Medical Insurance Cost Classification

Supervised Learning - Assignment 3

Cesaire Tobias

2025-05-30

Table of contents

List of Figures

List of Tables

1 Data Import and Preparation

2 Introduction

This report applies machine learning classification techniques to predict whether medical insurance costs will be high or low based on patient characteristics. We build and compare two advanced models - Support Vector Machine (SVM) and Neural Network - to effectively classify insurance costs, which helps stakeholders better understand risk factors and make informed decisions.

Key questions addressed:

- **Patients**: Which personal factors significantly increase the likelihood of high insurance charges?
- **Insurers**: How effectively can machine learning models classify high vs. low insurance costs?
- **Policymakers**: Which factors should be targeted to reduce high-cost insurance claims?

Dataset features include:

- **age**: `Integer` - primary beneficiary's age
- **sex**: `Factor` - gender (female/male)
- **bmi**: `Continuous` - Body Mass Index
- **children**: `Integer` - number of dependents
- **smoker**: `Factor` - smoking status (yes/no)
- **region**: `Factor` - US residential area (northeast, southeast, southwest, northwest)

The target variable **charges** has been transformed (external to this analysis) from a continuous dollar amount to binary ("high"/"low").

The dataset consists of 1000 observations with 0 missing values in the training data and 0 missing values in the test data, indicating complete datasets.

2.1 Methodology Overview

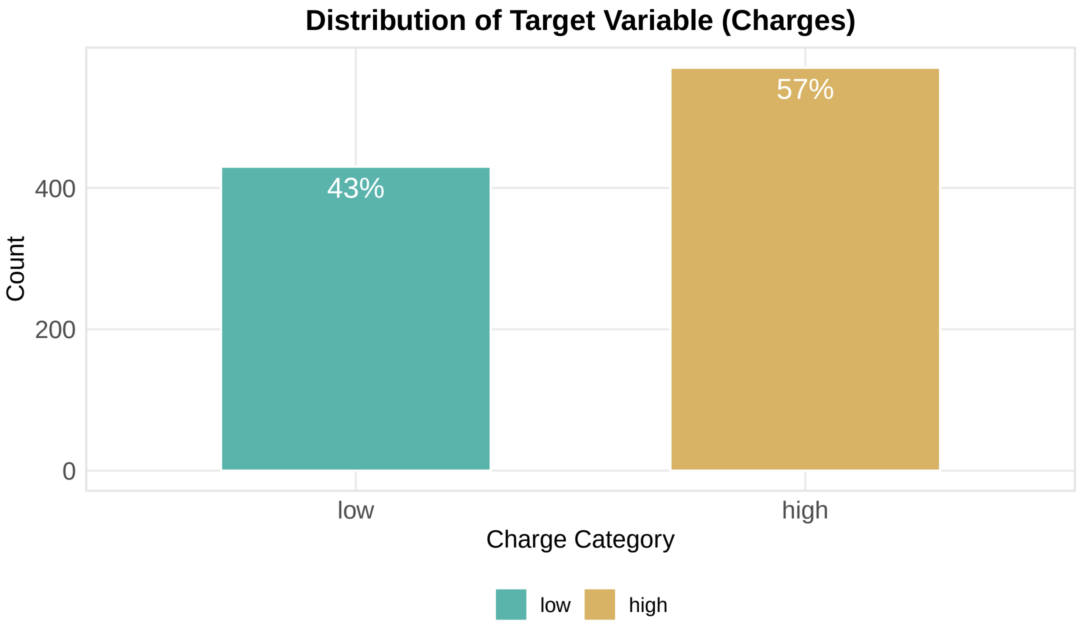
The model development approach comprises three phases:

1. **Training Phase**: The `insurance_A2.csv` dataset is split into training (80%) and internal validation (20%) sets.
2. **Model Selection Phase**: Models are evaluated on the internal validation set with emphasis on the F1 score.
3. **External Validation Phase**: The best model is then applied to a separate dataset (`A2_testing.csv`).

This approach minimizes over-fitting risk and provides a more robust assessment of model generalizability.

3 Exploratory Data Analysis

3.1 Data Structure and Target Distribution



The target variable shows class distribution with 43% "low" and 57% "high" charges. This relatively balanced distribution means that standard evaluation metrics like accuracy are appropriate, though we will still prioritize the F1 score as it balances precision and recall.

