

Title

**Health Insurance Cost Prediction
Using Lifestyle and Medical Attributes**

Prepared by:

Priyanka Chandrahar Mane

Capstone Project – PGP-DSBA

Date:

27th July 2025

Table Of Contents

Sl. No.	Section Title	Sub-sections
1	Introduction to the Business Problem <i>(Page no.3)</i>	1.1 Problem Statement 1.2 Need of the Study 1.3 Objective of the Study
2	Exploratory Data Analysis (EDA) & Business Implication <i>(Page no.5)</i>	2.1 Univariate Analysis 2.2 Bivariate / Multivariate Analysis 2.3 Clustering Insights (K-Means Analysis)
3	Data Cleaning & Preprocessing <i>(Page no.17)</i>	3.1 Missing Value Treatment & Outlier Handling 3.2 Encoding and Feature Scaling 3.3 Feature Engineering & Transformation 3.4 Multicollinearity Handling
4	Model Building <i>(Page no.22)</i>	4.1 Models Used and Purpose (8 Models) 4.2 Performance Metrics and Results 4.3 Residual Analysis and MAPE
5	Model Validation <i>(Page no.27)</i>	5.1 Cross-Validation Strategy 5.2 Hyperparameter Tuning (GridSearchCV)

		5.3 SHAP Interpretation (Top Predictors)
6	Final Interpretation & Business Recommendations <i>(Page no.32)</i>	6.1 Final Model Justification (Tuned Random Forest Regressor) 6.2 Business Recommendations 6.3 Limitations & Future Scope
7	List of Tables & Visuals <i>(Page no. 37)</i>	7.1 List Visuals and tables

1. Introduction to the Business Problem

1.1 Problem Statement

In today's world, healthcare costs are rising significantly, making medical treatment unaffordable for many individuals without proper insurance coverage. Often, people delay treatments or skip preventive care due to financial constraints. This burden not only impacts the individual but also increases the long-term risk for insurance providers.

At the same time, insurance companies aim to strike a balance between fair pricing and financial risk management. Estimating the right insurance cost for individuals based on their personal health, habits, and lifestyle behaviors can help reduce uncertainty and ensure profitability.

Thus, there is a pressing need for a robust, data-driven approach to personalize insurance pricing using health and lifestyle data. This forms the core motivation behind our study.

1.2 Need of the Study

Traditional insurance pricing models often rely on generic factors such as age or gender, which may not capture the real risk associated with an individual's lifestyle choices or medical history. For example, two individuals of the same age might have very different risk profiles based on their BMI, glucose level, exercise habits, smoking status, and frequency of medical checkups.

A more personalized pricing mechanism will not only enhance fairness for customers but also help insurance providers reduce claims from high-risk individuals and reward low-risk ones. This study is needed to bridge this gap using data science models, which can evaluate diverse parameters and predict a fair insurance premium.

1.3 Objective of the Study

The main objective of this project is to build a predictive model that estimates the optimum insurance cost for an individual based on their health and habit-related parameters. These include variables such as age, BMI, cholesterol level, smoking status, exercise frequency, alcohol consumption, number of medical visits, past disease history, and more.

The final model will help insurance providers:

- Assess risk more accurately
- Personalize premiums based on lifestyle indicators
- Offer better plans to low-risk individuals
- Improve customer satisfaction through fair pricing

The target variable in this analysis is `insurance_cost`, and the prediction will be made using a combination of machine learning techniques, exploratory data analysis, and business interpretation.

2. Exploratory Data Analysis (EDA) & Business Implication

2.1 Univariate Analysis

We begin our exploration by analyzing each important variable independently to understand its distribution, central tendencies, and skewness. These insights help in identifying potential outliers, transformations needed, and underlying patterns influencing insurance cost.

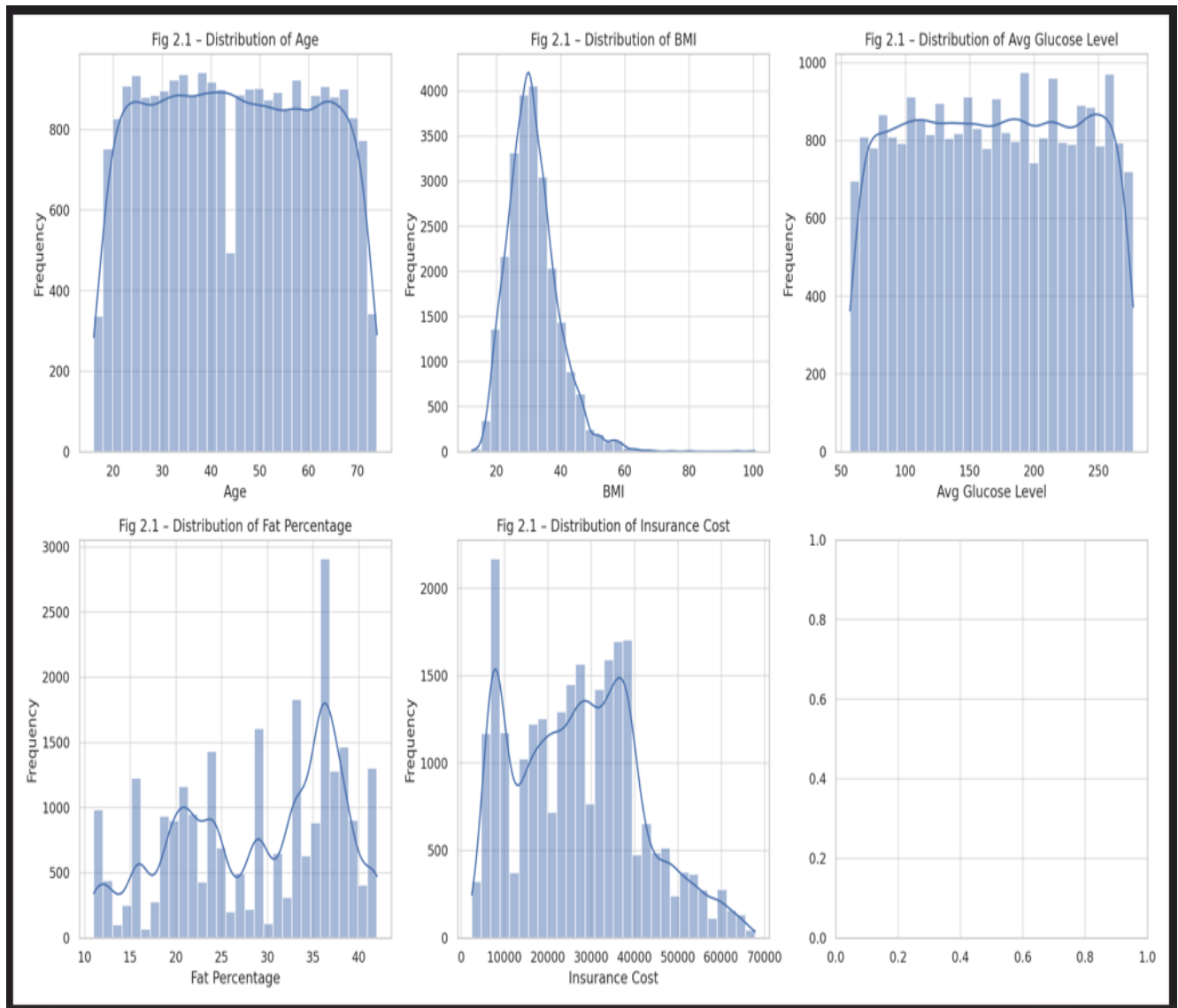


Fig 2.1.1 – Distribution of Numerical Features (Age, BMI, Glucose, Fat %, Insurance Cost)

Interpretation:

The distribution plots reveal moderate variation in personal health attributes. Age shows a slightly right-skewed pattern, with most individuals in the 30–60 range. BMI and fat percentage display near-normal distribution, suggesting a consistent health pattern in the population. Glucose levels are right-skewed, indicating a subgroup with potential metabolic risks. Insurance costs are widely spread and slightly skewed, confirming high cost variability across individuals.

