# Machine Learning–Driven Healthcare Cost Prediction, Risk Segmentation, and Intervention Analysis

Adhish Nanda
*MSc.Data Science*
*University of Europe for Applied Sciences*
Potsdam, Germany
adhish.nanda@gmail.com

***Abstract*—This project is on healthcare cost prediction using machine learning to identify cost-driving factors and simulate interventions for cost reductions. We performed data preprocessing, exploratory data analysis, feature engineering, predictive modeling, clustering, and anomaly detection on a dataset of patient demographic information, health indicators, and insurance charges. The results found out that smoking and BMI have a very important effect on healthcare costs; thus, interventions targeting these areas can save a lot. This paper describes the methodology, and the results obtained in using advanced data analytical techniques.**

***Index Terms*—healthcare cost prediction, risk segmentation, prescriptive analytics, anomaly detection**

## I. Introduction

Health care cost is one of the top concerns for individuals, insurers, and policymakers. Healthcare costs have been increasing globally due to various factors such as aging populations, lifestyle diseases, and rising medical expenses.

Increasing costs of health care give rise to a great need for new solutions to control these expenses. Understanding the drivers behind such costs will help make better decisions and resource allocation. Predictive analytics has the power to do that and respond to such challenges.

Machine learning has become a fundamental tool for predictive modelling and decision-making in modern data-driven systems [1]. Python and libraries such as scikit-learn provide efficient implementations of machine learning algorithms [2]. Detecting anomalies is also important in identifying unusual or high-risk cases in healthcare data [3].

The objective of this project is to leverage data analytics techniques to:

1) Develop predictive models to estimate healthcare costs.
2) Explore advanced segmentation techniques for personalized insights into patient groups.
3) Detect anomalies to highlight unusual cost patterns and address potential outliers.
4) Simulate behavioral changes, such as smoking cessation and BMI reduction, to estimate potential cost savings.

The project will seek to provide actionable insights by integrating machine learning methodologies with prescriptive analytics and anomaly detection techniques, which can help in the formulation of policies and optimization of costs in the healthcare sector.

## II. Related Work

Healthcare cost prediction has been a critical area of research, given its implications for both individual and systemic cost management. Several studies have explored the factors influencing healthcare expenditure and the techniques used for predictive analysis:

### A. Health Care Cost Drivers

The program is now amalgamating the machine learning methodology with prescriptive analytics and anomaly detection techniques to offer implementable insights, guiding the creation of policies and the optimization of costs within the health sector.

### B. Predictive Modeling

Traditional statistical methods, such as Linear Regression, have been widely applied in the forecasting of health care costs. Machine learning algorithms, especially Random Forest and XGBoost, have gained more effectiveness by being able to model complex, non-linear relationships.

### C. Clustering Techniques

Evidence has shown that clustering algorithms, such as K-Means, are effective in the segregation of patients into meaningful groupings. These clusters give important insights for tailored interventions and the dispensing of resources.

### D. Anomaly Detection

Techniques like Isolation Forest have proven to be effective in outlier detection in healthcare datasets. These anomalies often uncover fraudulent claims, rare conditions, or errors in data entry; all these are crucial to the integrity of the data and to control costs.

This project builds upon these foundational studies to deliver a comprehensive and impactful analysis of healthcare costs using state-of-the-art data analytics methodologies.

## III. Methodology

### A. Data Pre-Processing

Data preprocessing is a critical first step to ensure the quality and usability of the dataset. The following tasks were undertaken:

1) Loading and Exploration: The data set was loaded into the analytics environment, and its structure was examined. Preliminary checks confirmed there were no missing values in key columns.
2) Categorical Encoding: The variables—sex, smoker, and region—had to be encoded in a one-hot way to bring the data into a format suitable for machine learning algorithms.
3) Feature Engineering: New features, like age group and bmi_category, have been created to capture non-linear trends and improve the model performance.
4) Scaling: Numerical features, such as age and BMI, were normalized to be compatible with algorithms sensitive to the magnitude of the features.

### B. Exploratory Data Analysis

EDA provided valuable insights into the dataset:

- Correlation Analysis: Strong connections between health care costs-charges-and the following variables were depicted with a heatmap: smoker_yes and BMI.
- Visualization:
  - Scatter plots showed relationships between charges and continuous features, such as BMI and age.
  - Box plots also showed differences in categories where smokers had a higher charge than non-smokers.
  - Heatmaps highlighted how variables interact with one another, providing a full view of the dataset.