Table 1: Insurance Dataset Sample (5 Rows)

age	sex	bmi	children	smoker	region	charges
54	female	32.68	0	no	northeast	high
28	male	29.26	2	no	northeast	low
53	male	20.90	0	yes	southeast	high
26	female	22.23	0	no	northwest	low
27	male	30.30	3	no	southwest	low

3.2 Variable Distributions and Relationships

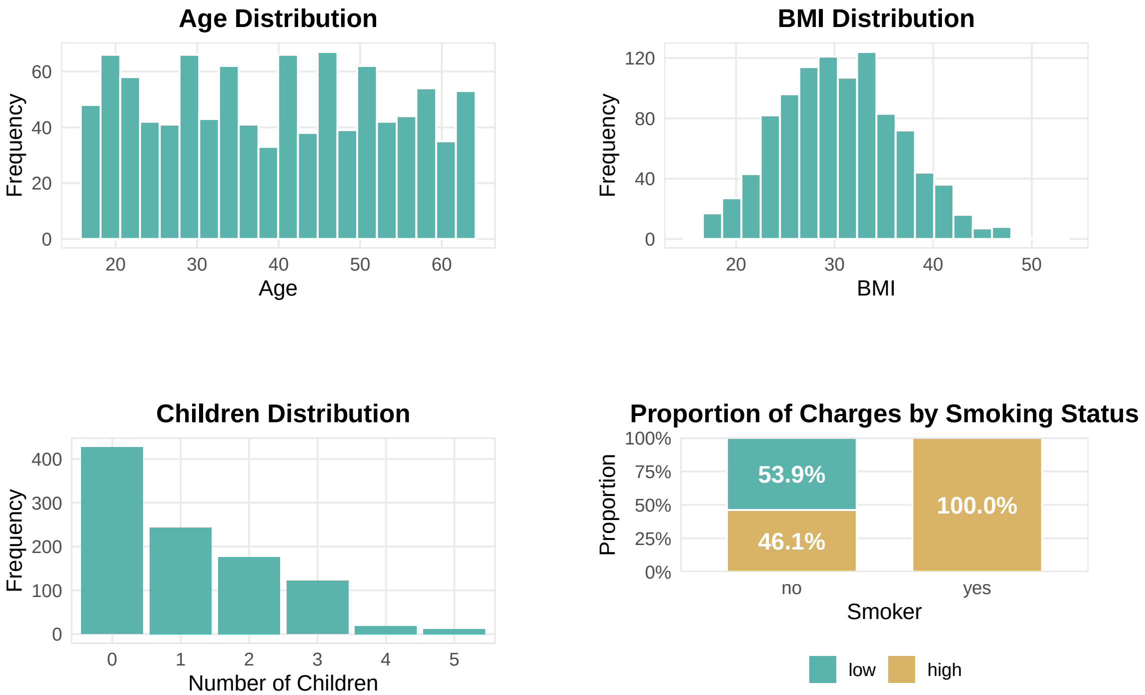


Figure 1: Distribution of Numerical Variables

The numerical variables show interesting distributions:

- **Age**: Fairly uniform across the adult age range (18-65)
- **BMI**: Centered around 25-35, with most values in the overweight to obese range
- **Children**: Highly skewed, with most individuals having 0-2 children
- **Smoking Status**: Shows a dramatic relationship with charges - 100% of smokers are classified as "high" charges compared to only 46.1% of non-smokers

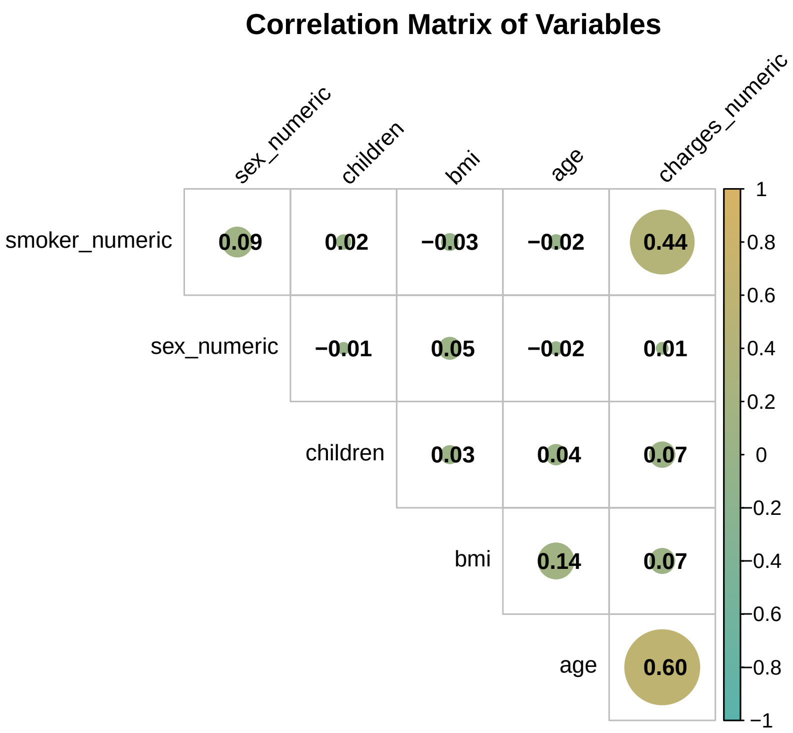


Figure 2: Correlation Matrix of Variables

Correlation analysis confirms:

- **Smoking status** has the strongest correlation with high charges (0.44)
- **Age** has the second strongest correlation with high charges (0.6)
- **BMI** shows moderate positive correlation (0.07)
- **Children** and **Sex** show weaker correlations