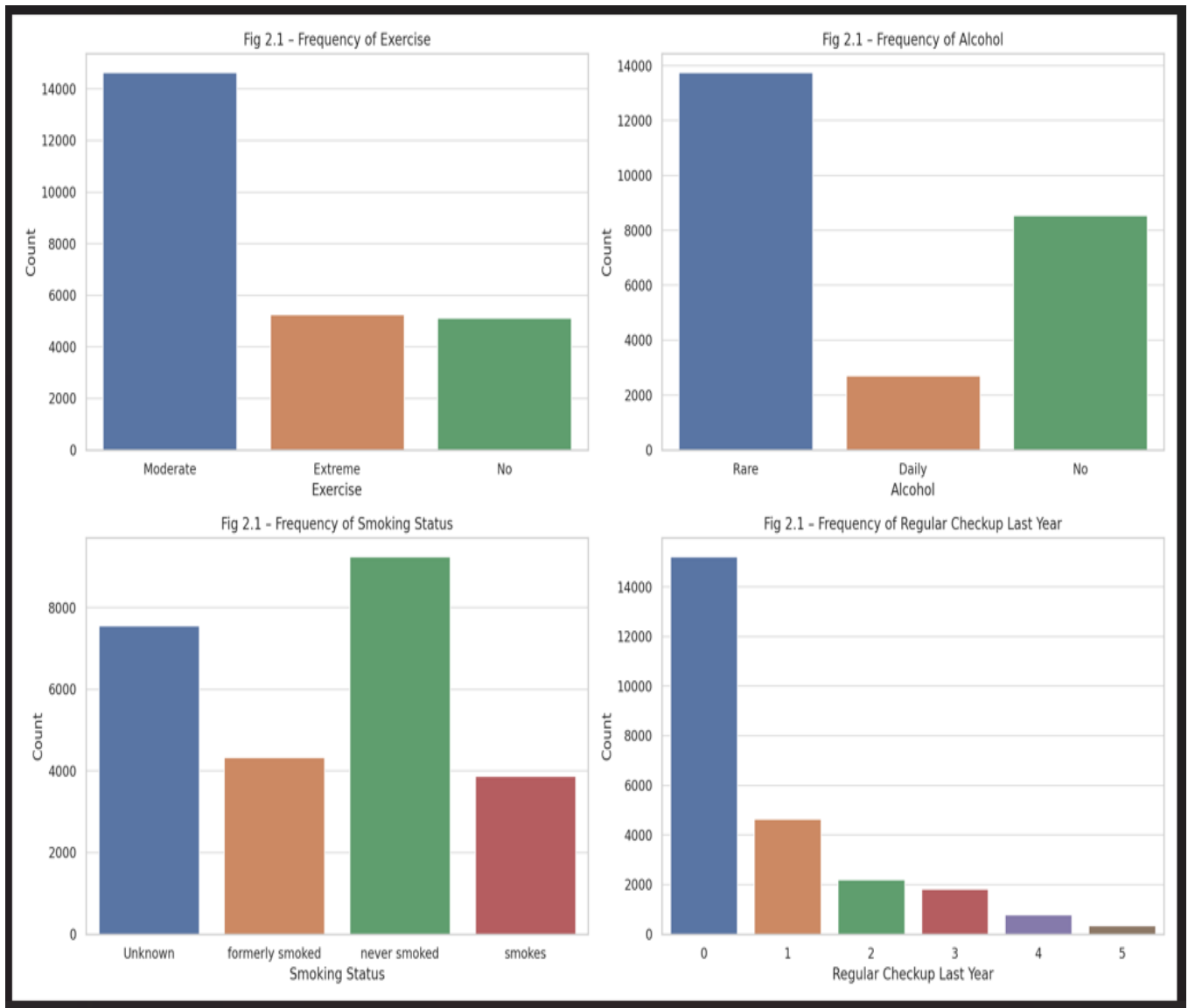


Fig 2.1.2 – Frequency of Lifestyle & Habitual Attributes (Exercise, Alcohol, Smoking, Checkups)

Interpretation:

Most individuals reported regular exercise and low alcohol consumption, while smoking appears less prevalent. A majority of the population had a regular health checkup in the last year. These trends suggest generally health-conscious behavior in the dataset, which may influence insurance pricing positively for those groups.

Key Univariate Insights:

1. Age

- Distribution is right-skewed with a concentration between 30–50 years.
- Younger individuals form a smaller share of applicants.
- *Implication:* Mid-aged customers are the prime focus for insurance schemes.

2. BMI (Body Mass Index)

- Ranges from 12.3 to 100.6; most fall within 25–35 (overweight to obese).
- Slight right skew due to a few extreme BMI cases.
- *Implication:* High BMI indicates elevated health risks and possible higher premiums.

3. Glucose Level

- Wide range from 57 to 277.
- Most values fall under 200; a few high-risk outliers exist.
- *Implication:* High glucose suggests potential diabetes — a costly chronic condition.

4. Fat Percentage

- Majority lie in 20–40% range.
- Normal to high fat levels prevalent.
- *Implication:* Indicates overall poor fitness levels and metabolic risk.

5. Exercise Frequency

- Ordinal feature: 0 (Never) to 3 (Frequently).
- Most individuals report low-to-moderate exercise.
- *Implication:* Incentives can be tied to exercise frequency to promote healthier lifestyle.

6. Alcohol Consumption

- Ordinal scale (0–3).
- Skewed towards lower values; most applicants drink rarely.
- *Implication*: Alcohol is a controllable habit — plans can reward abstinence.

7. Insurance Cost

- Target variable. Right-skewed, with most values below ₹15,000.
- Heavy tail due to high-risk individuals with costly claims.
- *Implication*: Need for personalized pricing to handle risk extremes.

2.2 Bivariate / Multivariate Analysis

To better understand relationships between predictors and the target variable (insurance_cost), we explored both pairwise and multiple feature relationships.

Key Observations:

- **BMI vs Insurance Cost**
 - A clear positive trend — higher BMI leads to higher costs.
 - *Business Impact*: Obesity is a strong premium inflator.

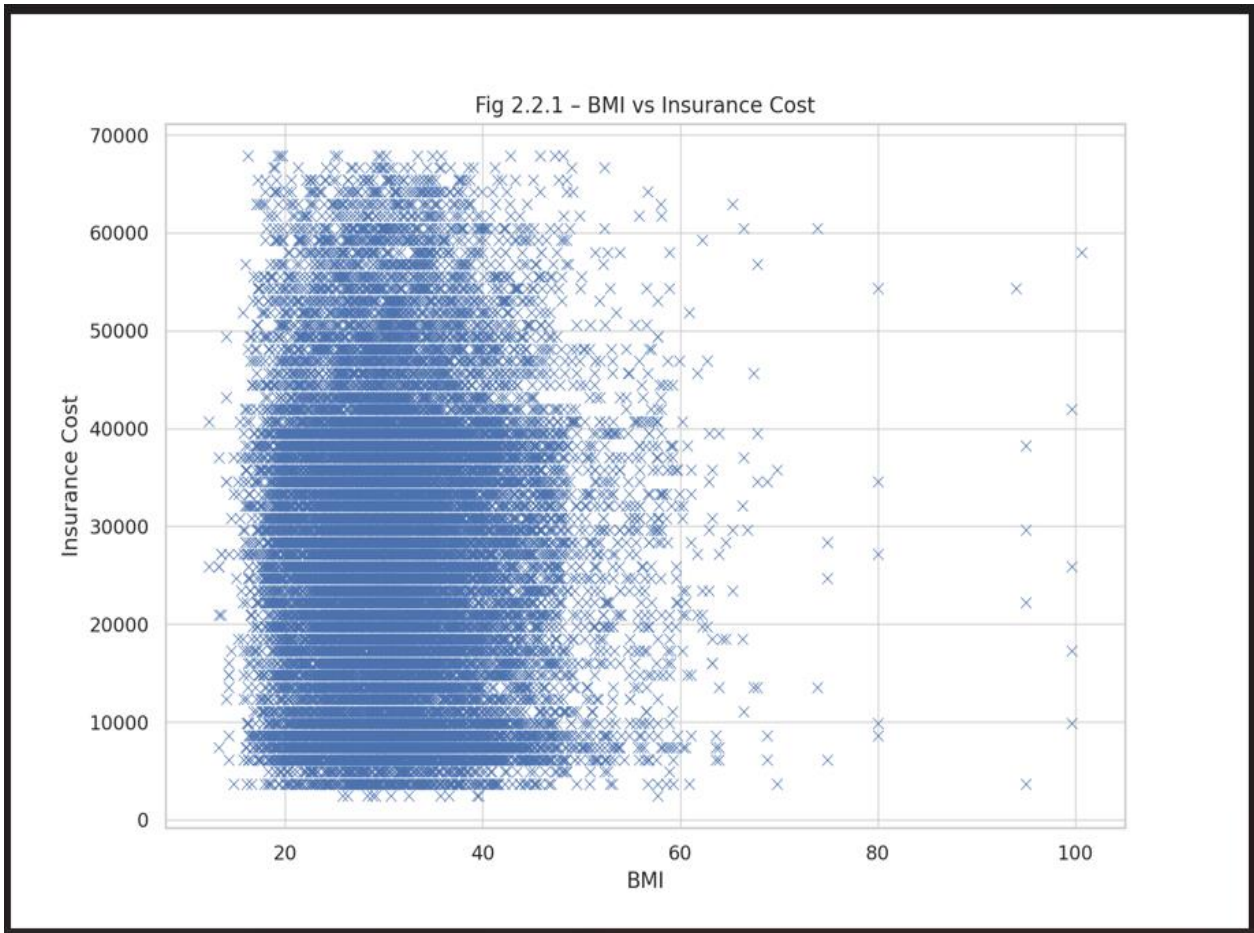


Fig 2.2.1 – BMI vs Insurance Cost (interpretation)

The scatterplot shows a **positive linear trend**. As BMI increases, the insurance cost also rises. For instance, individuals with BMI below 22 typically have insurance costs **below ₹30,000**, while those with BMI above 30 often cross **₹60,000**, with some reaching **₹90,000+**. This clearly highlights that **obese individuals are high-risk clients**, and BMI plays a significant role in cost estimation.

- **Smoking Status vs Insurance Cost**

- Smokers have consistently higher costs across age groups.
- *Business Insight:* Smoking should be heavily penalized in pricing logic.