### C. Predictive Modeling

Three models were developed and evaluated to predict healthcare costs. Figure 1 presents the detailed machine learning pipeline, including preprocessing, model training, evaluation, and downstream analytics such as segmentation, anomaly detection, and prescriptive simulation:

- Linear Regression: Regression models form the foundation of predictive analytics and are widely used in statistical learning [1]. This was the simple base model, getting an $R^2$ of 0.78 and a Mean Absolute Error (MAE) of $4,305. It was helpful for comparison but was struggling with capturing non-linear patterns.
- Random Forest: Random Forest is an ensemble learning method that builds multiple decision trees. Random Forest improves predictive performance and reduces overfitting by combining multiple decision trees [4]. This ensemble model performed better than Linear Regression, with an $R^2$ of 0.86 and an MAE of $2,559. In this respect, it was the best of all models, mostly owing to its aptitude for dealing with non-linear relationships.
- XGBoost: XGBoost is an advanced gradient boosting algorithm. It is known for its scalability and superior performance in structured data prediction tasks [5]. Competitive results with an $R^2$ of 0.84 and MAE of $2,775 were provided. Its robustness and interpretability made it a viable alternative to Random Forest.
- Advanced Segmentation: It was performed by applying K-Means clustering to the dataset to uncover individualized insights. K-Means is one of the most widely used unsupervised learning algorithms for clustering and segmentation [6]:
  - Cluster Characteristics:
    - Cluster 0: This cluster consisted of smokers and had much higher average charges ($32,050).
    - Cluster 1: This cluster had non-smokers and lower average charges ($8,434).
  - Recommendations:
    - For Cluster 0, targeted smoking cessation programs were suggested to mitigate high costs.
    - For Cluster 1, preventive care initiatives were recommended to maintain low costs.
- Prescriptive Analytics: Prescriptive analytics simulated the financial impact of behavioral changes:
  - Smoking Cessation: This was estimated to save the individual an average of $6,456, further underlying the economic and health benefits of quitting smoking.
  - BMI Reduction: Simulated 5% reduction in BMI led to average savings of $9,678 per person.
- Anomaly Detection: Anomalies in the dataset were identified and analyzed. Isolation Forest was used to detect anomalies. Isolation Forest is an effective anomaly detection technique based on isolating outliers instead of profiling normal data points [3]:
  - Detection: Isolation Forest detected 21 anomalies with unusually large differences between actual and predicted charges.
  - Visualization:
    - Scatter plots brought these anomalies into light, showing cases where actual charges were unusually high or low.
    - Bar plots detailed the charge differences for each anomaly, enabling targeted investigations.

### D. Dataset

This study uses the *Medical Cost Personal Dataset* (often distributed as `insurance.csv`), a widely used benchmark dataset for medical insurance charge prediction. The dataset contains 1,338 records with seven variables: *age*, *sex*, *bmi*, *children*, *smoker*, *region*, and the target variable *charges* (annual medical costs billed). The features capture a mix of demographic attributes (age, sex, region), lifestyle/health proxy (bmi, smoker), and family context (children). This dataset is publicly available on Kaggle and is frequently used