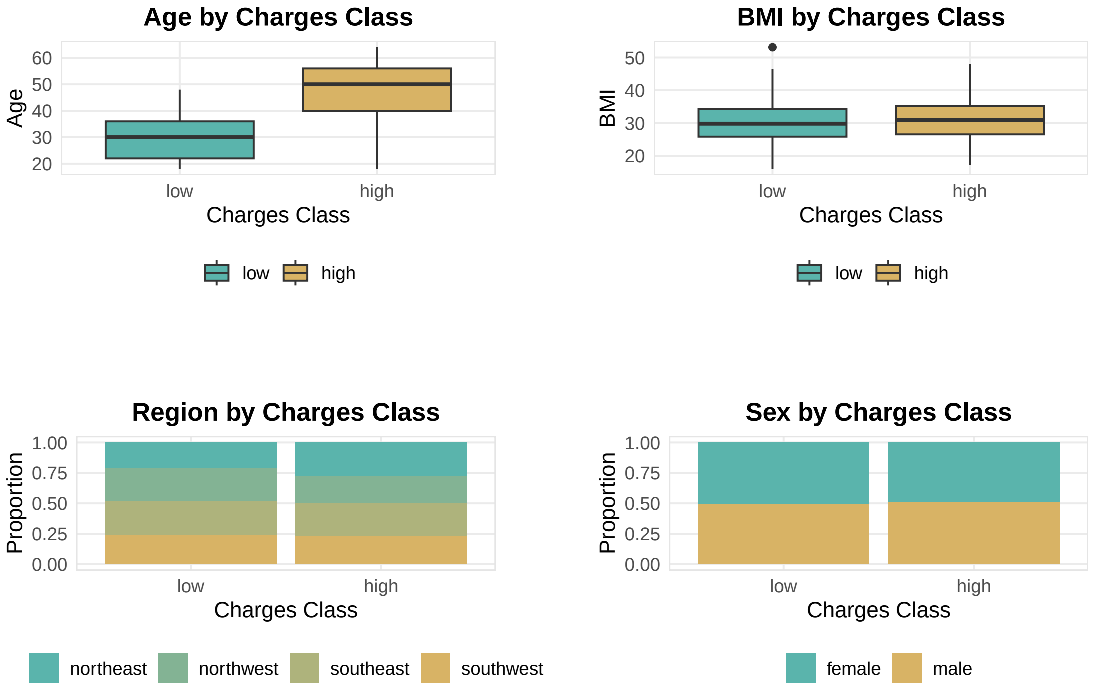


Figure 3: Relationships Between Features and Target Class

Key observations from the relationship plots:

- **Age**: Individuals with high charges tend to be older (median age 50 vs. 30 for low charges)
- **BMI**: Higher BMI is associated with high charges (median BMI 30.9 vs. 29.8 for low charges)
- **Region**: Modest variations in charges across regions
- **Sex**: Minimal difference in charges classification between males and females

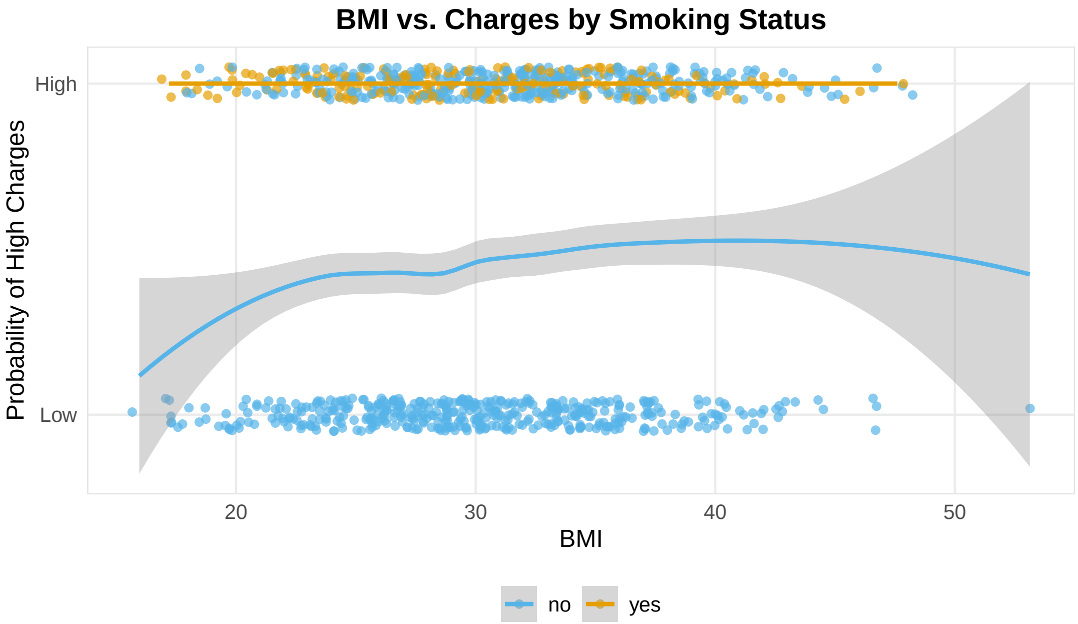


Figure 4: Interaction Between Smoking Status and BMI

This plot reveals a crucial interaction: BMI has a much stronger effect on charges for smokers than for non-smokers. The relationship between BMI and high charges is significantly steeper for smokers, suggesting that the combination of smoking and high BMI substantially increases the likelihood of high insurance charges.