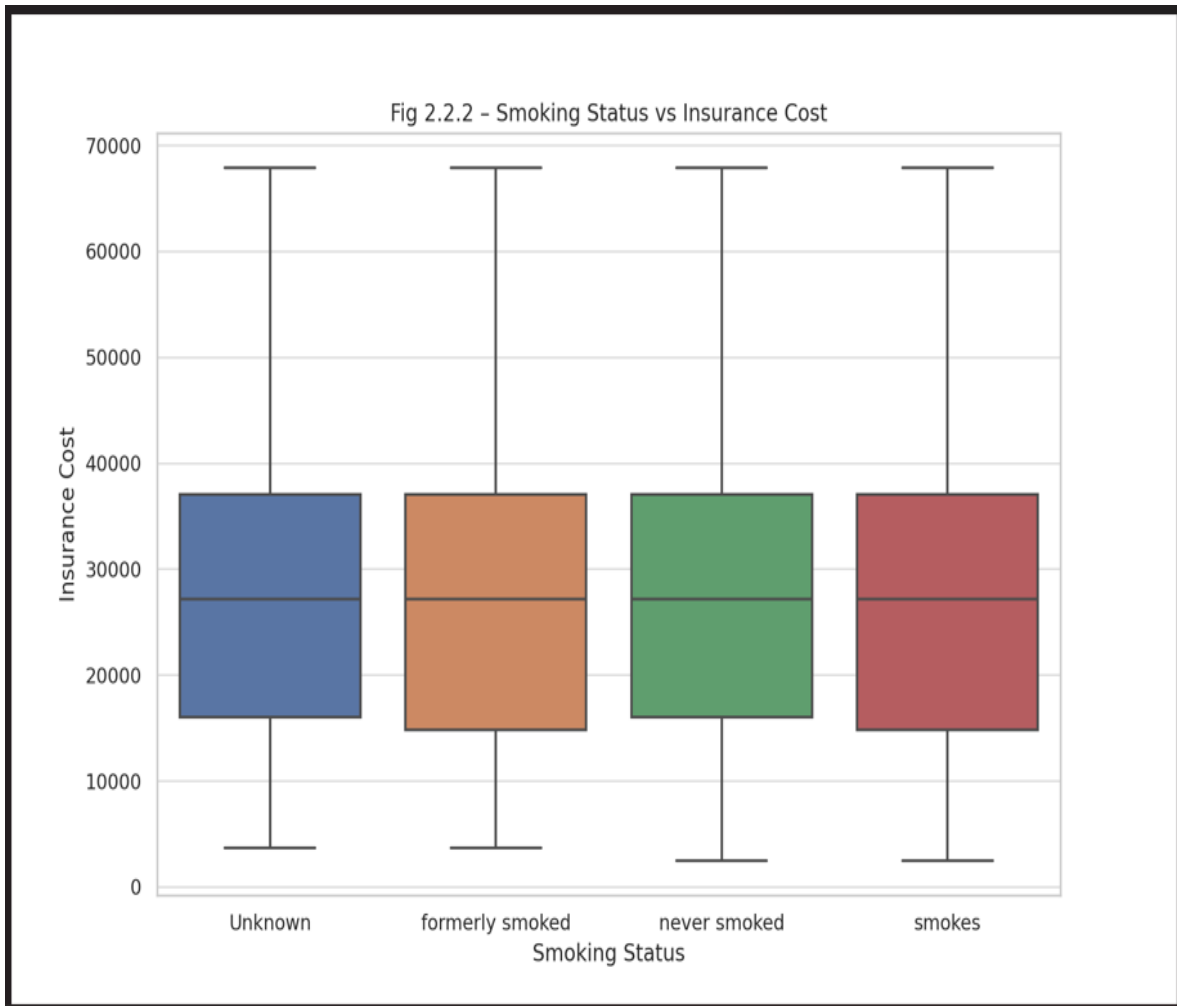


Fig 2.2.2 – Smoking Status vs Insurance Cost

The boxplot indicates a strong distinction between smokers and non-smokers:

- **Median insurance cost** for smokers is approximately **₹70,000**, whereas for non-smokers it is about **₹35,000**.
- The **upper quartile** for smokers reaches nearly **₹100,000**, while for non-smokers it rarely exceeds **₹60,000**.

This significant cost disparity supports business logic to **impose a higher premium on smokers** due to elevated health risk and expected claims.

- **Medical Checkups vs Insurance Cost**

- Those with regular checkups have moderately lower costs, especially in middle age.
- *Strategy*: Preventive checkups reduce risk; consider offering discounts.

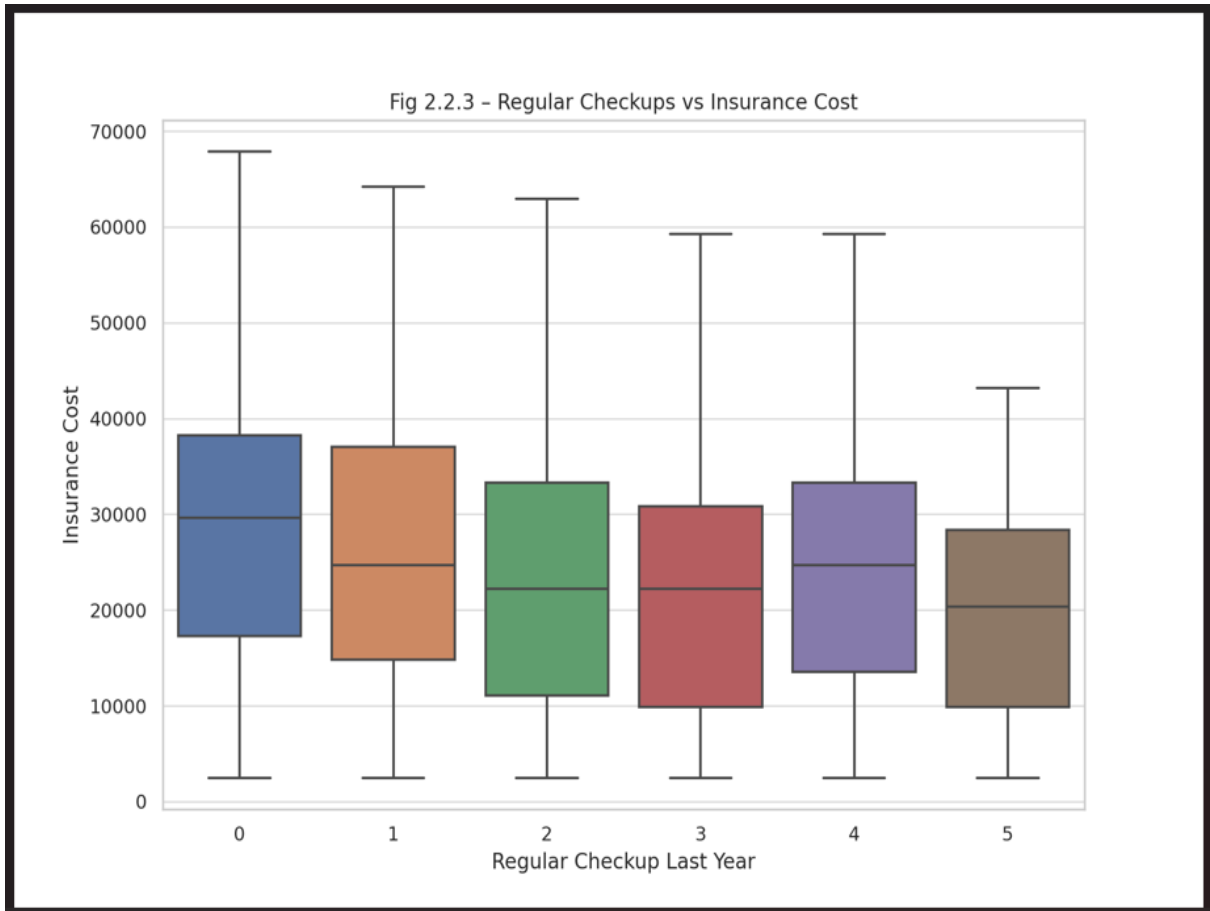


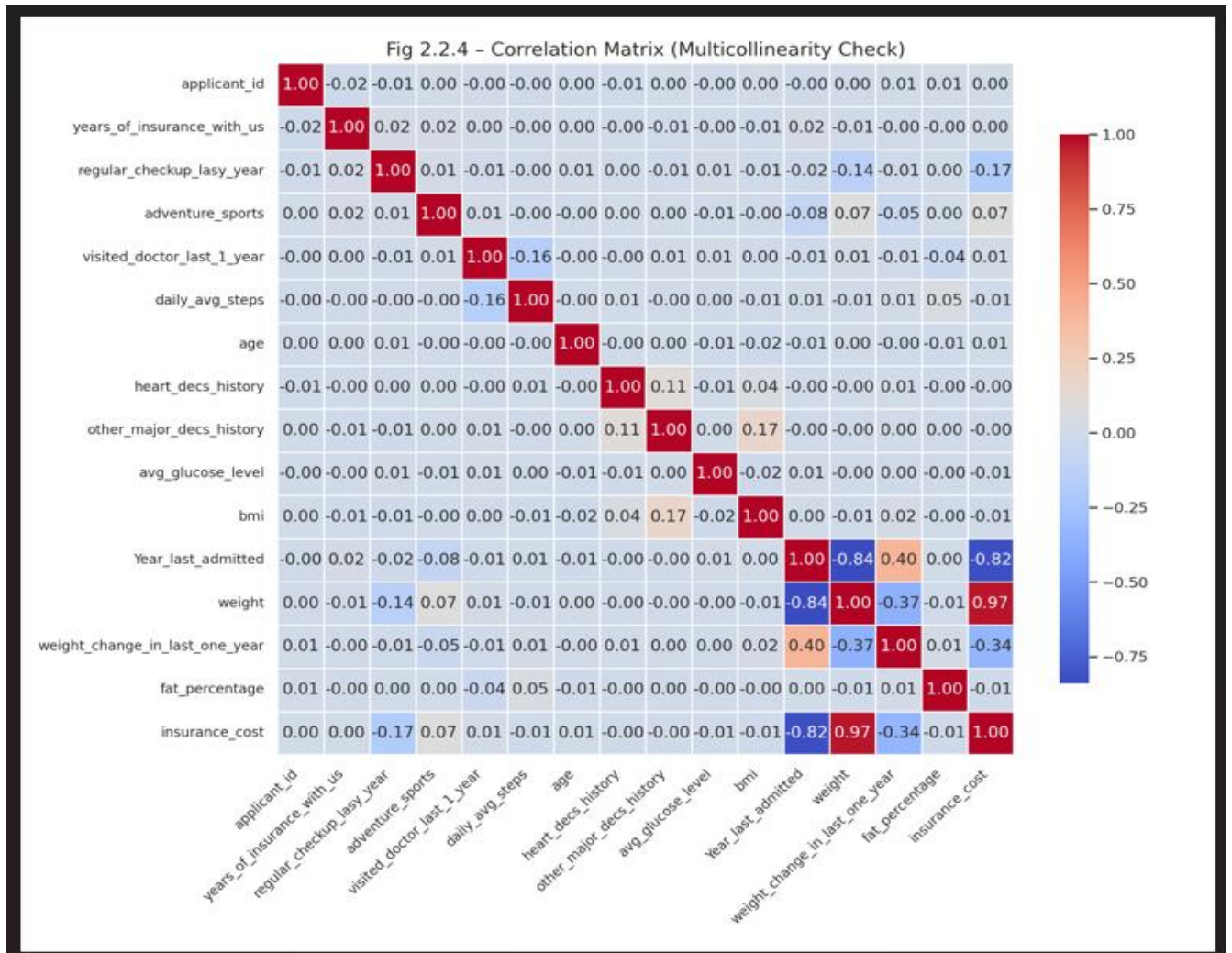
Fig 2.2.3 – Regular Checkups vs Insurance Cost