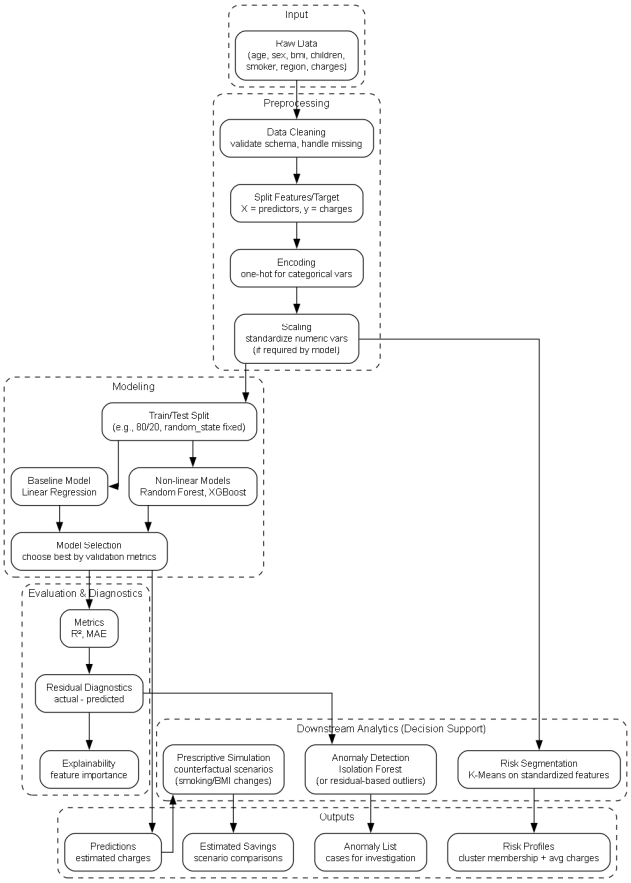


Fig. 1. Machine learning pipeline for healthcare cost prediction and downstream decision support.

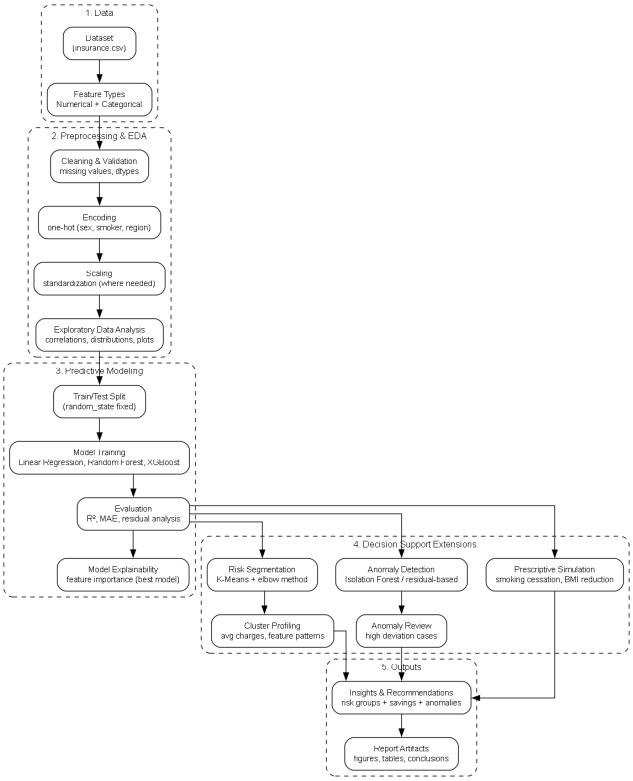


Fig. 2. Overall workflow of the healthcare cost analytics system.

to evaluate regression models and interpretable cost drivers in health insurance analytics [7].

### *E. Overall Workflow*

Figure 2 summarizes the end-to-end workflow. First, the dataset is ingested and cleaned, followed by encoding categorical variables and scaling numerical attributes. Exploratory analysis is then performed to understand distributional properties and cost drivers. Next, multiple regression models are trained and evaluated to predict *charges*. After predictive modeling, the study extends beyond prediction through: (i) risk segmentation via K-Means clustering to identify high- and low-cost groups, (ii) prescriptive simulations to estimate savings under behavioral changes such as smoking cessation and BMI reduction, and (iii) anomaly detection (Isolation Forest) to identify unusual cost patterns. Finally, results are translated into actionable insights and recommendations.

### *F. Experimental Settings*

All experiments were conducted using a reproducible train/test split and consistent preprocessing across models. Categorical variables (*sex*, *smoker*, *region*) were one-hot encoded, and numerical variables were scaled using standard normalization where required. The predictive task is supervised regression with *charges* as the target. Model performance was evaluated using $R^2$ and Mean Absolute Error (MAE). For segmentation, K-Means clustering was trained on the standardized feature space and the number of clusters was chosen using the elbow method. For anomaly detection, Isolation Forest was applied to identify rare cost patterns based on prediction residual behavior (difference between actual and predicted charges). Random seeds were fixed (e.g., `random_state=42`) to ensure replicability.

## IV. Results

### *A. Correlation Heatmap*

- Description: The correlation heatmap shown in Fig. 3 illustrates the relationships between features. It displays the correlation coefficients between numerical features in the dataset.
- Significance: Helps identify relationships between features, particularly between charges and other variables like smoker_yes and bmi.
- Insights:
  - Strong positive correlation between charges and smoker_yes
  - Moderate positive correlation between charges and bmi.
  - Low or no correlation with other features like region.