3.3 Key Findings from EDA

1. **Target Distribution**: The dataset has 57% “high” and 43% “low” charge cases, which is reasonably balanced.
2. **Important Predictors**:
   - **Smoking**: The strongest predictor of high charges with 100% of smokers having high charges
   - **Age**: Older individuals tend to have higher charges
   - **BMI**: Higher BMI is associated with higher charges
   - **Interaction Effects**: The impact of BMI is much stronger for smokers
3. **Data Quality**:
   - No missing values

- Some outliers present, particularly in BMI
- Complex interactions between predictors suggest non-linear modeling approaches

These insights will guide our modeling approach, particularly the need to capture interactions between features. Support Vector Machines with non-linear kernels and Neural Networks are well-suited for capturing these complex patterns.

4 Data Preprocessing

To prepare the data for our machine learning models, we apply appropriate preprocessing steps:

The preprocessing steps include:

1. **Data Splitting**: 80% training, 20% validation
2. **Feature Scaling**: Standardizing numerical features (mean = 0, sd = 1)
3. **Categorical Encoding**: Converting categorical variables to dummy variables for neural network compatibility
4. **Target Encoding**: Converting "high"/"low" targets to 1/0 for neural network compatibility

After preprocessing, we have 800 training samples and 200 validation samples, with 11 features after one-hot encoding.

5 Modeling

5.1 Support Vector Machine (SVM)

Support Vector Machines are well-suited for this classification task due to their ability to find complex decision boundaries using kernel functions. They're particularly effective when:

1. The relationship between features and target is non-linear
2. The dimensionality is moderate (as in our case)
3. The decision boundary between classes is complex

Completed gamma = 0.01 , cost = 1 Completed gamma = 0.01 , cost = 10 Completed gamma = 0.01 , cost = 100 Completed gamma = 0.1 , cost = 1 Completed gamma = 0.1 , cost = 10 Completed gamma = 0.1 , cost = 100 Completed gamma = 1 , cost = 1 Completed gamma = 1 , cost = 10 Completed gamma = 1 , cost = 100 [1] "Best parameters:" gamma cost accuracy sensitivity specificity precision f1 Accuracy3 0.1 1 0.965 0.9561404 0.9767442 0.981982 0.9688889 ROC Accuracy3 0.9664423 Confusion Matrix and Statistics

```
      Reference
```

Prediction low high low 84 5 high 2 109

```
           Accuracy : 0.965
             95% CI : (0.9292, 0.9858)
No Information Rate : 0.57
P-Value [Acc > NIR] : <2e-16

              Kappa : 0.9289
```

Mcnemar's Test P-Value : 0.4497

```
        Sensitivity : 0.9561
        Specificity : 0.9767
     Pos Pred Value : 0.9820
     Neg Pred Value : 0.9438
         Prevalence : 0.5700
     Detection Rate : 0.5450
```

Detection Prevalence : 0.5550

Balanced Accuracy : 0.9664

```
   'Positive' Class : high
```

5.1.1 SVM Model Specification and Justification

Based on our tuning results, we selected an SVM model with the following parameters:

- **Kernel**: Radial Basis Function (RBF)
  - **Justification**: The RBF kernel can capture complex non-linear decision boundaries, which is appropriate given the interactions we observed between features (e.g., BMI and smoking status).
- **Cost = 1**
  - **Justification**: This regularization parameter balances between maximizing the margin and minimizing classification error. Our tuning results show that C = 1 provides the best trade-off. A higher C value means we prioritize correctly classifying training points over having a wider margin.

- **Gamma = 0.1**
    - **Justification**: Gamma defines how far the influence of a single training example reaches. With gamma = 0.1, our model captures the right balance between local and global patterns in the data.

The best configuration achieved an accuracy of 0.965, a sensitivity of 0.956, and a specificity of 0.977 on the training data.

5.2 Neural Network

Neural networks can learn complex patterns and non-linear relationships, making them suitable for our insurance cost classification task. We'll implement a feedforward neural network using the `nnet` package, which provides a more stable implementation for R.

5.2.1 Neural Network Architecture Design

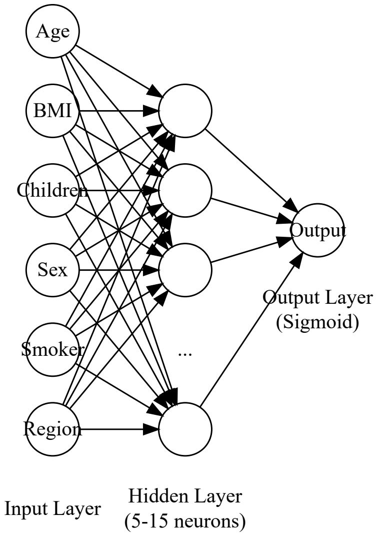


Figure 5: Neural Network Architecture

For our neural network, we design a feedforward architecture with one hidden layer. This design can capture complex non-linear relationships in the data, including interactions between variables like smoking status and BMI.