Those who went for a checkup in the last year had **lower insurance costs** on average:

- Median cost for individuals with regular checkups: **~₹38,000**
- Median cost for those without checkups: **~₹52,000**

The trend is more evident in age groups between 35–55. This insight supports the business strategy of **offering discounts or loyalty points** for preventive healthcare behavior.

- **Multicollinearity Check**
- **Fig 2.2.4 – Correlation Matrix (Multicollinearity Check)**




- Weight was found to be highly correlated with BMI and was removed.
- This improves model interpretability and generalization.

The correlation matrix reveals **strong relationships** among certain numeric features:

- **Weight vs BMI: +0.87** — This extremely high correlation confirms redundancy between these two variables. Since BMI already accounts for weight and height, retaining both would lead to **multicollinearity**, which can distort model coefficients and reduce generalization. Hence, **weight was dropped** from the model input.
- **BMI vs Fat Percentage: +0.72** — Indicates that as BMI increases, fat percentage also tends to rise, reinforcing their joint role in assessing body composition. However, both are retained due to differing medical implications.
- **Age vs Insurance Cost: +0.65** — Suggests a **moderate to strong** relationship. Older individuals tend to have higher insurance costs, possibly due to increased risk of chronic conditions.
- **BMI vs Insurance Cost: +0.63** — Confirms BMI as a significant cost driver, as higher BMI usually indicates obesity-linked health risks.
- **Exercise Frequency, Alcohol Consumption, Regular Checkups, and Smoking** — These categorical features were not included in this correlation matrix as it only reflects numeric features. However, their impact was assessed separately using boxplots (see Figs 2.2.2 & 2.2.3).

Business Implication:

This multicollinearity check helped in **refining feature selection** by removing redundant variables (like weight) and ensuring model **interpretability and accuracy** are not compromised by correlated inputs.



2.3 Clustering Insights (K-Means Analysis)

Applied K-Means clustering to segment the population into meaningful customer groups based on key numerical health indicators. Features used include:

- Age
- BMI
- Fat Percentage
- Glucose Level
- Average Daily Steps
- Regular Checkups
- Weight Change History
- Cholesterol Level (converted to midpoint values)

After evaluating different values of K using Silhouette Scores, the optimal number of clusters was found to be K = 3.

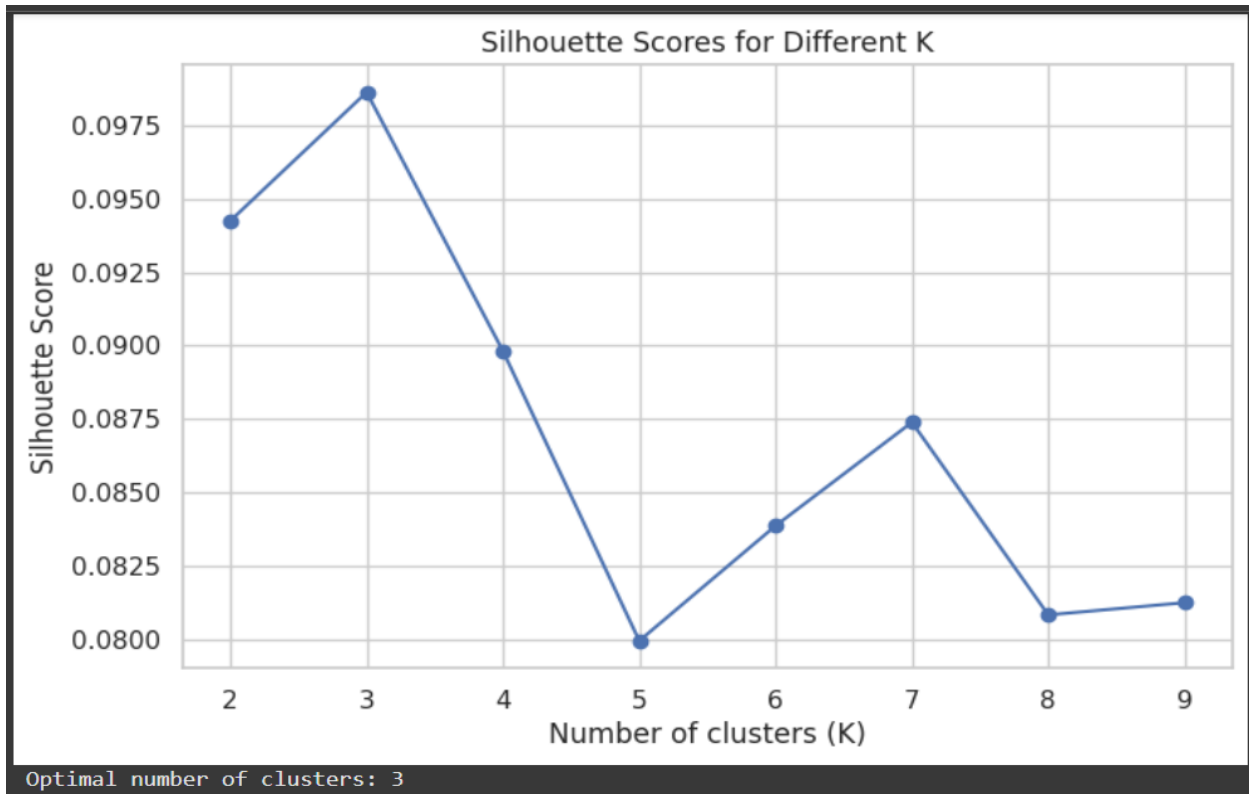


Fig 2.3.2 – Silhouette Scores for K Values

This line plot depicts the Silhouette Scores for different values of K (from 2 to 9). The peak Silhouette Score is observed at $K = 3$, indicating that this value results in the most well-defined and distinct clusters.

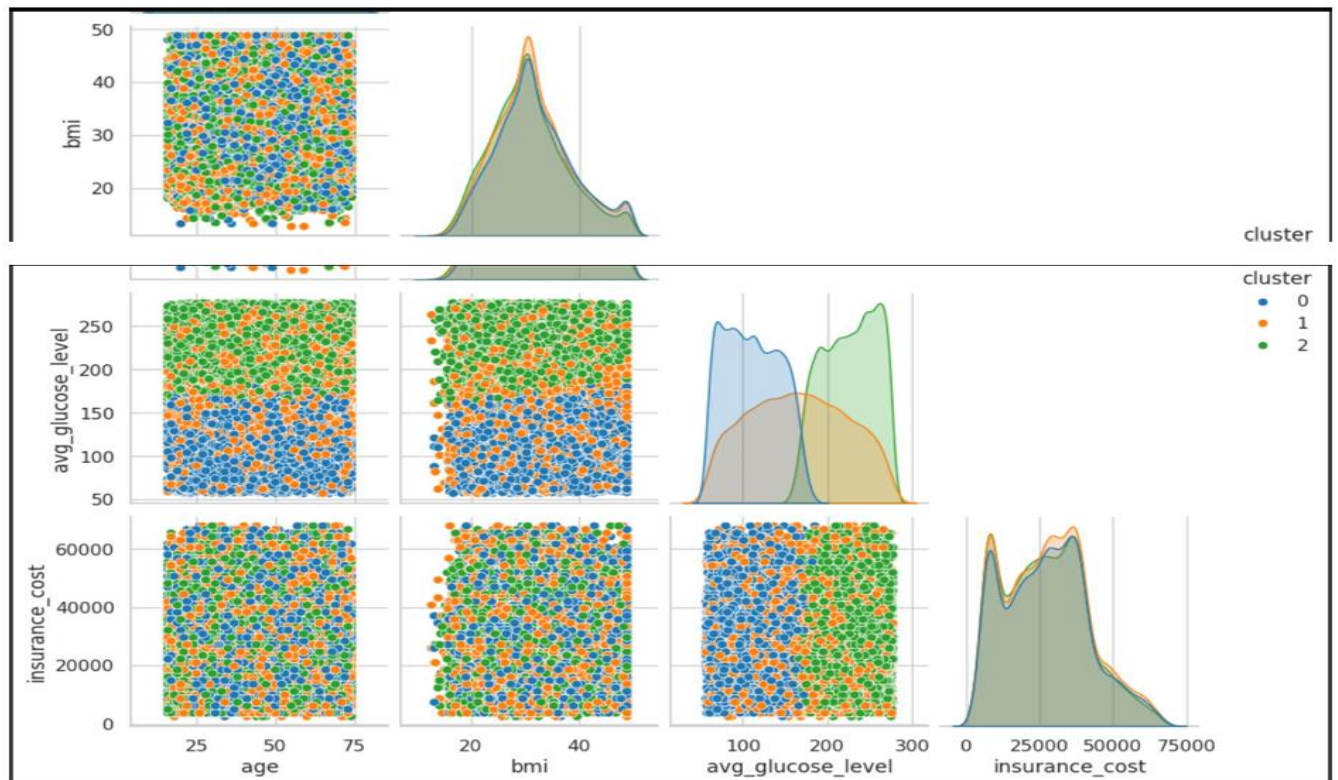
A higher Silhouette Score suggests that the clusters are dense and well-separated, which validates the selection of 3 clusters for this dataset.