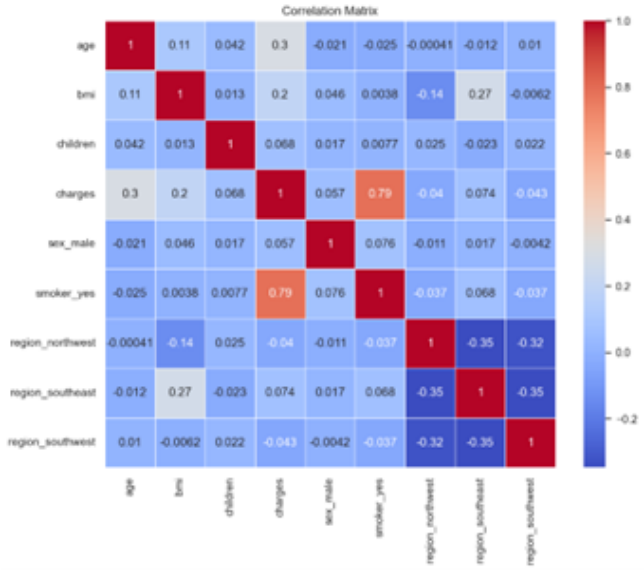


Fig. 3. Correlation heatmap of numerical features, highlighting strong positive association between healthcare charges and smoking status, and a moderate association with BMI.

### B. Distribution of Healthcare Charges

- Description: A histogram displayed with Fig. 4 shows the distribution of healthcare charges.
- Significance: Highlights the range and skewness of charges in the dataset.
- Insights:
  - Highly skewed towards lower charges with a few outliers on the higher end.
  - Indicates the presence of high-cost anomalies.

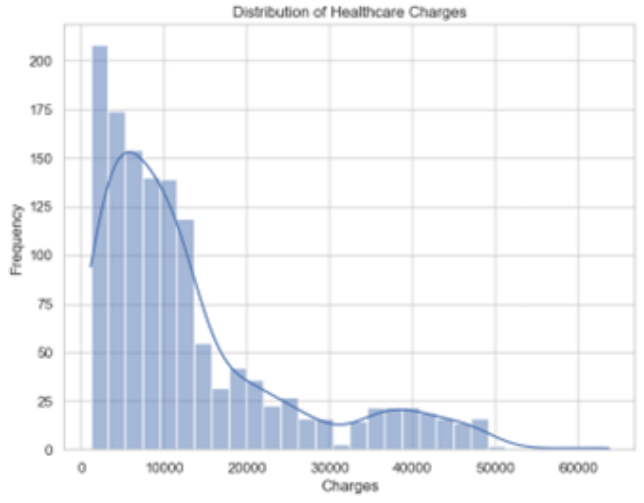


Fig. 4. Histogram of healthcare charges showing a right-skewed distribution with a small number of high-cost outliers.

### C. Scatter plot (Charges Vs. BMI)

- Description: Fig. 5 visualizes the relationship between BMI and healthcare charges.
- Significance: Explores whether BMI impacts the cost of healthcare.
- Insights:
  - A positive trend suggests higher BMI is associated with higher charges.
  - Smokers tend to cluster at higher BMI and charges.

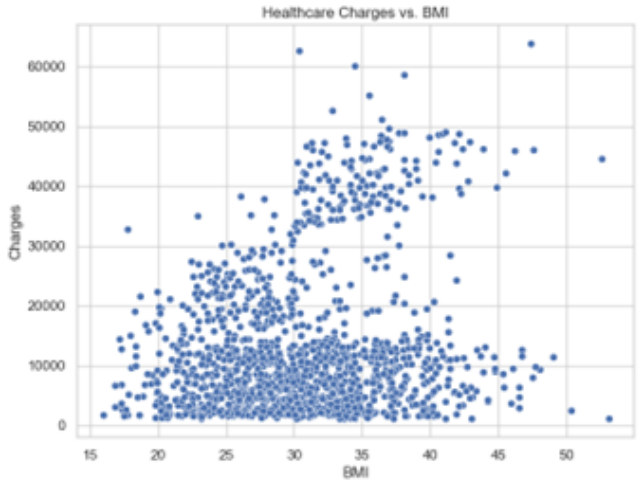


Fig. 5. Scatter plot of charges versus BMI, illustrating an upward cost trend with increasing BMI and visible high-cost cases.