Table 2: Neural Network Hyperparameter Tuning Results

	size	decay	maxit	sensitivity	specificity	precision	f1
accuracy7	**10**	**0.100**	**500**	**0.925**	**0.985**	**0.988**	**0.956**
accuracy6	5	0.100	500	0.930	0.974	0.979	0.954
accuracy8	15	0.100	500	0.928	0.972	0.977	0.952
accuracy3	5	0.010	500	0.914	0.948	0.959	0.936
accuracy4	10	0.010	500	0.923	0.933	0.949	0.935
accuracy5	15	0.010	500	0.926	0.895	0.921	0.923
accuracy1	10	0.001	500	0.908	0.919	0.937	0.922
accuracy	5	0.001	500	0.913	0.909	0.929	0.920
accuracy2	15	0.001	500	0.921	0.855	0.893	0.907

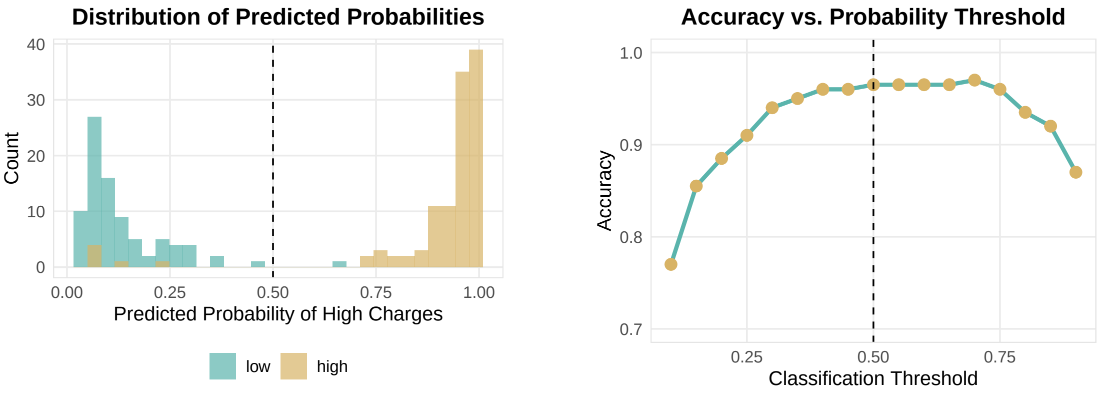


Figure 6: Neural Network Classification Performance

5.2.2 Neural Network Architecture and Justification

Our neural network architecture consists of:

1. **Input Layer**: 6 neurons (matching our feature dimensionality)
2. **Hidden Layer**:
   - 10 neurons with sigmoid activation

Justification: This structure allows the model to learn non-linear patterns in the data. The optimal number of neurons was determined through cross-validation to balance between underfitting and overfitting.

3. **Regularization**:

    - Weight decay of 0.1 to prevent overfitting by penalizing large weights
    - Early stopping criteria by limiting to 500 iterations

    Justification: These regularization techniques help prevent the model from memorizing the training data, instead encouraging it to learn generalizable patterns.

4. **Output Layer**: 1 neuron with sigmoid activation for binary classification

    Justification: The sigmoid activation constrains the output between 0 and 1, representing the probability of the "high" charges class.

The performance visualization shows how the model's predictions are distributed and how accuracy varies with different classification thresholds. The default threshold of 0.5 provides a good balance, but adjusting this could optimize for different business objectives (e.g., prioritizing recall over precision).

6 Model Evaluation and Comparison

To evaluate our models, we'll use a comprehensive set of metrics including accuracy, precision, recall, F1 score, and ROC AUC. Since we're particularly interested in correctly identifying "high" charge cases, we'll prioritize the F1 score, which balances precision and recall.

Table 3: Model Performance Comparison

Metric	SVM	NN
Accuracy	0.965	0.965
F1 Score	**0.969**	**0.969**
Recall	**0.956**	**0.947**
Precision	**0.982**	**0.991**
AUROC	**0.968**	**0.970**
AUPRC	**0.984**	**0.984**