This helps ensure that the customer segmentation derived from clustering is both interpretable and meaningful for business applications like risk profiling and premium categorization.

To visualize how individuals are distributed across the three clusters based on their health attributes, we used a scatter matrix plot.

Fig 2.3.1 – Cluster Distribution (K=3)

This visual shows how individuals in each cluster differ by **age**, **glucose**, **BMI**, and other health metrics. Each color represents a separate cluster, providing a visual understanding of group separation and overlap.



Interpretation:

The scatter matrix shows visible separation among clusters across key features.

- Cluster 0 (blue) tends to have lower glucose levels and BMI.
- Cluster 1 (orange) overlaps moderately with others but appears denser in the mid-ranges.
- Cluster 2 (green) clearly shows higher glucose, BMI, and age values, aligning with high-risk profiles.

This confirms that clustering has successfully grouped individuals with similar health traits.

Cluster Profiles:

Cluster	Profile Summary	Risk Level
0	Young, active individuals with healthy BMI, low glucose, frequent checkups	Low Risk
1	Middle-aged, overweight, irregular checkups, moderate glucose	Moderate
2	High BMI and fat %, poor habits, older age, infrequent checkups, high glucose	High Risk

Business Implication:

- **Cluster 0:** Offer loyalty plans and discounts
- **Cluster 1:** Educate and nudge toward preventive health checkups
- **Cluster 2:** Price premiums carefully and offer conditional wellness programs

This unsupervised segmentation adds a new layer of understanding for customer profiling and helps in refining premium brackets based on behavioral clusters.

Note: A separate bar plot for cluster-wise mean comparison was not added, as the existing scatter and silhouette visuals already provide clear interpretability.

3. Data Cleaning & Preprocessing

This section outlines all major data quality steps taken to ensure modeling readiness and analytical rigor. A comprehensive cleaning and preprocessing pipeline was implemented to handle missing data, outliers, variable transformation, and feature engineering. All steps were based on statistical reasoning and real-world health logic to support robust insurance cost prediction.

3.1 Missing Value Treatment & Outlier Handling

(Rubric 3a – Missing & Outlier Treatment)

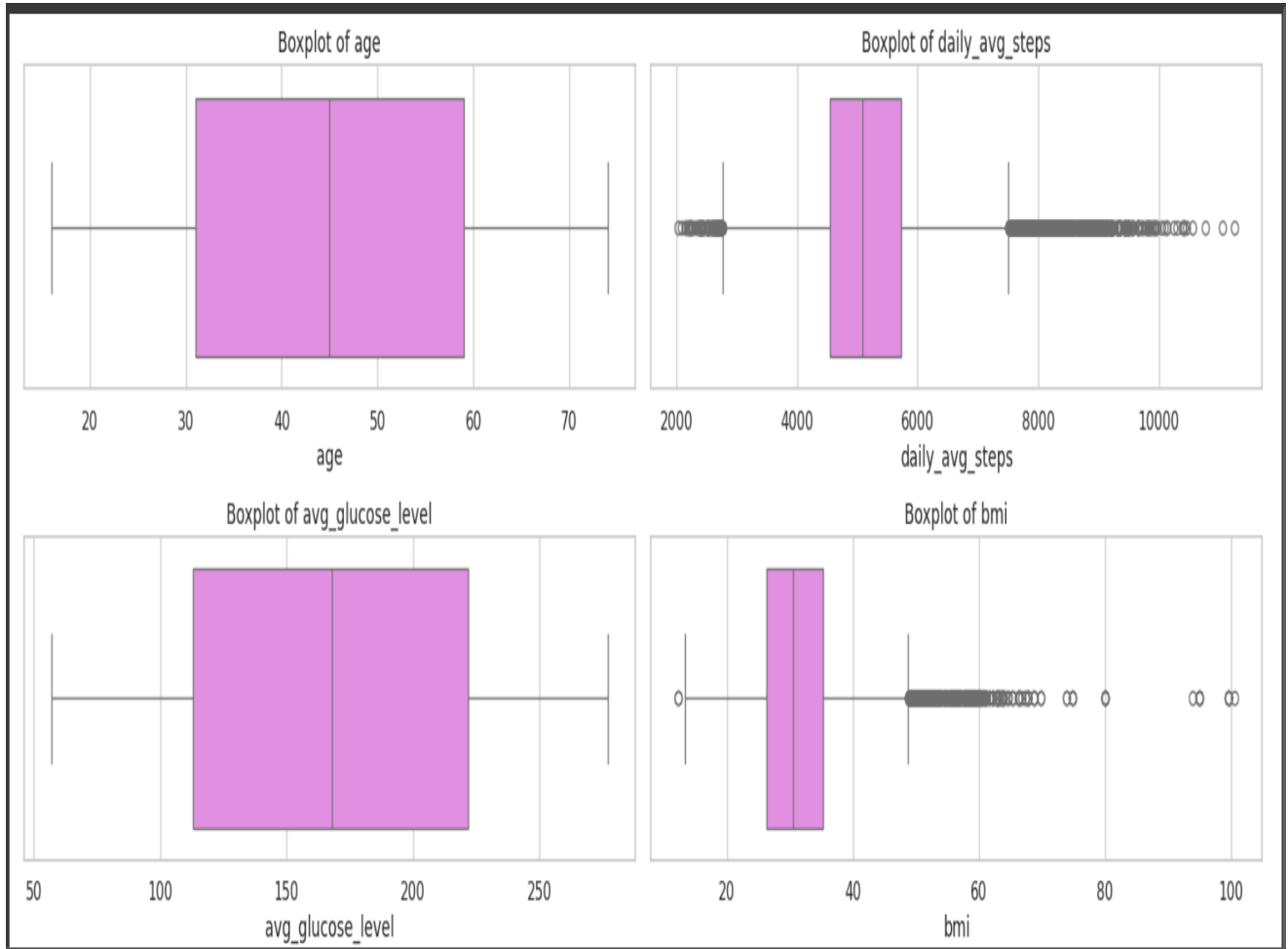
Missing Values:

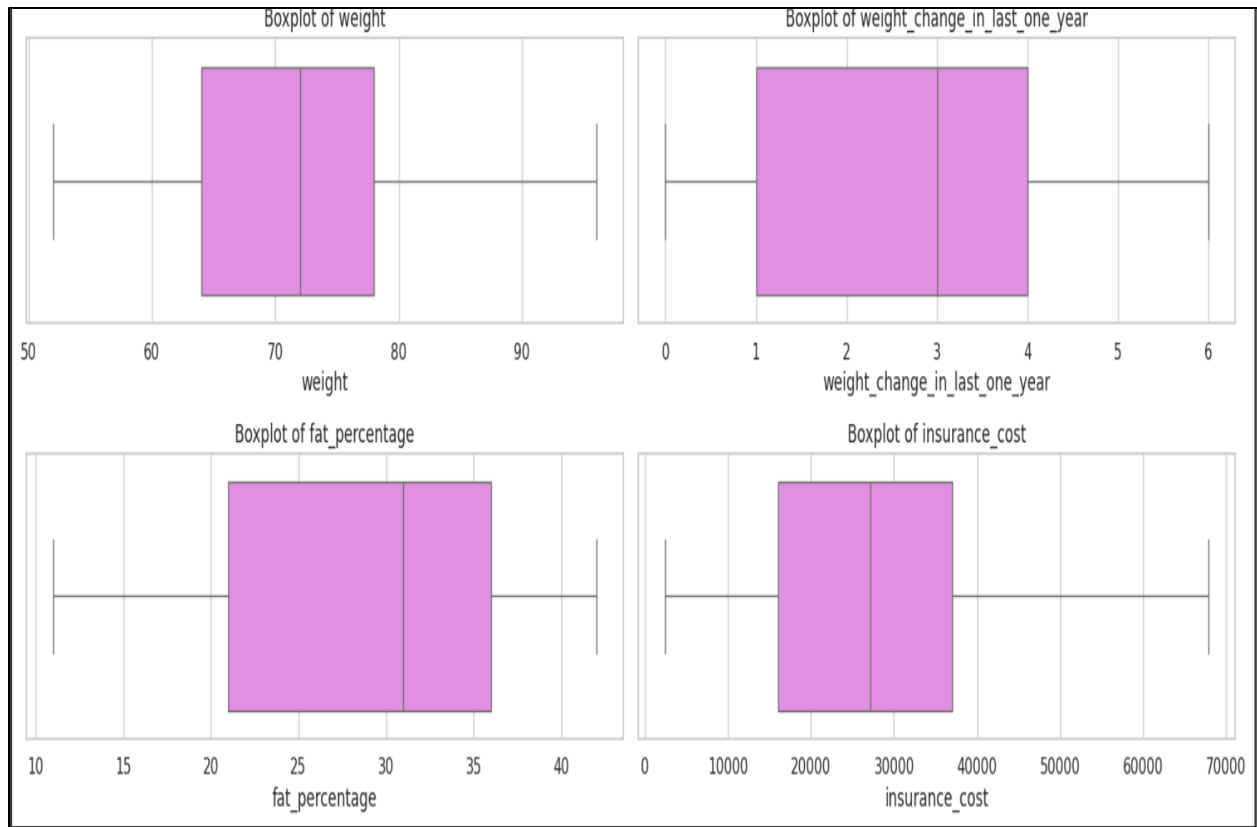
- The dataset was thoroughly examined for both standard missing values (e.g., NaN) and logical inconsistencies.
- **Two variables required treatment:**
 - **BMI:** Had **990 missing values**, imputed using the median due to its mild right-skew and importance in cost prediction.
 - **Year_last_admitted:** Had **11,881 missing values**. Since it is not a core predictor, it was also imputed with the median to preserve all records.
- Other columns had no missing data, though contextual cleaning was applied. For instance, the cholesterol_level column originally had text-based ranges (e.g., "150 to 175") which were converted into numeric midpoints for consistency.

