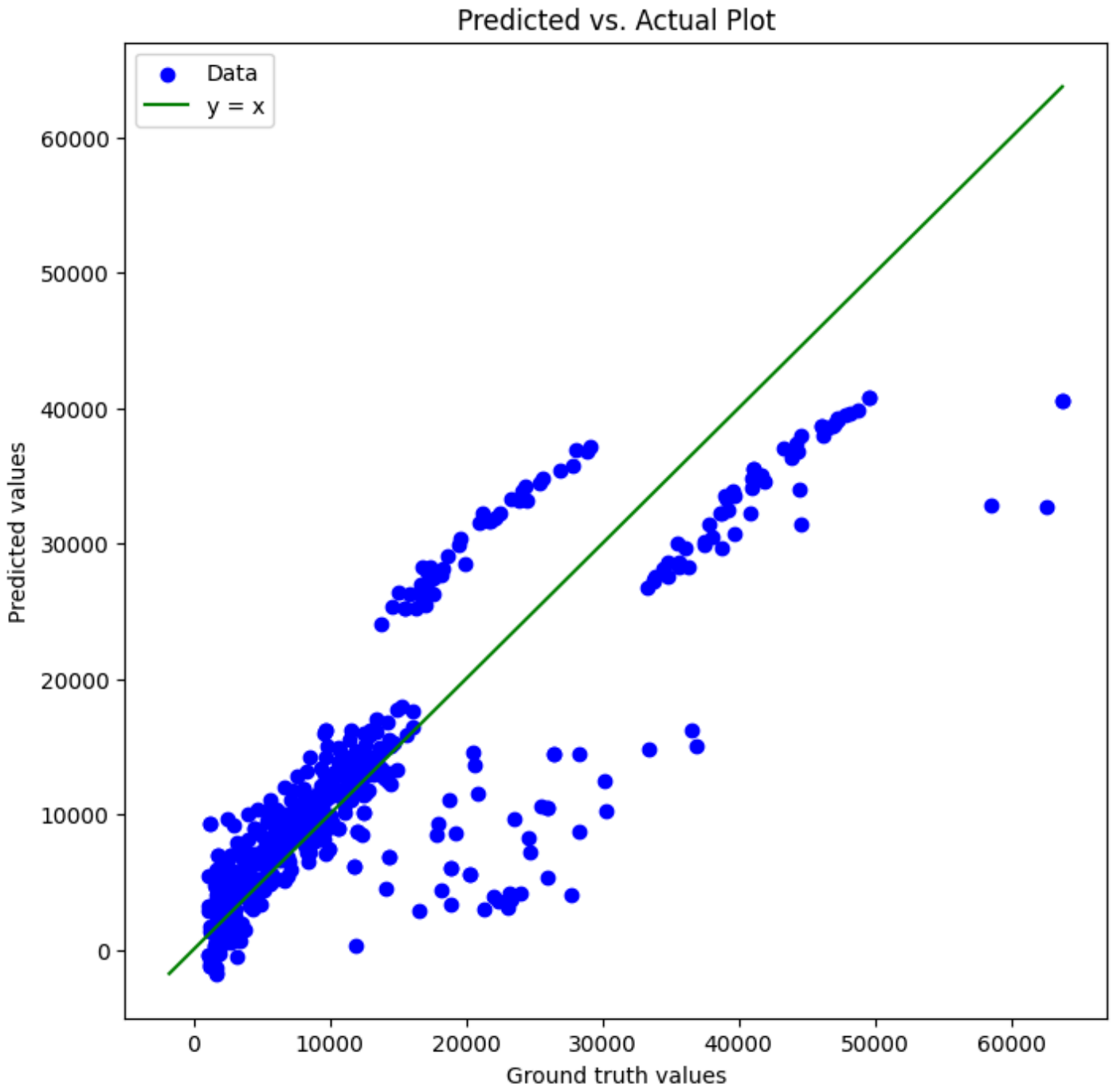
K-Means Clustering Model:

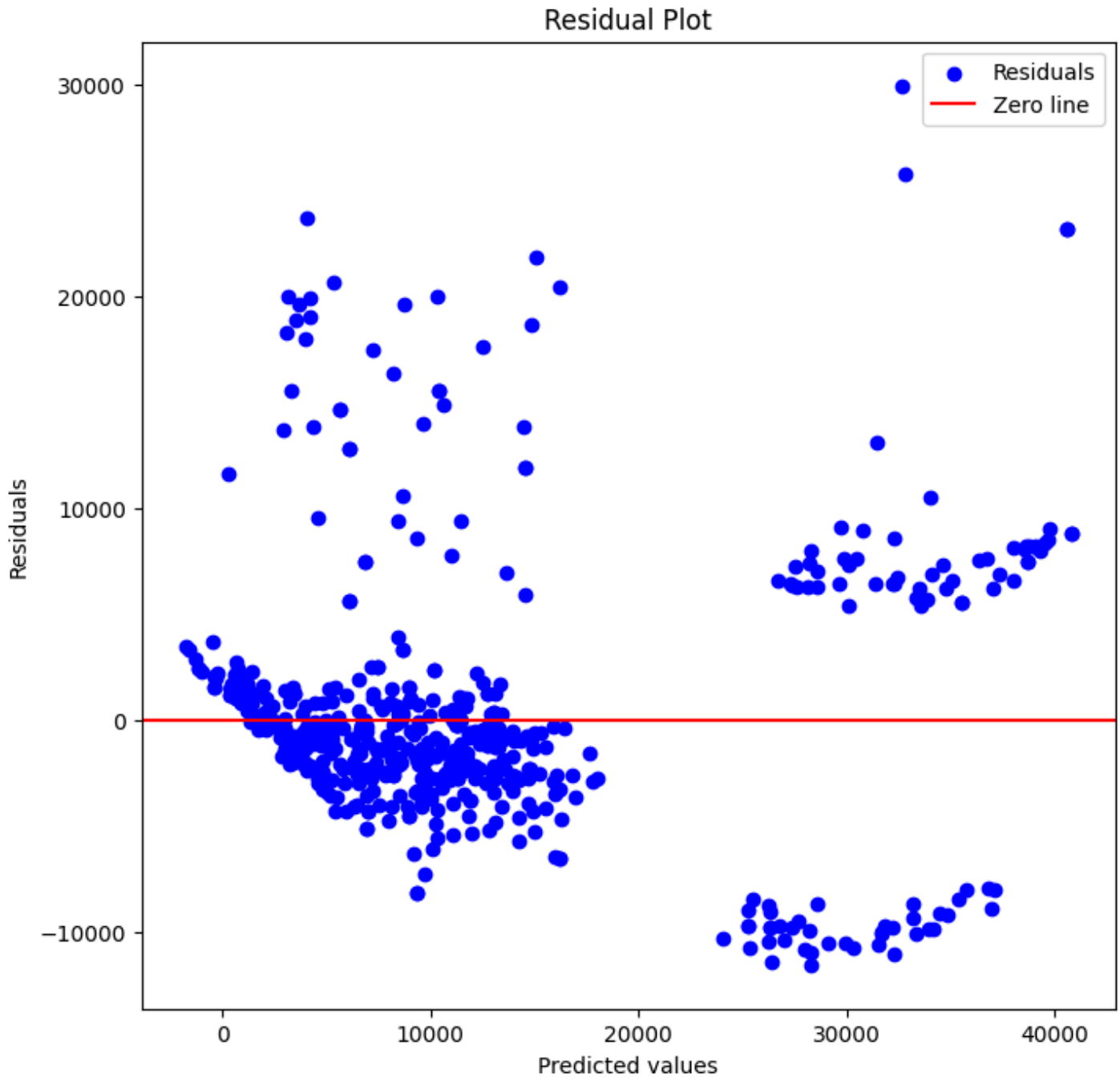
Silhouette Score: 0.3130

Visualizations

