Project 2: Insurance Premium Classification & Modeling

Jadon Chanthavong

9/29/2025

Dataset Link: https://www.kaggle.com/datasets/mosapabdelghany/medical-insurance-cost-dataset/data

Introduction to Data: This dataset contains 1,338 records with patient demographics (age, sex, BMI, number of children, smoking status, region) and their corresponding insurance charges.

Introduction of Problem: For this classification project, I will analyze medical insurance cost data to predict risk categories for insurance customers.

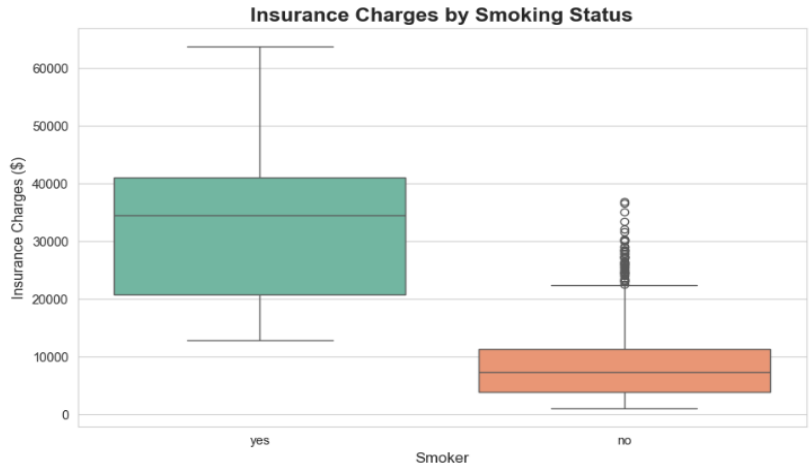
What categories do these individuals fall in and how does that have an effect on their insurance premiums?

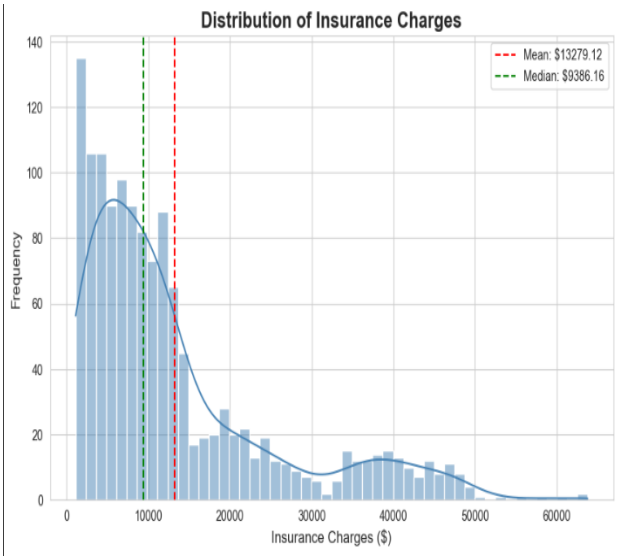
Classification Target: I plan to create three risk categories:

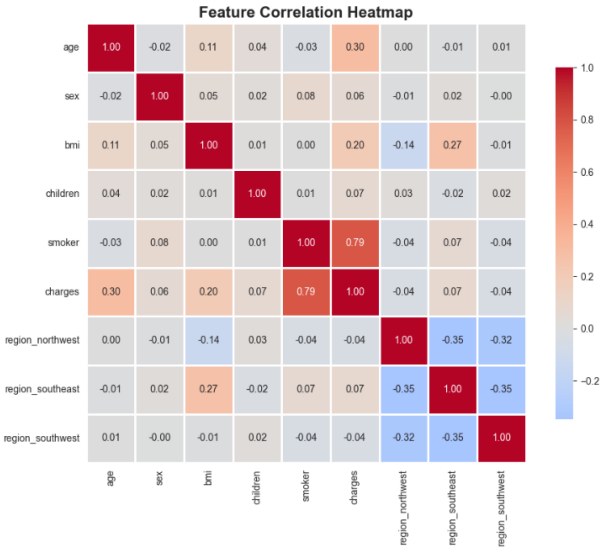
- Low Risk: Bottom percentile of insurance costs (lowest 33%)
- Medium Risk: Middle percentile of insurance costs (middle 33%)
- High Risk: Top percentile of insurance costs (highest 33%)