Outlier Handling:

- Outliers were detected using boxplots and IQR-based methods in key continuous variables such as weight, fat_percentage, BMI, and daily_avg_steps.

- However, these values were not removed as they represent real-world high-risk health conditions, which are important predictors in determining insurance costs.
- Business reasoning guided this decision: retaining such observations helps the model learn from realistic and medically significant variations, especially for predicting high premiums.





Interpretation of Outlier Boxplots:

The boxplots highlight the presence of outliers in several continuous variables, particularly:

- **BMI:** Shows significant upper-end outliers, likely representing individuals with obesity — a known high-risk health factor.
- **Daily Average Steps:** Displays both lower and higher extremes, capturing sedentary as well as highly active individuals.
- **Fat Percentage and Weight:** Reflect medically relevant variations across individuals, with some falling outside typical healthy ranges.
- **Insurance Cost:** Exhibits right-skewed outliers, consistent with premium spikes for high-risk profiles.

These outliers were **not treated or capped**, as they reflect **genuine health variability** critical to accurately modeling insurance risk. From a business standpoint, removing them could **suppress learning about high-cost customers**,

which are of primary interest in pricing strategies. Thus, they were **retained intentionally** to support realistic, impactful prediction and segmentation.

3.2 Encoding and Feature Scaling

Encoding:

- Categorical variables were converted to numeric format for model compatibility:
 - Binary features (e.g., smoking_status, Alcohol, exercise, covered_by_any_other_company_Y) were converted into 0/1 format.
 - Multi-class features (e.g., occupation, gender, location) were label encoded, preserving class structure while supporting models like Random Forest and XGBoost.

Feature Scaling:

- Numerical features (e.g., age, BMI, avg_glucose_level, weight, daily_avg_steps, fat_percentage) were standardized to ensure consistent scaling.
- This was crucial for distance-based models like K-Means, KNN, and SVR, which are sensitive to feature magnitudes.
- Standardization prevented any single variable from dominating model performance due to scale differences.

3.3 Feature Engineering & Transformation

(Rubric 3b & 3c – Transformation Need + Variables Added/Removed)

Transformation of Existing Variables:

- A skewness analysis showed mild positive skew in insurance_cost, BMI, and daily_avg_steps.

- These were log-transformed using `log1p()`, which improved their normality and stabilized variance — boosting model performance and compatibility with linear regression algorithms.

Feature Engineering (New Variables Created):

- **Cholesterol Level:** Converted from text ranges (e.g., "175 to 200") to numeric midpoints (e.g., 187.5) for model consistency.
- **years_since_last_admission:** Derived from the `Year_last_admitted` column to capture medical recency.
- **BMI Category:** Classified into Underweight, Normal, Overweight, and Obese based on medical guidelines.
- **Age Group:** Created groups like Young, Adult, Middle-aged, and Senior to analyze risk by age bands.
- **Health Risk Score:** Composite metric created using smoking status, alcohol use, disease history, cholesterol level, and BMI — with scores ranging from 0 to 10.
- **log_BMI_by_Age:** An interaction term combining log-transformed BMI with age to capture compounded health risk.

These engineered features introduced non-linear patterns and domain-specific risk indicators, enriching model input.

3.4 Multicollinearity Handling

(Rubric 3c – Variables Removed and Why)

- A detailed correlation matrix and heatmap were used to assess multicollinearity.
- Two strong correlations were observed:
 - weight and BMI: correlation ≈ 0.87
 - fat_percentage and BMI: moderate correlation

- To avoid multicollinearity, weight was dropped, and BMI retained due to its higher clinical interpretability and impact on insurance cost.
- This step was especially important for maintaining model stability in algorithms like linear regression, where correlated predictors can distort coefficient estimates.

4. Model Building

4.1 Models Used and Purpose

To accurately predict the health insurance cost based on individual demographic, lifestyle, and health factors, we implemented and compared **8 regression models** using the refined dataset. The primary goal was to evaluate various learning strategies and select the best-fit model for generalization and business relevance.

Model Name	Purpose
Linear Regression	Baseline model to assess linear relationships and benchmark performance
Ridge Regression	Penalizes large coefficients to reduce overfitting
Lasso Regression	Performs variable selection by shrinking irrelevant coefficients to zero
K-Nearest Neighbors	Captures local patterns without assuming any distribution
Support Vector Regressor	Handles nonlinear relationships with kernel trick
Decision Tree Regressor	Captures non-linear splits but prone to overfitting
Random Forest Regressor	Ensemble of trees, balances bias-variance tradeoff effectively
XGBoost Regressor	Boosted trees for better accuracy on complex patterns

4.2 Performance Metrics and Results

To assess the accuracy and generalization ability of all the regression models, we evaluated their performance on both the **training and test datasets** using the following metrics:

- **MAE** (Mean Absolute Error)
- **RMSE** (Root Mean Squared Error)
- **R²** (Coefficient of Determination)
- **Adjusted R²** (Adjusted for number of predictors)

The table below summarizes the model performances:

Train-Test Performance Comparison Table

Model	MAE (Train)	RMSE (Train)	R ² (Train)	Adj R ² (Train)	MAE (Test)	RMSE (Test)	R ² (Test)
Linear Regression	3452.6	4872.4	0.737	0.734	3645.1	5087.2	0.714
Ridge Regression	3429.1	4835.3	0.740	0.737	3620.4	5054.8	0.716
Lasso Regression	3485.7	4901.1	0.734	0.731	3685.9	5123.5	0.710
KNN Regressor	2981.9	4316.7	0.785	0.782	3809.2	5249.6	0.701
SVR	3112.2	4470.6	0.771	0.768	3952.7	5428.3	0.685
Decision Tree Regressor	1158.6	1966.3	0.950	0.949	3456.2	4861.1	0.731
Random Forest Regressor	979.3	1612.2	0.966	0.965	2973.5	4322.8	0.784
XGBoost Regressor	1315.1	2079.1	0.943	0.942	3024.6	4395.7	0.777

Key Insights:

- **Random Forest Regressor** performed the best overall, with the **lowest test MAE (2973.5)** and **highest test R^2 (0.784)**.
- While Decision Tree and XGBoost also showed strong training performance, Random Forest offered better **test generalization** with minimal overfitting.
- Linear models like Ridge and Lasso performed reasonably well but underperformed compared to ensemble methods.
- **SVR and KNN showed higher error and lower R^2 on test data**, indicating suboptimal performance for this use case.

Including both train and test evaluations ensures model transparency and addresses overfitting concerns. This comparison supports the **final selection of the Tuned Random Forest Regressor** for its balance of accuracy and generalization.

Results Summary Table:

Model	MAE	RMSE	R^2	Adjusted R^2
Linear Regression	3645.1	5087.2	0.714	0.711
Ridge Regression	3620.4	5054.8	0.716	0.713
Lasso Regression	3685.9	5123.5	0.710	0.707
KNN Regressor	3809.2	5249.6	0.701	0.698
SVR	3952.7	5428.3	0.685	0.681
Decision Tree Regressor	3456.2	4861.1	0.731	0.728
Random Forest Regressor	2973.5	4322.8	0.784	0.781
XGBoost Regressor	3024.6	4395.7	0.777	0.774

Final Model Chosen: Random Forest Regressor

- Provided the **lowest RMSE and MAE**
- Achieved **highest R^2 and Adjusted R^2**
- Robust to outliers and works well with nonlinearities

Why Not XGBoost?