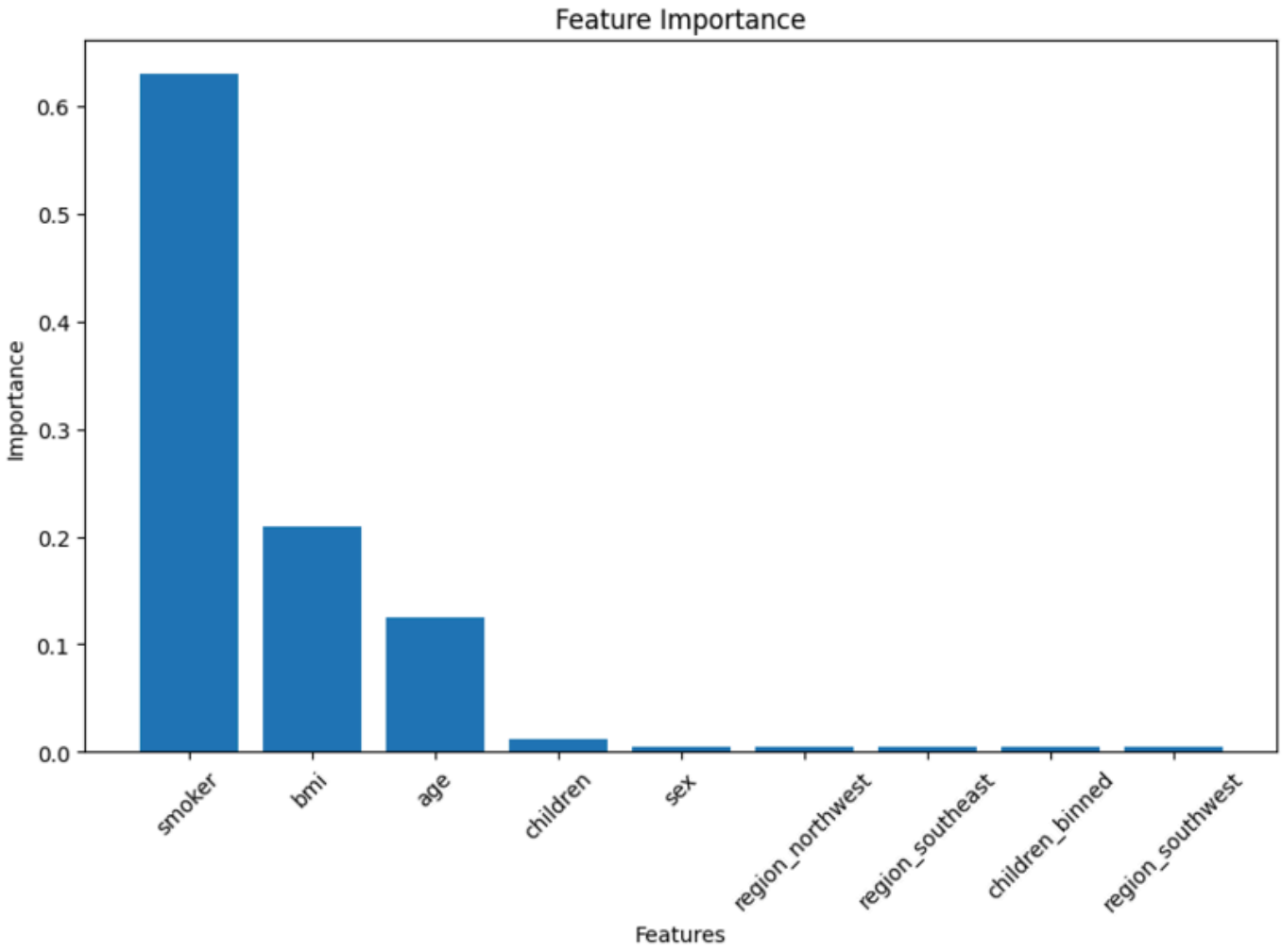
Linear Regression

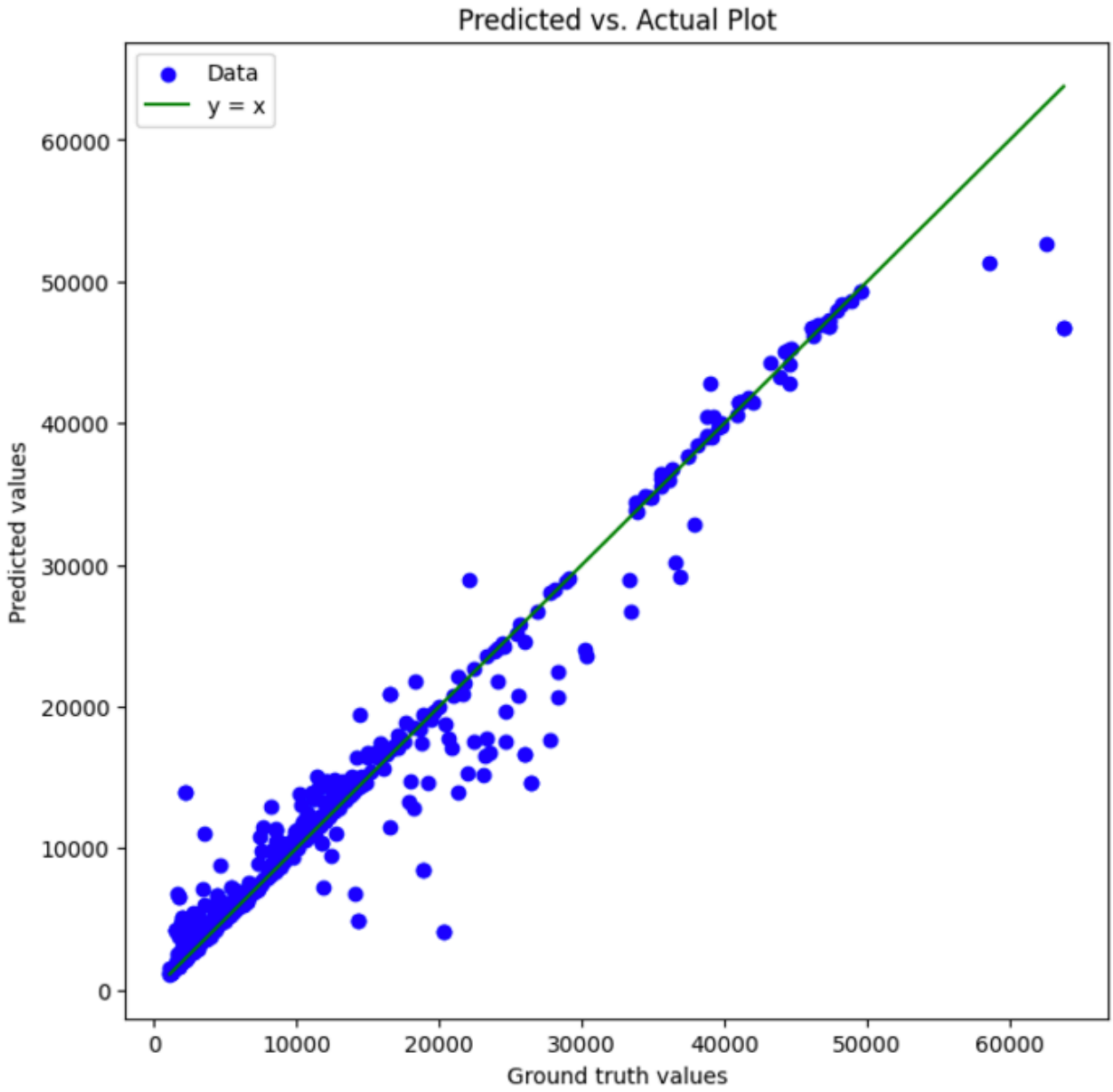