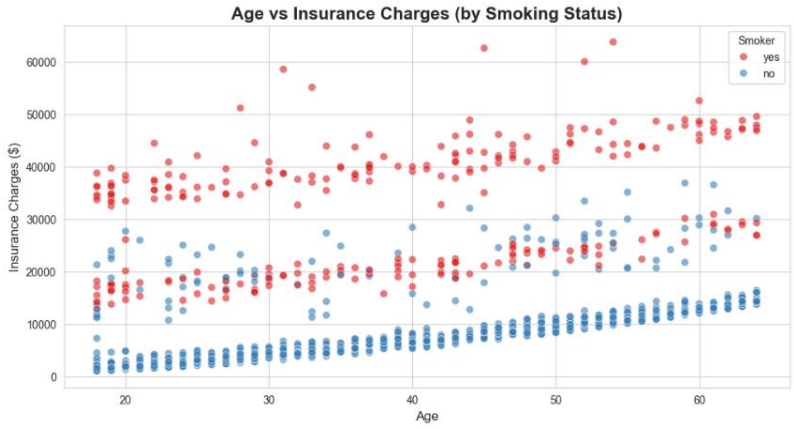
Pre-processing the Data: Data preprocessing serves as a critical foundation that ensures the integrity and reliability of machine learning models by identifying potential data quality issues that could compromise classification accuracy and lead to misleading conclusions about insurance risk factors. Without preprocessing validation, analysts risk training models on corrupted data where missing values could introduce bias, duplicate records could skew feature importance calculations, and inconsistent categorical encodings could prevent algorithms from learning meaningful patterns. The preprocessing assessment of this insurance dataset revealed exceptional data quality with zero missing values across all 1,338 records, no duplicate entries, and realistic ranges for all variables (ages 18-64, BMI 15.5-53.1, charges $1,121-$63,770), confirming the dataset's readiness for analysis. However, preprocessing also revealed critical insights that shaped our modeling approach: the dramatic right-skew in insurance charges (mean $13,270 exceeding median $9,386) validated our decision to use tercile-based risk categories rather than equal-dollar bins, ensuring balanced classes for training, while the vastly different feature scales (age: 18-64, BMI: 15-50, charges: $1,000-$64,000) necessitated StandardScaler application to prevent larger-valued features from dominating model training. This thorough data validation provided confidence that our models' identification of smoking status as the dominant predictor which reflected genuine risk patterns rather than data artifacts, enabling reliable risk stratification that could inform insurance pricing strategies and health intervention programs while raising important ethical questions about behavior-based premium calculations.

Data Visualizations:









How does this step relate to the modeling?

The exploratory data analysis directly informed our modeling decisions in three critical ways:

Identified Key Predictors: The box plot and correlation heatmap revealed smoking status as the dominant feature (correlation of 0.79 with charges), followed by age and BMI. This told us to expect high model accuracy but potential over-reliance on smoking status alone.

Justified Preprocessing Choices: The right-skewed distribution histogram validated our tercile-based risk categories, ensuring balanced classes (33% each) rather than using equal-dollar bins. Additionally, visualizations showing vastly different feature scales (age: 18-64, BMI: 15-50, charges: $1K-$64K) confirmed the need for StandardScaler to prevent larger values from dominating model training.

Revealed Feature Interactions: The age vs. charges scatter plot showed that age affects smokers and non-smokers differently; smokers have consistently high costs regardless of age, while non-smokers show gradual cost increases with age. This interaction pattern suggested our models would capture distinct risk profiles for these subgroups.

In summary, visualization shaped our preprocessing strategy, set performance expectations (>80% accuracy anticipated), and revealed why certain features matter more than others. This prevents us from building models blindly.

Model Selection and Rationale

For this classification problem, I implemented three algorithms to compare performance and interpretability:

1. Logistic Regression

A linear classification algorithm that predicts probabilities using a sigmoid function to map inputs to risk categories.

2. Random Forest Classifier

An ensemble method that builds multiple decision trees and aggregates their predictions through majority voting.

3. Gradient Boosting Classifier

An ensemble technique that builds trees sequentially, with each new tree correcting errors from previous ones.

All three models were trained on identical preprocessed data (encoded, scaled, 80/20 split) to ensure fair comparison. This multi-model approach allows us to balance interpretability (Logistic Regression), robustness (Random Forest), and maximum performance (Gradient Boosting).