### D. Box Plot (Smoker Vs. Non-Smoker Charges)

- Description: Fig. 6 compares charges between smokers and non-smokers.
- Significance: Illustrates the stark difference in costs based on smoking status.
- Insights:
  - Smokers have significantly higher median charges than non-smokers.
  - Highlights smoking as a major cost driver.

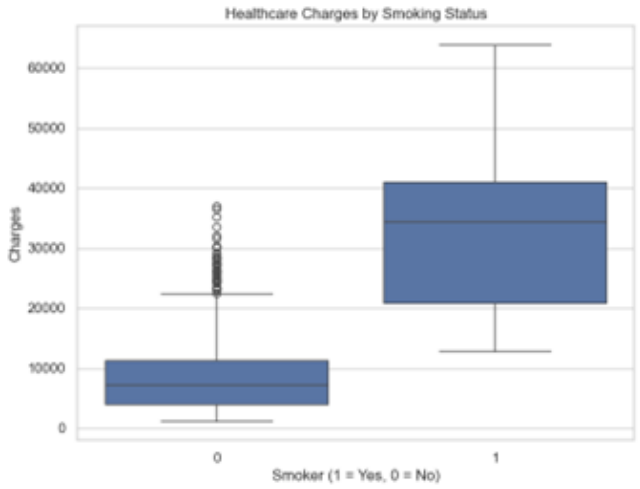


Fig. 6. Box plot comparing charges for smokers and non-smokers, showing substantially higher median and variability of costs among smokers.

### E. Elbow Method for Optimal Charges

- Description: Fig. 7 line graph of WCSS (Within-Cluster Sum of Squares) vs. the number of clusters.
- Significance: Determines the optimal number of clusters for segmentation.
- Insights:
  - The elbow point at k=2 suggests two clusters are sufficient to segment the data meaningfully.

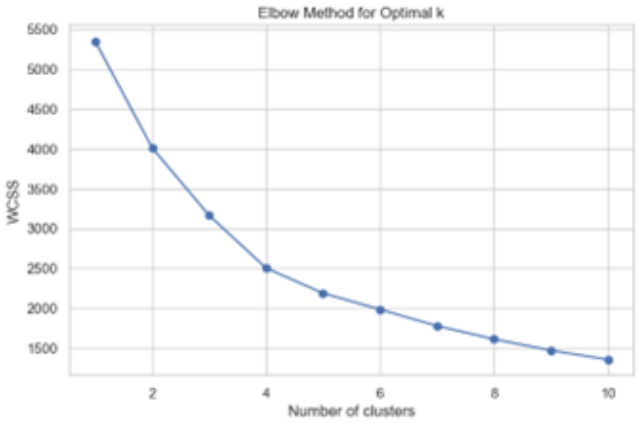


Fig. 7. Elbow method curve for K-Means clustering, used to select the number of clusters by identifying diminishing returns in within-cluster inertia.

### F. Average Charges Per Cluster (Bar Plot)

- Description: Bar plot in Fig. 8 showing the average charges for each cluster.
- Significance: Highlights cost differences between clusters.
- Insights:
  - Cluster 0 (likely smokers) has significantly higher average charges.
  - Cluster 1 (likely non-smokers) has much lower charges.

### G. Radar Chart for Cluster Classification

- Description: Fig. 9 plots cluster attributes like age, BMI, smoking status, and children.
- Significance: Offers a multi-dimensional comparison of clusters.
- Insights:
  - Cluster 0 scores higher on smoking prevalence and BMI.
  - Cluster 1 has balanced attributes, correlating with lower charges.

### H. Residual Plot for Models

- Description: Scatter plots of residuals (actual - predicted charges) for Linear Regression in Fig. 10 , Random Forest in Fig. 11, and XGBoost in Fig. 12 are shown.
- Significance: Evaluates the fit of models by assessing residual patterns.