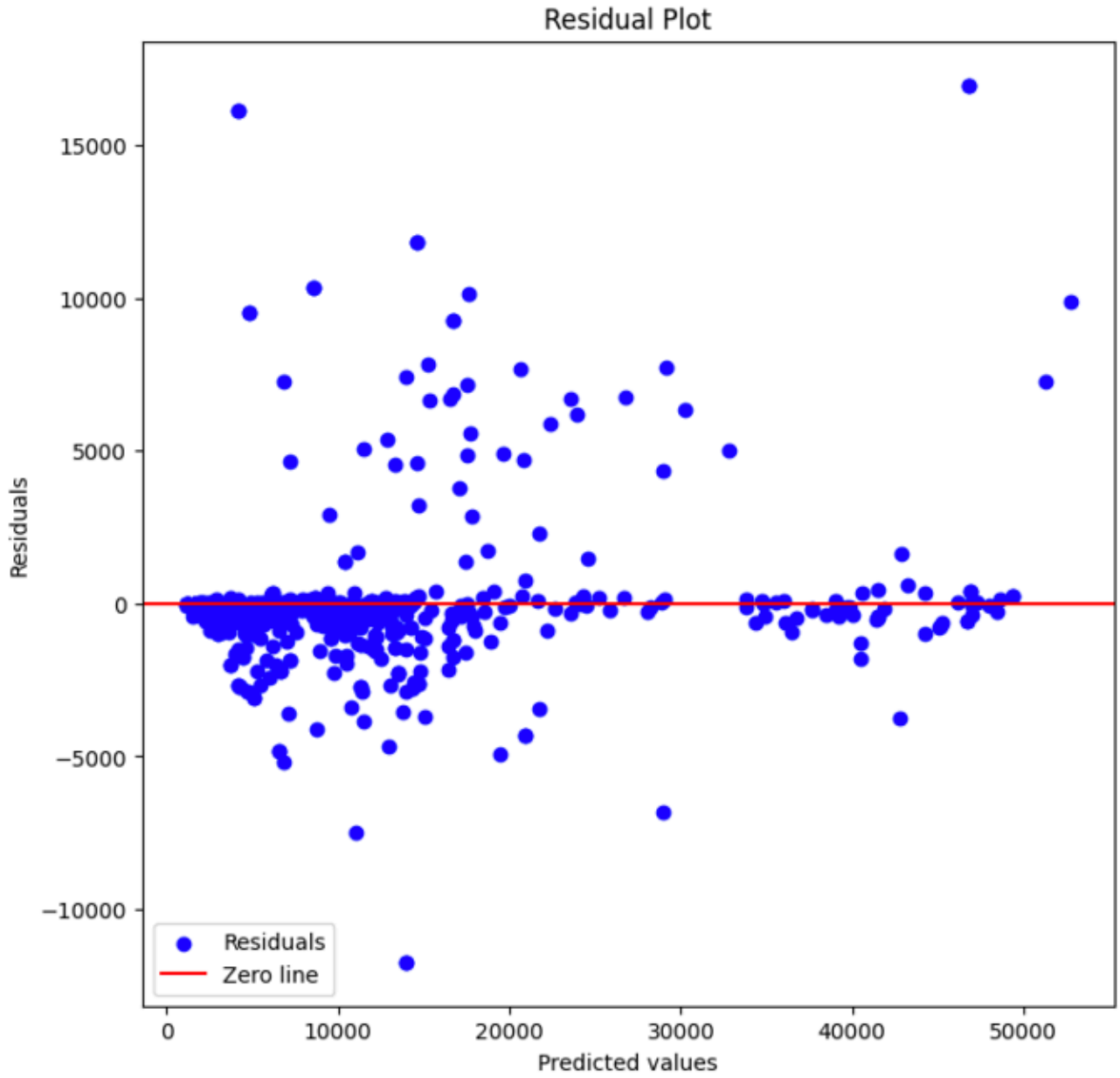

Predicted vs. Actual Plot
Data
y = x
Predicted values
Ground truth values
0
10000
20000
30000
40000
50000
60000