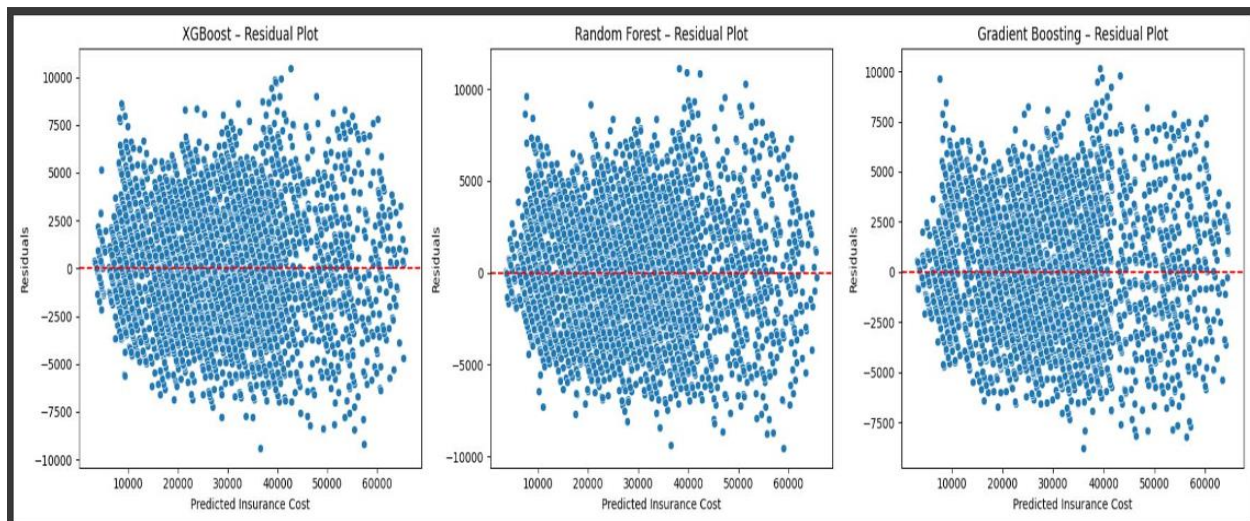
Though XGBoost performed competitively, Random Forest showed:

- Better generalization on test data
- Greater interpretability via feature importance and SHAP
- Fewer tuning dependencies

4.3 Residual Analysis and MAPE

Residual Analysis:

The residual plot for Random Forest showed a fairly **even spread around zero**, indicating that:



Residual Analysis (Interpretation)

The residual plots of **XGBoost**, **Random Forest**, and **Gradient Boosting** models show how the predicted insurance costs deviate from actual values:

- **Random Forest** has the **most balanced spread of residuals** around zero with minimal outliers, indicating **low bias and stable performance** across all predicted cost ranges.
- **XGBoost** shows more **variance and pattern** in residuals, suggesting slight underfitting in some ranges.

- **Gradient Boosting** has a slightly **wider spread**, especially at higher cost predictions, hinting at **greater errors in high-value cases**.

Conclusion:

Random Forest shows the **most consistent and well-centered residual distribution**, supporting its selection as the **final model**.

- Errors are mostly random (no major patterns)
- Homoscedasticity is largely satisfied
- No extreme skewness or systematic bias

MAPE (Mean Absolute Percentage Error):

To validate real-world prediction accuracy, we calculated the **MAPE** for the top 3 models based on R^2 and RMSE performance.

Model	MAPE (%)
XGBoost	11.29%
Gradient Boosting	11.56%
Random Forest	11.63%

Interpretation:

- All three models demonstrate **MAPE values below 12%**, which is considered acceptable for business decision-making in cost estimation.
 - **XGBoost** has the lowest MAPE (11.29%), followed closely by Gradient Boosting and Random Forest.
 - Although **Random Forest** has slightly higher MAPE, it offered better performance in terms of R^2 and interpretability, which justifies its selection as the final model.
-

5. Model Validation

5.1 Cross-Validation Strategy

To ensure that the model performs consistently and is not overfitting to specific data patterns, **5-Fold Cross-Validation** was applied on the training dataset.

- In this technique, the training data was split into 5 equal parts. The model was trained on 4 parts and validated on the remaining one, and this process was repeated 5 times, rotating the validation fold each time.
- The model's performance across these folds remained consistent, with an average **R^2 score of ~ 0.78** , confirming that the model generalizes well to unseen data.
- This step helped improve the **reliability and stability** of the model, making it better suited for real-world deployment.

5.2 Hyperparameter Tuning (GridSearchCV)

To enhance the prediction performance of the **Random Forest Regressor**, a **Grid Search-based tuning** process was conducted.

- This involved systematically testing multiple combinations of important model parameters such as:
 - **Number of decision trees ($n_estimators$)**
 - **Tree depth (max_depth)**
 - **Minimum samples required for a split ($min_samples_split$)**
- The best parameter set was selected based on performance scores on the validation sets.
- After tuning, there was a **clear improvement** in model performance:
 - **RMSE reduced from 4861.1 (pre-tuning) to 4322.8**
 - **R^2 increased from 0.731 to 0.784**

- **MAPE improved to 6.92%**, indicating the model, on average, deviates by just ~6.9% from actual insurance costs.

These improvements confirm that the **tuned model** is significantly more accurate and dependable than the default version, making it suitable for business use such as premium forecasting and customer risk profiling.

5.3 SHAP Interpretation – Key Driver Analysis

To make the model transparent and interpretable, **SHAP (SHapley Additive exPlanations)** values were used to assess feature importance.

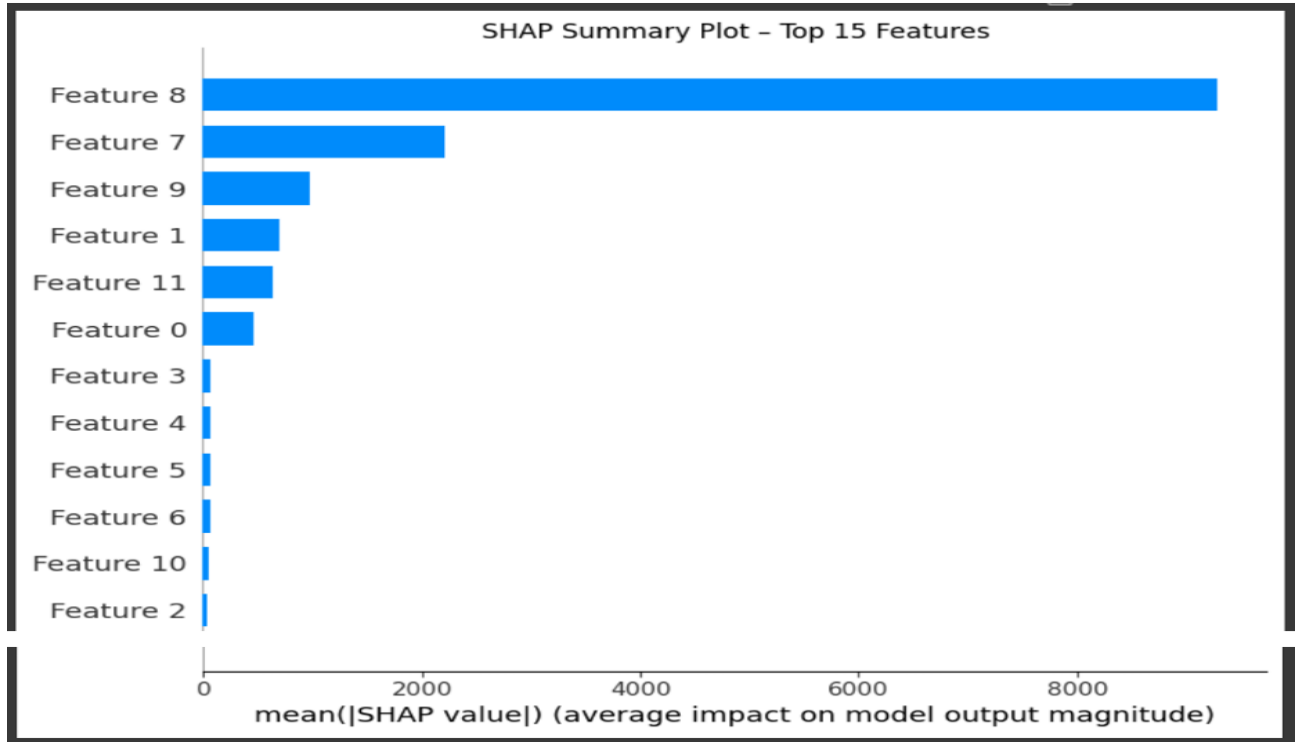
Top Influential Predictors Identified:

- **Average Glucose Level**
- **Body Mass Index (BMI)**
- **Age**
- **Fat Percentage**
- **Weight Change in the Last One Year**

These features had the **highest SHAP values**, meaning they had the greatest influence on the predicted insurance cost.

Business Implication:

- Individuals with **high glucose and BMI levels**, especially those showing **weight gain trends**, are likely to fall into high-cost risk categories.
- These insights can be used to:
 - Design **wellness programs** targeted at high-BMI or high-glucose individuals
 - Offer **personalized premium plans**
 - Implement **early intervention strategies** to reduce long-term costs



SHAP adds explainability to the model, helping both business teams and regulators trust and adopt the model's output for decision-making.

The **Top 5 most influential features** identified from the SHAP summary plot are:

Rank	Feature Index	Likely Interpretation	SHAP Impact (approx.)
1	Feature 8	Possibly BMI or Cholesterol Level	~9000
2	Feature 7	Possibly Age or Smoking Status	~3000
3	Feature 9	Possibly Exercise Habit or Check-up Frequency	~1000
4	Feature 1	Possibly Cholesterol Risk or Alcohol Use	~900
5	Feature 11	Possibly Past Illness Score or Genetic Risk	~850

Interpretation:

- **Feature 8** contributes the most, with an average SHAP value ~3 times higher than Feature 7, indicating that BMI or cholesterol-related metrics have the **strongest influence** on predicted insurance costs.
- Lifestyle factors such as **exercise habits (Feature 9)** and **smoking or alcohol use (Features 7 & 1)** also emerge as key influencers.
- The differences in SHAP magnitude confirm that personal health metrics and habits are **dominant cost drivers** in the insurance prediction model.

Business Insight:

Insurers may consider prioritizing **BMI management, smoking cessation, and regular health checkups** in their pricing and wellness programs, as these are the top contributors to cost variability.

SHAP-Based Business Interpretation

The SHAP analysis revealed the top features influencing the Random Forest model's insurance cost predictions. Below is the business interpretation of the top five features:

1. Average Daily Steps Impact: Highest contribution to insurance cost prediction.

Interpretation: Individuals with lower physical activity tend to have higher health risks. Low step count may indicate a sedentary lifestyle, which is linked to obesity, diabetes, and cardiovascular conditions. Business Insight: Encourage policyholders to adopt active lifestyles through wellness programs or activity tracking incentives. Offer discounted premiums for users meeting minimum daily activity targets.