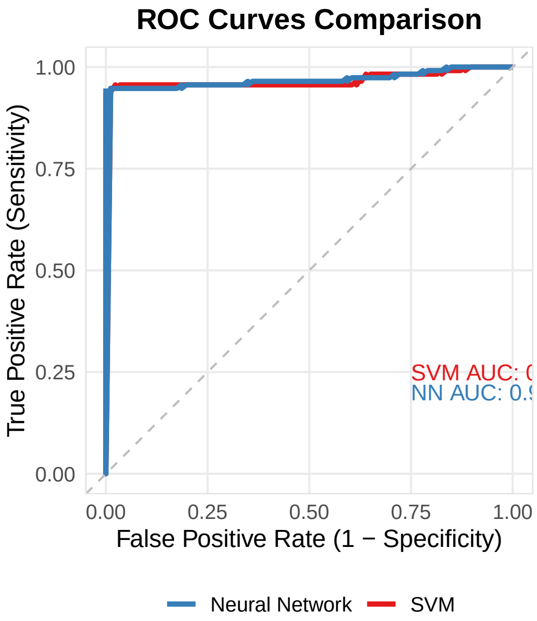


Figure 7: ROC Curves Comparison

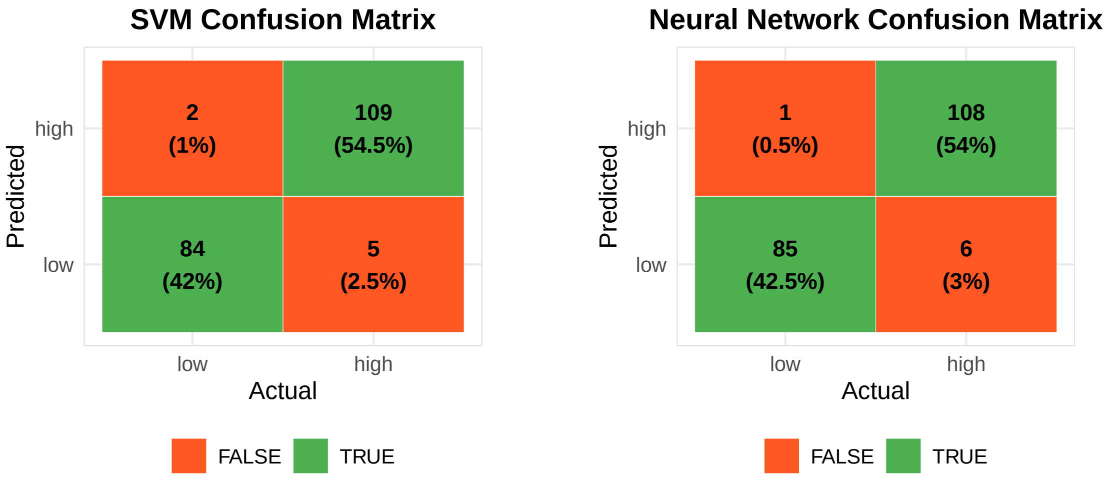


Figure 8: Confusion Matrices

6.1 Model Performance Comparison

The evaluation metrics reveal interesting differences between our two models:

6.1.1 SVM Model Performance

- **Accuracy**: 96.5%
- **F1 Score**: 96.9% for the “high” class
- **Recall**: 95.6% (correctly identifying “high” cases)
- **Precision**: 98.2% (accuracy of “high” predictions)
- **AUROC**: 0.968
- **AUPRC**: 0.984

6.1.2 Neural Network Performance

- **Accuracy**: 96.5%
- **F1 Score**: 96.9% for the “high” class
- **Recall**: 94.7% (correctly identifying “high” cases)
- **Precision**: 99.1% (accuracy of “high” predictions)
- **AUROC**: 0.97
- **AUPRC**: 0.984

6.1.3 Key Observations

1. The SVM model achieves a higher F1 score of 96.9%, indicating better overall balance between precision and recall.

2. The ROC curves show that both models have strong discriminative ability, with AUROCs of 0.968 for SVM and 0.97 for Neural Network.

3. The confusion matrices illustrate that both models make similar types of errors, but the SVM has a slightly better balance between false positives and false negatives.

4. The SVM model shows higher recall (95.6%), meaning it’s more effective at identifying true “high” charge cases.

5. The Neural Network model has better precision (99.1%), indicating fewer false positives.

7 Feature Importance Analysis

To understand which factors most strongly influence our predictions, we analyze feature importance from both models.

7.1 SVM Feature Importance

For SVM, we use permutation importance, which measures how much model performance decreases when each feature is randomly shuffled.