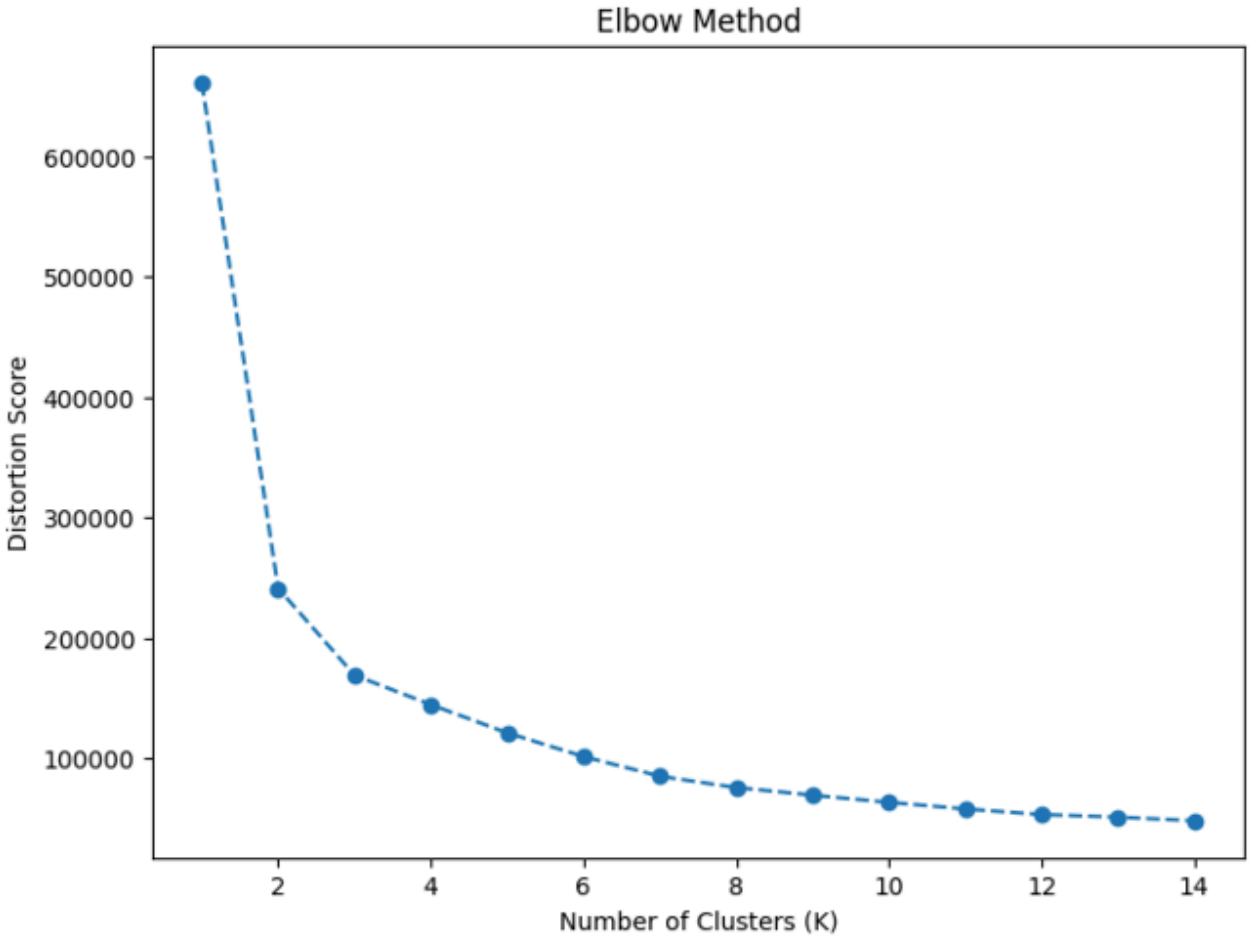
Random Forest

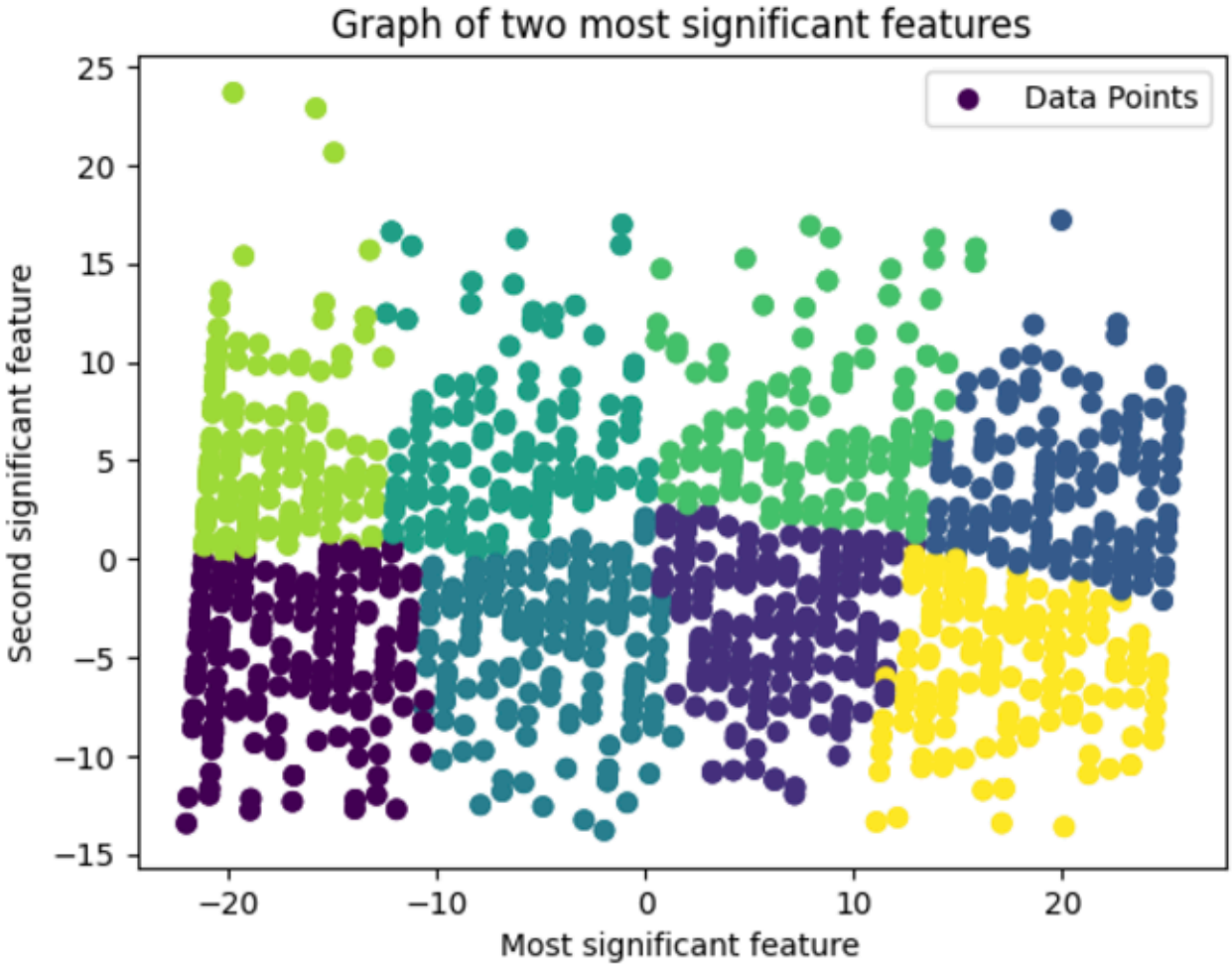

Feature Importance
Importance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
smoker
bmi
age
children
sex
region_northwest
region_southeast
children_binned
region_southwest
Features


Predicted vs. Actual Plot
Data
y = x
Predicted values
Ground truth values
0
10000
20000
30000
40000
50000
60000



K-Means Clustering


Elbow Method
Distortion Score
600000
500000
400000
300000
200000
100000
2
4
6
8
10
12
14
Number of Clusters (K)



Analysis

Analysis of 3+ Algorithms/Models:

Linear Regression (Supervised)

Based on the visualizations and results, our simple linear regression model is performing relatively well on this dataset. The R^2 value of ~0.74 suggests that our model explains around 74% of the variance in this dataset. The unexplained variance could be due to a variety of factors, such as non-linear relationships or other influencing factors perhaps not in this dataset. Looking at the coefficient values of the linear regression terms, we see that the most influential factor is smoker by far, followed by BMI and Age. This makes sense as smoking is injurious to health, making it a higher risk for insurance to cover and increasing insurance cost more than any other feature in the model. Age and BMI are also indicators of health, which is why they are weighted more heavily in the results of the linear regression. The positive coefficients show that these factors contribute to higher insurance charges. Another factor that contributes to higher insurance based on the model is having children. Looking at the Predicted vs. Ground truth visualization, we see that our model fits the test data up for smaller predictions - up to around 15000 dollars. For data points with ground truth y values that are greater than that, we see that our model is less accurate in making those predictions. Looking at

the Residuals vs Predicted values visualization, there are two clusters with a predicted value of around 30000 dollars that shows model bias. This suggests that there are other patterns in the data that our model does not account for. Our Linear Regression model performed well for the most part but it performed poorly when attempting to predict higher insurance costs because there are some patterns in the data that it did not account for.