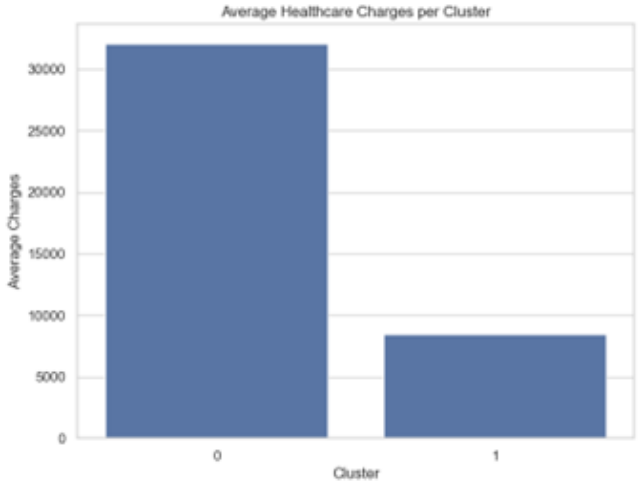


Fig. 8. Average healthcare charges by cluster, demonstrating clear separation between lower-cost and higher-cost risk groups.


Fig. 9. Radar chart summarizing normalized cluster profiles across key features, enabling visual comparison of risk-group characteristics.

- Insights:
  - Random Forest and XGBoost show less clustering and randomness, indicating better fits compared to Linear Regression.

### I. Feature Importance (Bar Plot)

- Description: Bar plot showing the importance of features in Random Forest and XGBoost models. Feature importance analysis shown in Fig. 13 indicates smoking is a key cost driver.
- Significance: Highlights which feature most influence predictions
- Insights:

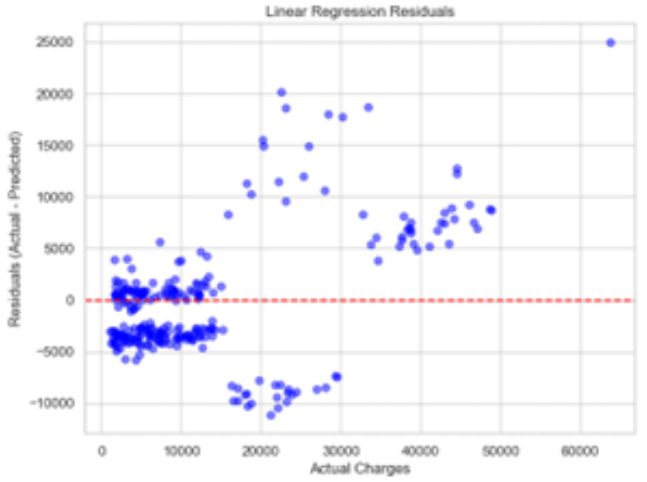


Fig. 10. Residual plot for Linear Regression, showing systematic structure in residuals that suggests limited ability to capture non-linear patterns.

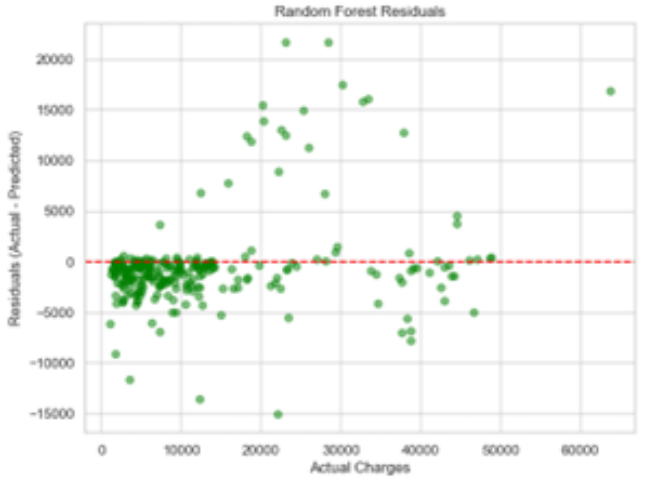


Fig. 11. Residual plot for Random Forest, showing reduced residual magnitude and improved fit compared to the linear baseline.