Takeaways: The Dominant Discovery: Smoking Status Reigns Supreme

The most striking finding emerged immediately during exploratory analysis and was reinforced throughout modeling: smoking status is not just a factor in insurance costs; it's the MOST significant factor. Smokers average $32,050 in annual charges compared to non-smokers' $8,434, a nearly 4x difference.

This finding has profound implications: the traditional view of insurance risk as a complex interplay of multiple demographic and health factors is oversimplified. In reality, smoking status creates two almost entirely separate populations within the dataset. Even young, healthy smokers fall into the high-risk category, while older non-smokers with elevated BMI often remain in low or medium risk tiers.

Model Performance: High Accuracy with Expected Patterns

All three models achieved strong performance. This high success rate validates our tercile-based classification approach and confirms that the features we selected effectively predict risk categories. However, the confusion matrices revealed an important nuance: models occasionally confused Medium and High risk categories, which makes intuitive sense given that both groups may include smokers with varying ages and BMI levels. The clear separation occurred primarily between smokers (High risk) and non-smokers (Low/Medium risk).

Secondary Factors: Age and BMI Matter Differently by Group

While smoking dominates overall, our scatter plot analysis and model predictions revealed that age and BMI play different roles for smokers versus non-smokers. Among non-smokers, age shows a gradual positive relationship with costs where older individuals in particular face incrementally higher premiums. For smokers, costs start high regardless of age, with only modest increases over time. This suggests that insurance companies treat smoking as an immediate, severe risk factor that overshadows typical age-related cost increases.

BMI showed a similar pattern: it correlates moderately with costs for non-smokers but adds relatively little predictive power for smokers, whose baseline risk is already elevated. This interaction effect suggests that traditional health metrics (age, BMI) matter most when the individual doesn't smoke; once smoking enters the equation, it dominates the risk calculation.

Our classification system successfully stratifies customers into actionable risk tiers:

- Low Risk (33%): Primarily non-smokers with lower-than-median costs who are ideal candidates for standard premium rates
- Medium Risk (33%): Mixed group of non-smokers with age/BMI concerns and occasional younger smokers that require moderate premium adjustments
- High Risk (33%): Predominantly smokers regardless of other factors which warrants significantly higher premiums or targeted intervention programs

This stratification enables insurance providers to quickly identify which customers need specialized attention. However, it also raises an important consideration: if smoking is this

dominant, should insurers invest more in smoking cessation programs rather than simply charging higher premiums? The cost difference ($24,000 annually) suggests that supporting smokers in quitting could be both ethically sound and financially beneficial.

Answering the Initial Question

Yes, I successfully answered my core question. Individuals fall into risk categories primarily based on smoking status, with age and BMI serving as secondary modifiers. The effect on premiums is dramatic and non-linear: crossing the smoker/non-smoker boundary has a far greater impact than any other demographic or health factor in the dataset. Our models demonstrate that insurance companies can accurately predict and justify premium differences using a relatively simple set of features, with smoking status alone providing the majority of predictive power.

The Unexpected Insight

Perhaps the most surprising finding was how little other factors mattered once smoking was accounted for. Region, sex, and number of children showed negligible correlation with costs and minimal importance in model predictions. This suggests that insurance risk, at least in this dataset, is far more about personal health choices (smoking) and natural aging than about demographic characteristics or family circumstances. This insight challenges the complexity often assumed in insurance pricing models and suggests that simpler, more transparent pricing structures focused on modifiable risk factors might be both fairer and more effective.

Impact

Potential Positive Impacts:

- Transparent Risk-Based Pricing: The findings could help insurance companies justify premium differences with clear, data-driven evidence. Showing that smoking accounts for a 4x cost increase provides objective rationale for higher premiums, potentially reducing disputes and making pricing more transparent for consumers who understand exactly why their rates differ.
- Targeted Health Intervention Programs: Identifying smoking as the dominant cost driver could shift insurance company strategies from simply charging more to actively supporting smoking cessation. The $24,000 annual cost difference creates a strong financial incentive for insurers to invest in quit-smoking programs, nicotine replacement therapy coverage, and behavioral health support; these are interventions that benefit both the company's bottom line as well as customer health.
- Simplified Risk Assessment: Understanding that age and BMI matter primarily for non-smokers could streamline the insurance application process. Rather than collecting extensive health data that ultimately has minimal predictive power, companies could focus on the few factors that actually matter, reducing paperwork burden and processing time for customers.