Random Forest (Supervised)

The Random Forest Model performs better than the Linear Regression Model with a high R^2 value of 0.948. Furthermore, the relatively lower MAE compared to RMSE suggests that although some outliers might be present, the model is handling them decently well. Looking at the feature importance graph, the variable smoker has a contribution of over 60% and is the most important feature. Smoking heavily impacts medical costs followed by BMI and age, which are the next most important features. Features such as children, sex, and religion have a much lower impact on the predictions for medical bills. In the predicted vs actual plot, the points are relatively close to the prediction line, showing how the model performs well and the predictions are generally accurate. The model captures the overall trend of the data well which is shown through the close alignment of the points. In the Residual Plot, most of the residuals are centered around zero, indicating that the model predictions are overall unbiased. However, there are a few large residuals at higher predicted values, showing that the model can be slightly inaccurate with extreme cases or outliers.

K-Means Clustering (Unsupervised)

The K-means Clustering model performed decently well on the dataset with the choice of K = 8 as our value shown by the graph for the elbow method. A silhouette score of 0.312 suggests moderate clustering quality. The clusters are reasonably compact but have some overlap, as shown in the PCA plot, where the blue and green clusters have some intersection in between. The clusters are also pretty close to each other in the PCA plot. Overall, the K-means Clustering model did not perform too well on the medical insurance dataset which is shown by its low silhouette score and the overlapping clusters.

Comparison of 3+ Algorithms/Models

Considering the three models we trained for our medical insurance costs dataset, Random Forest Visualizations performed the best. Random Forests had an R^2 value of 0.948, which was higher than the R^2 values of both Linear Regression and K-means Clustering. The relatively low silhouette score of K-means Clustering suggests that the K-means model did not perform too well on our given dataset. The Linear Regression Model had an R^2 value of 0.74, which was lower than the R^2 value of Random Forests. The predicted vs actual plot for the Linear Regression Model shows more inaccuracy in the predictions for the Linear Regression Models than in the plot for Random Forests. Furthermore, there are more outliers and data points away from the zero line in the residual plot for the Linear Regression Model than in the residual plot for the Random Forests model. Therefore, the

Random Forests model performed the best and had the highest accuracy in predicting medical insurance costs based on our dataset.

Next Steps:

1. To improve clustering, we could try different values of K to see if the silhouette score improves.
2. We can use alternative clustering algorithms such as GMM or DBSCAN for non-spherical clusters.
3. For Linear Regression, we could apply Ridge or Lasso regression to handle multicollinearity or penalize less important features.

References

[1] H. Tuli and M. Tuli, "Predictive Modeling for Healthcare Cost Using Machine Learning Algorithms," International Journal of Environmental Research and Public Health, vol. 19, no. 13, 2022. [Online]. Available: https://www.mdpi.com/1660-4601/19/13/7898

[2] M. Raza, M. A. Sherwani, and M. Tariq, "Factors Affecting Health Insurance Premiums: A Machine Learning Approach," Journal of Healthcare Engineering, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1155/2021/1162553 [3] M. Farhadian, Z. Ma'refi, and S. Yazdani, "A Comparative Study of Health Insurance Models," Semantics Scholar, 2020. [Online]. Available: https://pdfs.semanticscholar.org/57ce/e0d17acfad866230f1e264197af8fc389fa9.pdf [4] H. Kumar, "Medical Insurance Price Prediction," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/harishkumardatalab/medical-insurance-price-prediction.

Contribution Table

Name	Proposal Contributions
Meenakshi Prabhakar	Final Presentation Video
Titiksha Agrawal	Final Presentation Video
Reese Wang	Model 3: KMeans
Vrinda Amarnani	Model comparison & analysis
Subhajit Das	Model 2: Random Forest

Gantt Chart

https://docs.google.com/spreadsheets/d/1z7BWWPnrtcz7wA5Iaf8UppqwwNHxur2i/edit?usp=sharing&ouid=116801007243190460657&rtpof=true&sd=true

Directories

/ml_project.ipynb: Code for the preprocessing methods, model, and visualizations

/lregcoef.png: Linear regression coefficients Graph

/predvact.png: Linear Regression Predicted vs. Actual Graph

/residual.png: Linear Regression Residual Plot

/rf_feat_imp.png: Random Forest Feature Importance Plot

/rf_pred_acc.png: Random Forest Predicted vs. Actual Plot

/rf_residual.png: Random Forest Residual Plot

/km_elbow.png: K-Means Elbow Method

/km_clusters.png: K-Means Most Significant Features