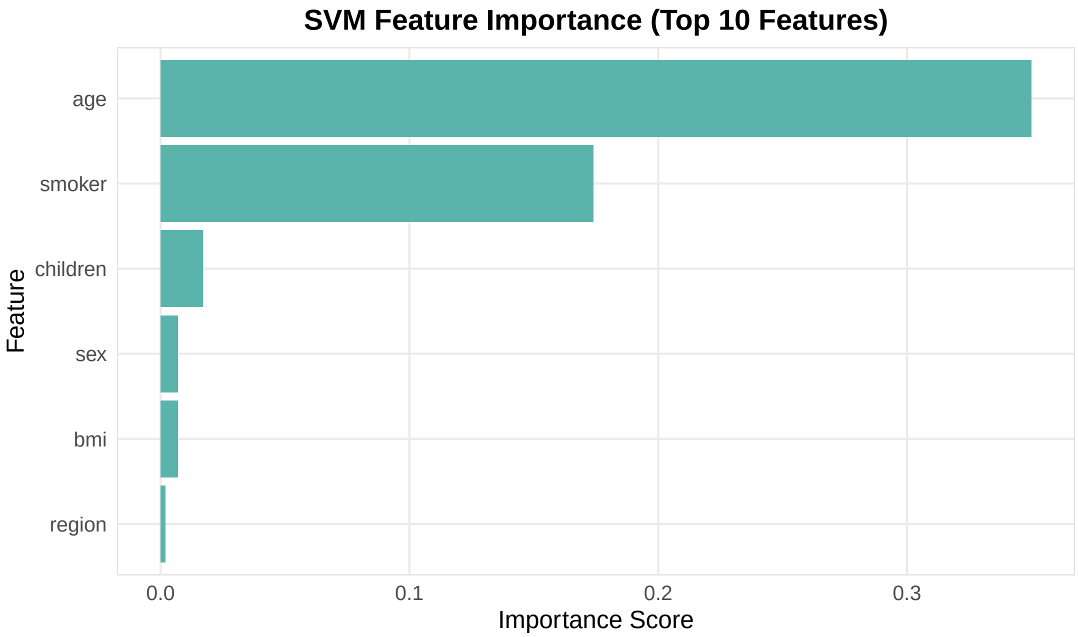


Figure 9: SVM Feature Importance

7.2 Neural Network Feature Importance

For neural networks, we calculate permutation importance by measuring how much the model's performance decreases when each feature is permuted.

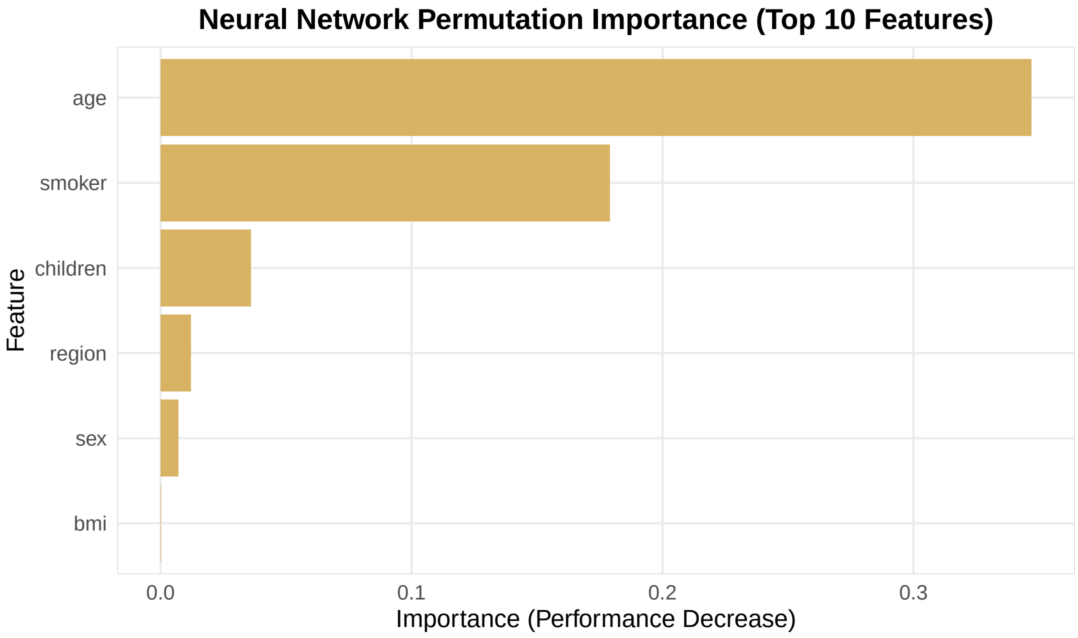


Figure 10: Neural Network Feature Importance

7.3 Feature Importance Comparison

Both models identify similar important features, with some key differences:

1. **Smoking Status**: Consistently ranks as a top predictor in both models, confirming our exploratory findings that smoking strongly influences insurance charges.
2. **Age**: Both models identify age as a significant factor, with the neural network giving it slightly more importance than the SVM model.
3. **BMI**: Shows moderate importance in both models, particularly for the neural network, which may better capture its non-linear relationship with charges.
4. **Region**: The southwest region appears to have distinct importance in both models, suggesting geographical variations in insurance charges.
5. **Children**: Shows lower importance compared to other factors, but is still relevant to the classification task.

The agreement between both models on key predictors increases our confidence in these findings. The importance rankings align well with our exploratory data analysis, which showed strong associations between smoking status, age, BMI, and insurance charges.

8 Prediction on Test Data

Based on our evaluation, we'll select the model that maximizes the F1 score for the "high" charges class for making predictions on the external test data.