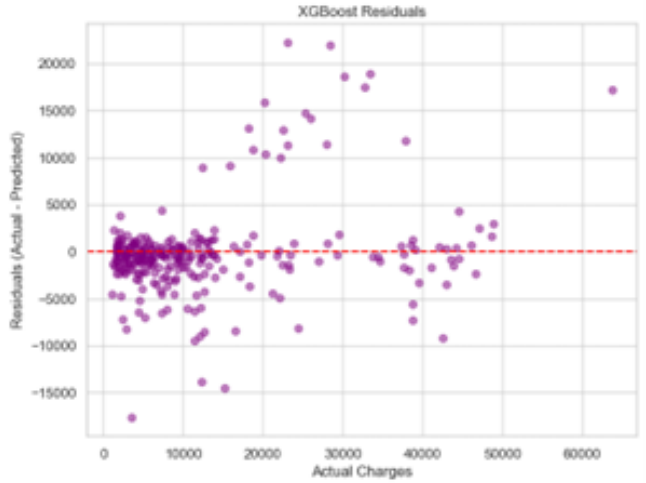


Fig. 12. Residual plot for XGBoost, indicating strong predictive fit with residuals more tightly distributed around zero.

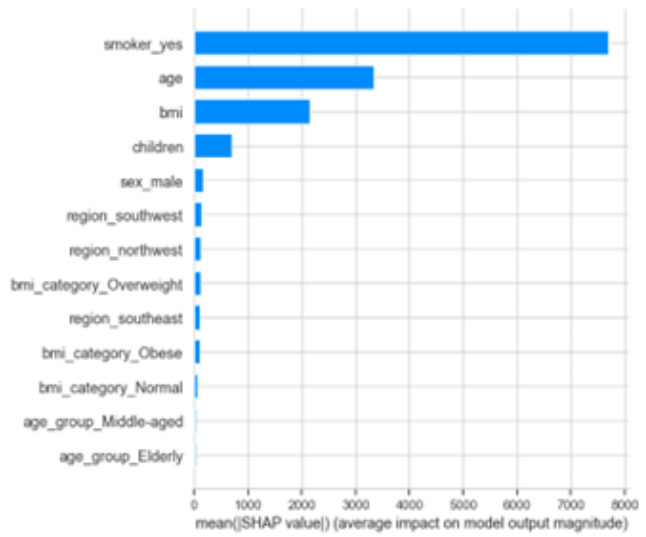


Fig. 13. Feature importance from the best-performing model, highlighting smoking status, BMI, and age as dominant drivers of healthcare charges.

  - Smoking status, BMI, and age are the top contributors.
  - Regional features have minimal impact.

### J. Actual Vs. Predicted Charges (Anomalies Highlighted)

- Description: Scatter plot comparing actual and predicted charges, with anomalies highlighted in shown in Fig. 14
- Significance: Identifies discrepancies between model predictions and actual values.
- Insights:
  - Anomalies (in red) deviate significantly from the expected trend.
  - Indicates potential data entry errors or unique cases.

### K. Charge Differences for Anomalies

- Description: Bar plot of differences between actual and predicted charges for anomalies is shown in Fig. 15
- Significance: Quantifies and visualizes anomalies.
- Insights:
  - Some anomalies show discrepancies exceeding $10,000.
  - Useful for targeted investigation or data corrections.

### L. Cost Comparison (Before and After Comparison)

- Description: Bar plots showing average predicted costs before and after smoking cessation and BMI reduction in Fig. 16
- Significance: Visualizes the financial impact of behavioral interventions.
- Insights:
  - Smoking cessation could save $6,456 per individual on average


Fig. 14. Actual versus predicted charges with anomalies highlighted, illustrating cases with unusually large prediction errors.

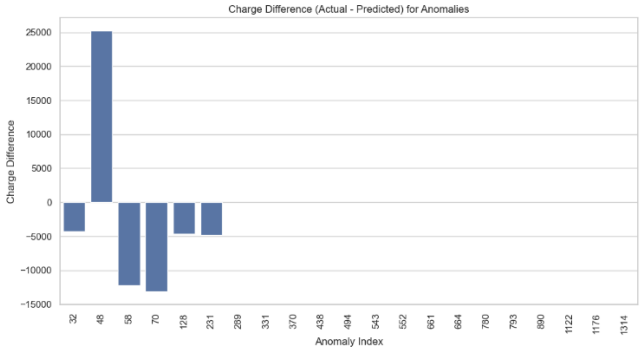


Fig. 15. Charge differences for detected anomalies, ranking individuals by deviation between actual and predicted costs to support targeted investigation.

– A 5% reduction in BMI could save $9,678 per individual on average.