Potential Negative Impacts:

- Discrimination Against Addiction: While smoking is technically a "choice," nicotine addiction is a medically recognized condition often rooted in socioeconomic factors, mental health issues, and childhood exposure. Using smoking status to categorize people as "high risk" could unfairly penalize individuals struggling with addiction, making healthcare less accessible to those who may need it most. This creates a cruel cycle: those with addiction pay more, have less money for treatment, and remain trapped in high-risk categories.
- Privacy and Honesty Concerns: If smoking status becomes the primary determinant of insurance costs, individuals might be incentivized to hide their smoking habits or provide false information on applications. This could lead to insurance fraud, strained trust between insurers and customers, and potential policy cancellations if dishonesty is discovered; people could suddenly be uninsured when they need coverage at their most dire times.
- Oversimplification of Health: This analysis treats smoking as a binary yes/no variable, ignoring critical nuances like former smokers, frequency of smoking, vaping versus traditional cigarettes, and secondhand smoke exposure. Someone who smoked briefly years ago might be categorized identically to a pack-a-day smoker, leading to unfair premium assignments. The model also completely ignores mental health, family medical history, occupation hazards, and exercise habits which are significant factors that impact health outcomes but aren't captured in this dataset.

Visualization #1: Box Plot of Insurance Charges vs. Smoking Status

Figure 1 below examines the relationship between smoking status and insurance charges using a box plot visualization. This chart displays the median, quartiles, and outliers for insurance costs within each smoking category. After analyzing this plot, a dramatic conclusion emerged: smokers consistently show significantly higher median charges (approximately $34,000) compared to non-smokers (approximately $7,000) - nearly a 4x difference. This addresses the core research question about whether smoking status is the primary cost driver. The visualization reveals minimal overlap between the two groups, with even the lowest-cost smokers paying more than the median non-smoker, demonstrating that smoking is the single most influential factor in insurance pricing.

Visualization #2: Histogram of Insurance Charges Distribution

Figure 2 is a histogram displaying the distribution of insurance charges across all customers, which answers the question of the overall cost patterns in the dataset. The histogram effectively shows the shape and spread of insurance costs by displaying frequency bins. After examining this plot, key conclusions include that the distribution is heavily right-skewed, with most customers (approximately 60%) clustered in the lower cost range (under $10,000), while a smaller segment experiences dramatically higher charges (up to $64,000). The mean ($13,270)

exceeds the median ($9,386), confirming this positive skewness. This graph displays a clear pattern that justifies using tercile-based risk categories rather than equal-dollar bins, ensuring balanced classification groups despite the skewed underlying cost structure.

Visualization #3: Correlation Heatmap of Key Variables

Figure 3 below is a correlation heatmap showing the relationships between demographic, health, and cost variables including age, BMI, sex, smoking status, number of children, region, and insurance charges. This visualization uses color coding to display correlation strengths, with darker colors indicating stronger relationships. The heatmap reveals that smoking status has by far the strongest correlation with charges (0.79), followed by age (0.30) and BMI (0.20). Sex, children, and region show negligible correlations (all under 0.15). This addresses the question of which factors most influence insurance costs. The visualization demonstrates that while multiple demographic factors exist in the dataset, smoking status stands as the overwhelmingly dominant predictor, with age and BMI serving as secondary factors that operate within the broader smoking/non-smoking divide.

Visualization #4: Scatter Plot of Age vs. Charges by Smoking Status

Figure 4 below is a scatter plot examining the relationship between age and insurance charges, with smoking status indicated by color coding. This visualization displays each of the 1,338 customers as individual points to show the distribution of costs across age groups and smoking categories. After analyzing this plot, several important conclusions emerged: two distinct populations are clearly visible, with smokers (red points) forming a band between $12,000-$64,000 regardless of age, while non-smokers (blue points) show a more gradual upward trend from $1,000-$37,000 as age increases. This reveals that age's impact on costs differs dramatically by smoking status - for non-smokers, aging is associated with steadily increasing costs, while for smokers, costs start high even at young ages and increase more modestly. The separation between these groups suggests that a unified model may mask important subgroup patterns, potentially justifying separate analyses for smokers and non-smokers to capture nuanced risk factors beyond the smoking effect alone.