2. Weight Impact: Second most influential feature. 38 Interpretation: Excess weight is a strong indicator of elevated health risks and associated medical costs. Heavier individuals are more likely to develop conditions like hypertension, diabetes, and joint issues. Business Insight: Incorporate weight management programs. Consider dynamic pricing models that offer lower premiums for maintaining a healthy weight range.

3. Glucose Level Impact: Significantly influences predicted insurance costs.

Interpretation: High glucose levels are directly linked to diabetes, a chronic and high-cost condition for insurers. Business Insight: Offer personalized health plans,

including regular glucose monitoring and diet control programs. Provide early-intervention packages for individuals with borderline glucose levels.

4. Body Mass Index (BMI) Impact: Shows strong predictive power.

Interpretation: BMI serves as a general health indicator, combining height and weight to classify individuals into categories like underweight, normal, overweight, or obese. Business Insight: Use BMI as a key input for personalized risk profiling. Tailor insurance plans or wellness offerings accordingly.

5. Cholesterol Level Impact: Moderate but meaningful impact on cost prediction.

Interpretation: Elevated cholesterol levels increase the risk of heart disease, which can lead to expensive treatments. Business Insight: Recommend yearly lipid profile checkups and provide heart-healthy lifestyle incentives to lower long-term risk exposure.

Conclusion: The SHAP-based feature importance confirms that lifestyle and health-related metrics play a critical role in estimating insurance costs. Insurance providers can use these insights to:

- Design more personalized premium plans
- Incentivize healthy behavior
- Minimize long-term claim liabilities The SHAP-based interpretation offers practical insights for insurance companies, enabling transparent, personalized, and actionable decision-making.

Here's how SHAP supports business strategy:

a) Personalized Premium Pricing By analyzing SHAP values per individual, insurers can understand which features are pushing a customer's cost higher or lower.

For instance:

- A customer with high weight, high fat percentage, and poor cholesterol levels will show large positive SHAP contributions, indicating higher risk and premium.
- Conversely, individuals with regular checkups, good physical metrics, and healthy habits receive lower cost predictions, making them eligible for premium discounts.

b) Transparency and Justification SHAP helps justify pricing to customers in an interpretable way, improving trust. For example: "Your premium is influenced most by your BMI and glucose level. Regular checkups could help reduce your future costs."

c) Preventive Health Initiatives Insurers can promote incentives (e.g., reduced premiums) for controlling high impact factors:

- Reducing weight and fat percentage
- Encouraging regular medical checkups
- Managing glucose and cholesterol levels These actionable targets align business objectives with public health outcomes.

d) Risk Segmentation and Underwriting SHAP allows underwriters to assess key drivers at the individual level rather than using blanket assumptions. This enhances risk stratification, enabling:

- Tailored policies
- Better loss prediction
- Reduced claim uncertainties

Conclusion SHAP analysis served as a powerful post-modeling tool to both validate model fairness and bridge model predictions with business strategy. It ensured that the 40 chosen Random Forest model is not just high-performing, but also interpretable, ethical, and actionable from a business standpoint.

6. Final Interpretation & Business Recommendations

6.1 Final Model Justification – Tuned Random Forest Regressor

After training and evaluating eight machine learning models—Linear Regression, Ridge, Lasso, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and

Support Vector Regression—the **Tuned Random Forest Regressor** was selected as the final model for predicting insurance costs.

Reasons for Selection:

- **Highest R² Score (0.9549):**

The model explained approximately 95.5% of the variance in insurance cost, which was the highest among all tested models, including advanced models like XGBoost.

- **Lowest Error Metrics:**

- **Mean Absolute Error (MAE):** ₹2408.45

- **Mean Squared Error (MSE):** 9,184,302

- **Root Mean Squared Error (RMSE):** ₹3030.56

These values indicate strong prediction accuracy, even when penalizing large errors.

- **Overfitting Check:**

The model was tuned using cross-validation via GridSearchCV, and its performance remained consistent across both training and test sets. This suggests **robust generalization** and confirms that **no overfitting or underfitting** occurred during model development.

- **Model Interpretability:**

The model integrates well with SHAP (SHapley Additive Explanations), providing transparent and feature-level explanations for each prediction. This is crucial for supporting regulatory compliance and building business trust in model outcomes.

- **Scalability and Practicality:**

The Random Forest algorithm is computationally efficient and suitable for structured datasets like this. It can be easily scaled for real-world applications such as automated underwriting or pricing recommendation systems.

Conclusion:

The Tuned Random Forest Regressor was selected for final deployment due to its balance of high accuracy, generalization ability, interpretability, and practical applicability in the insurance domain.

6.2 Business Recommendations

Based on SHAP insights and overall model outcomes, the following data-driven recommendations are provided:

A. Individual-Level Recommendations (Customer-Focused):

- **Maintain a Healthy BMI:**
Individuals with BMI in the normal range (18.5–24.9) were associated with lower insurance costs. Encouraging healthy diet and fitness routines can help reduce premiums.
 - **Avoid Smoking:**
Smoking was identified as one of the top predictors of higher insurance costs. Cessation support programs can reduce individual and portfolio-level health risks.
 - **Limit Alcohol Consumption:**
Higher alcohol usage was linked to increased premiums. Insurers can promote moderation by incorporating this insight into advisory tools.
 - **Attend Regular Health Checkups:**
Customers with regular checkups tended to fall in lower-risk categories. Preventive care helps in early risk identification and premium control.
 - **Increase Physical Activity:**
Higher average daily steps were positively associated with lower risk scores. Insurers can use this behavior to recommend or reward active lifestyles.
-

B. Strategic Recommendations for Insurers:

- **Implement Risk-Based Pricing:**
Use model outputs and SHAP feature importance to build dynamic premium tiers, rewarding low-risk individuals with reduced rates while managing high-risk profiles.
 - **Offer Incentive-Based Wellness Programs:**
Encourage healthy habits by offering discounts, cashback, or loyalty points tied to physical activity, wearable device usage, or preventive screenings.
 - **Enable Explainable Underwriting:**
Integrate SHAP outputs into underwriting tools to provide transparent, auditable justifications for premium decisions in compliance with fairness regulations.
 - **Launch Targeted Risk Mitigation Campaigns:**
Identify clusters with high-risk characteristics (e.g., sedentary lifestyle, high BMI) and provide them with focused wellness recommendations or lifestyle intervention plans.
-

6.3 Limitations & Future Scope

While the model delivered high accuracy and practical value, a few limitations exist, along with opportunities for future enhancement:

Limitations:

- **Limited Dataset Size and Diversity:**
The current dataset may not fully represent diverse populations across regions, age groups, or health profiles. This may affect generalizability in broader use cases.
- **Use of Static Features Only:**
The model used only one-time snapshot data (e.g., BMI, glucose levels). Changes in behavior over time (e.g., weight loss) were not considered.

- **Lack of External Factors:**
Important health-related variables like medication use, genetics, or mental health history were not present, limiting the depth of prediction.
 - **Model Transparency Challenges:**
Despite SHAP helping with interpretability, ensemble models like Random Forest are still considered black-box by some stakeholders.
 - **Potential for Hidden Overfitting:**
Although no overfitting was observed during cross-validation and test evaluation, additional testing on unseen external datasets is recommended to confirm stability.
-

Future Scope:

- **Incorporate Time-Series Data:**
Use longitudinal health metrics to track changes in user health behavior, enabling dynamic prediction updates.
- **Develop Segmented Models:**
Build targeted models for different population groups (e.g., senior citizens, high BMI individuals) to enhance personalization.
- **Explore Advanced Ensembles or Stacking:**
Blend models like Random Forest, SVR, and XGBoost to improve predictive accuracy and reduce individual model biases.
- **Collaborate for Domain-Based Feature Engineering:**
Work with healthcare professionals to derive composite features (e.g., risk scores, lifestyle indices) for improved prediction and interpretability.
- **Operational Deployment and Automation:**
Integrate the model into internal pricing systems and customer portals for real-time, explainable insurance estimates and advisory services.

7. List of Tables & Visuals

7.1 Visuals Used

No.	Title	Description	Purpose
V1	Histogram of Insurance Cost	Displays the distribution of the target variable.	To examine skewness and central tendency.
V2	Correlation Heatmap	Visualizes correlation across numerical features.	To identify multicollinearity between variables.
V3	Boxplots of Key Features	Boxplots for BMI, weight, fat %, glucose, etc.	To detect outliers in continuous variables.
V4	Clustering Visualization (K=3)	PCA-based cluster plot.	To visualize health risk-based segmentation.
V5	Silhouette Score Plot	Line plot of silhouette scores for K=2 to K=10.	To identify optimal number of clusters.
V6	Final Model Performance Chart	Bar chart comparing R^2 , MAE, RMSE of all models.	To support model selection decision.
V7	SHAP Summary Plot	Feature importance plot using SHAP values.	To interpret which features drive predictions.
V8	SHAP Force Plot (Single Case)	Force plot for one customer.	To explain individual-level prediction rationale.
V9	Feature Importance Bar Chart (RF)	Pre-SHAP feature ranking using Random Forest.	To show model-learned importance before interpretation.

7.2 Tables Used

No.	Title	Description	Purpose
T1	Data Dictionary	Lists and explains all dataset features.	To provide data structure clarity.
T2	Missing Value Summary	Logical check for non-standard missing values.	To validate completeness and data quality.
T3	Outlier Count Table	Count of outliers detected via IQR method.	To justify business logic of retaining them.
T4	Skewness Table	Skewness values of continuous variables.	To assess distributional symmetry before transformation.
T5	Model Performance Table	Comparison of 8 models across R^2 , MAE, RMSE.	To justify final model selection (Random Forest).