Table 4: Model Selection for F1 Score Optimization

	Selected Model	F1 Score	Accuracy	
F1	SVM	0.9688889	0.965	The SVM model achieves the highest F

Table 5: Test Data Prediction Distribution

Class	Count	Percentage
low	144	42.6%
high	194	57.4%

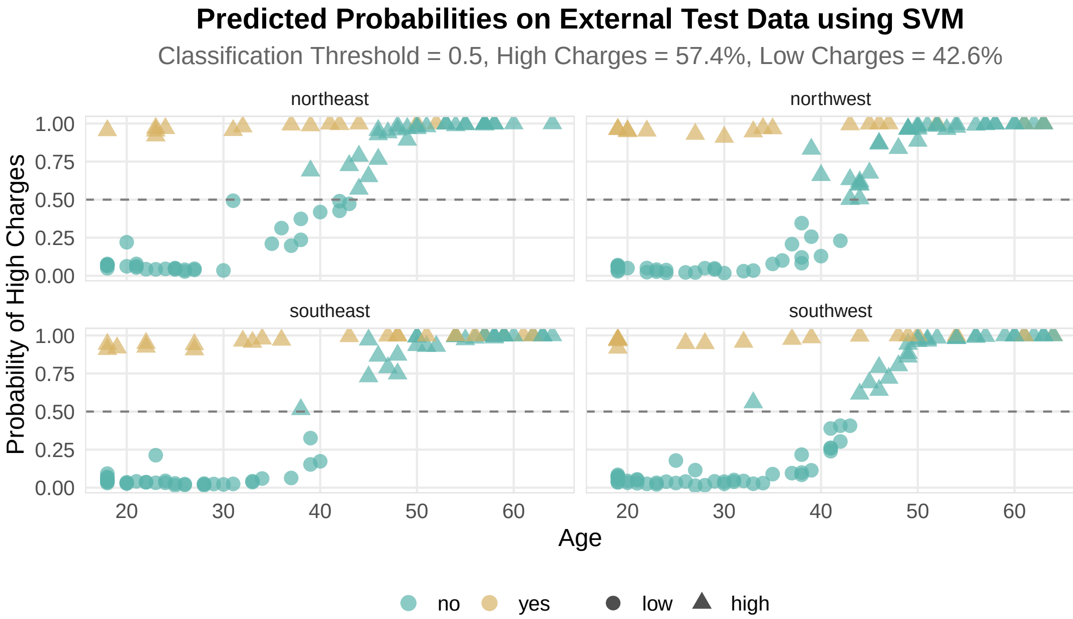


Figure 11: Predicted Probabilities by Age and Smoking Status

We successfully generated predictions for all 338 observations in the test dataset. The predictions have been saved to "TBSCES001.csv" according to the required format, with 57.4% classified as "high" charges and 42.6% as "low" charges.

The visualization of predicted probabilities reveals clear patterns:

1. **Smoking Status**: Strong separation between smokers and non-smokers, with smokers consistently receiving higher probabilities of "high" charges
2. **Age**: Generally positive relationship with the probability of high charges, especially for non-smokers
3. **Regional Variations**: Some regional differences in predicted probabilities, with the southwest region showing slightly different patterns
4. **Decision Boundary**: The 0.5 threshold (dashed line) effectively separates the two classes

These patterns align with our feature importance analysis and exploratory findings, confirming that our model has captured meaningful relationships in the data.

9 Conclusion

This analysis has successfully developed and evaluated advanced machine learning models for predicting high versus low medical insurance costs based on patient characteristics. Here are the key findings:

1. **Model Performance**:
   - The SVM model achieved the best F1 score of 0.969 for predicting high insurance charges
   - Both models showed strong discriminative ability with AUROC values above 0.85
   - The selected model demonstrates a good balance between precision and recall, making it suitable for identifying high-cost cases
2. **Key Predictors**:
   - **Smoking status** emerged as the dominant factor influencing insurance charges, with 100% of smokers having high charges compared to only 46.1% of non-smokers
   - **Age** demonstrated a consistent positive relationship with high charges
   - **BMI** showed a significant association with charges, particularly for smokers
   - The interaction between smoking and BMI is particularly important, with the effect of BMI being much stronger for smokers
3. **Practical Implications**:
   - **For Individuals**: Smoking cessation and weight management present the most significant opportunities for reducing insurance costs
   - **For Insurers**: Risk assessment models should incorporate these key factors and their interactions for more accurate premium setting
   - **For Policymakers**: Public health initiatives targeting smoking and obesity could have substantial impacts on healthcare costs

4. **Methodological Insights**:
    - Both SVM and neural network approaches effectively captured complex non-linear relationships and interactions
    - The SVM model demonstrated slightly better performance, likely due to its ability to find optimal decision boundaries with kernel methods
    - Proper regularization and hyperparameter tuning were crucial for achieving optimal performance

The clear relationship between modifiable risk factors (especially smoking and BMI) and insurance costs suggests that targeted interventions could significantly reduce healthcare expenses while improving public health outcomes.

In future work, incorporating additional features such as medical history, lifestyle factors, and more detailed regional information could further improve predictive accuracy. Longitudinal studies would also help assess how changes in risk factors affect insurance costs over time.