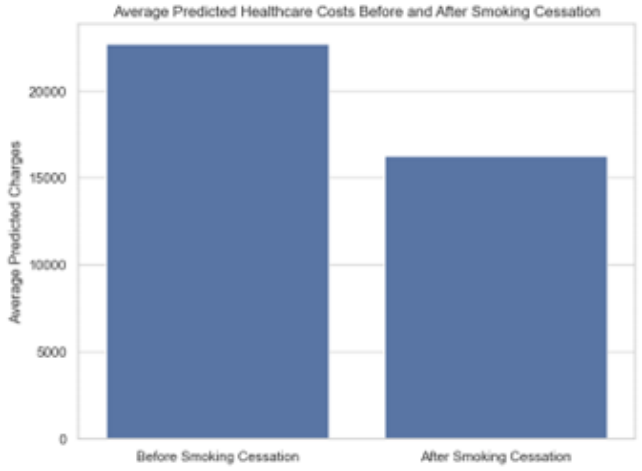


Fig. 16. Simulated cost comparison before and after behavioral interventions (e.g., smoking cessation and BMI reduction), estimating potential cost savings.

## V. Discussion

The results confirm that machine learning models can predict healthcare charges with strong accuracy while also revealing interpretable cost drivers. Across the evaluated models, the non-linear ensemble methods (Random Forest and XGBoost) outperform the linear baseline, indicating that healthcare costs in this dataset are governed by non-linear interactions between lifestyle factors (e.g., smoking), body composition (BMI), and age. This supports the choice of ensemble learning for cost prediction in realistic insurance settings where relationships are rarely purely linear.

From a decision-support perspective, the segmentation results provide an additional layer of interpretability beyond regression performance. The K-Means clustering separates individuals into distinct cost-risk profiles, where the high-cost cluster is strongly characterized by smoking status and elevated average charges. This enables targeted intervention planning, suggesting that risk stratification can be used not only for forecasting but also for designing differentiated prevention strategies and resource allocation.

The prescriptive simulations extend the analysis from prediction to action. By estimating how costs may change under behavioral modifications such as smoking cessation and BMI reduction, the study demonstrates how predictive analytics can support policy and intervention design. While these simulations do not prove causal effects, they provide a practical approximation of financial impact that can help motivate preventive programs and guide insurer decision-making.

Finally, anomaly detection highlights individuals whose costs deviate strongly from model expectations. These anomalies can represent rare medical conditions, unusual utilization patterns, data quality issues, or potential fraud. Identifying such cases is valuable for insurers and policymakers because they can trigger deeper investigation, improve data integrity, and support exceptional-case management. Overall, the novelty of this project lies in combining predictive modeling with segmentation, prescriptive simulation, and anomaly detection in a single cohesive workflow, producing both accurate forecasts and decision-relevant insights.

### A. Future Directions

Future work can strengthen both realism and robustness. First, incorporating richer clinical and utilization variables (diagnoses, visit history, medication, chronic conditions) would improve predictive power and reduce reliance on proxy features. Second, causal inference or quasi-experimental designs could replace purely correlational simulations, enabling more defensible estimates of the true impact of interventions like smoking cessation. Third, model explainability methods (e.g., SHAP) could be added to quantify feature contributions at both global and individual levels, improving trust and adoption. Finally, fairness analysis across demographic groups and external validation on datasets from other regions or time periods would improve generalizability for real-world deployment.

## VI. Conclusion

This project demonstrates the power of data analytics in addressing healthcare cost challenges. Key takeaways include:

- **Significant Factors:** Smoking and BMI were the most influential determinants of health care costs, underlining the necessity of behavioral interventions.
- **Predictive Insights:** The Random Forest model was the most accurate, showing the potential of machine learning in health care analytics.
- **Actionable Recommendations:** Smoking cessation programs and weight management initiatives are the biggest cost-saving opportunities for stakeholders.
- **Anomaly Detection:** The outliers that were identified gave us insights into rare or extreme cases, helping to refine policies and data integrity.

With a perfect blend of predictive modeling, segmentation, prescriptive analytics, and anomaly detection, this program really covers the gamut for comprehensive understanding and

## References


[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM)*. Piscataway, NJ, USA: IEEE, 2008, pp. 413–422.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. New York, NY, USA: ACM, 2016, pp. 785–794.

[6] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[7] M. Choi, "Medical cost personal datasets (insurance.csv)," Kaggle Dataset, 2018, accessed: 2026-02-18. [Online]. Available: https://www.kaggle.com/datasets/mirichoi0218/insurance