Non-smoker

or

Smoker?

STEVENS
INSTITUTE OF TECHNOLOGY
1870

Predicting Smoking Status Using Health Indicators

CS 513 A - Knowledge Discovery & Data Mining (Spring 2025)

Group 4

Himanshu Hemanth Paithane

Vishnu Vardhan Putta

Aditya Yadav

AGENDA

1. Problem Statement & Project Goal
2. Project Workflow
3. Dataset Description
4. Data Cleaning & Exploratory Data Analysis
5. Model Implementation & Evaluation
6. Model Comparison & Conclusion
7. Future Work
8. References

PROBLEM STATEMENT & PROJECT GOAL

Problem Statement:

- Smoking contributes to numerous diseases including cancer, cardiovascular diseases, and respiratory disorders
- Self-reported smoking status is often unreliable due to social stigma and personal biases
- Healthcare providers need objective methods to identify potential smokers for appropriate interventions
- Early detection of smoking habits can significantly improve targeted interventions and reduce health risks
- Biological indicators and health metrics may reveal smoking habits even when not self-reported

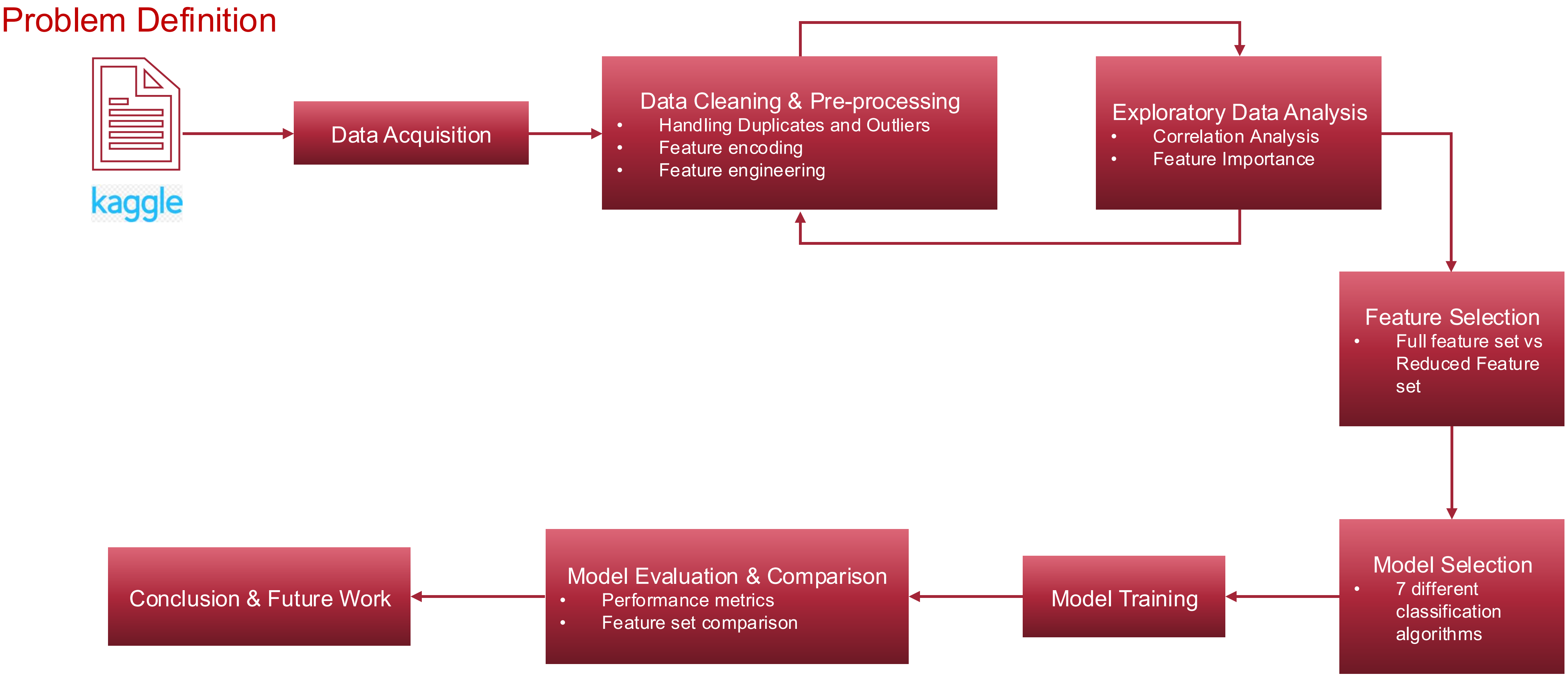
Project Goal:

- Develop classification models to accurately predict smoking status based on health indicators
- Identify the most predictive biological signals of smoking
- Determine the minimum set of health indicators needed for reliable prediction
- Compare multiple machine learning approaches to identify optimal models for different use cases

Real-World Applications:

- Healthcare screening to identify potential smokers who don't disclose their habit
- Insurance risk assessment for more accurate premium calculations
- Public health research to study smoking prevalence and effects
- Preventive care and targeted smoking cessation interventions
- Evidence-based policy development for tobacco control

PROJECT WORKFLOW



Problem Definition

kaggle

Data Acquisition

Data Cleaning & Pre-processing
- Handling Duplicates and Outliers
- Feature encoding
- Feature engineering

Exploratory Data Analysis
- Correlation Analysis
- Feature Importance

Feature Selection
- Full feature set vs Reduced Feature set

Model Selection
- 7 different classification algorithms

Model Training

Model Evaluation & Comparison
- Performance metrics
- Feature set comparison

Conclusion & Future Work

DATASET DESCRIPTION

Dataset Overview:

- Source: Kaggle - Body Signal of Smoking Dataset from health screening centers
- URL: https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking/data
- Original Size: 55,692 records with 27 features
- After cleaning: 44,084 records (removed duplicates and outliers)
- Kaggle Usability Score: 7.06
- Objective: Determine smoking status through biological signals and health indicators
- Target Variable: "smoking" (Binary: 0 = Non-smoker, 1 = Smoker)

Feature Categories:

- Demographics: gender, age, height, weight, waist circumference
- Vital Signs: blood pressure (systolic, diastolic), fasting blood sugar
- Blood Tests: cholesterol (total, HDL, LDL), triglycerides, hemoglobin
- Liver Function: AST, ALT, GTP (liver enzymes)
- Sensory Tests: eyesight (left/right), hearing (left/right)
- Other Health Indicators: dental caries, tartar, urine protein, serum creatinine

Class Distribution:

- 63.3% non-smokers (27,972 records)
- 36.7% smokers (16,112 records)
- Moderate class imbalance requiring appropriate evaluation metrics

DATA CLEANING & PREPROCESSING

Initial Data Assessment:

- 55,692 records with 27 features
- No missing values (all fields complete)
- 11,140 duplicate records identified and removed
- Renamed columns to snake_case format for coding consistency
- Removed unnecessary features:
  - ID column (just an index identifier)
  - 'oral' column (constant value across all records)

Categorical Variable Encoding:

- Gender: F → 0, M → 1
- Tartar: Y → 1, N → 0
- Hearing status: Values standardized as 1 = normal, 2 = abnormal

Data Quality Assessment:

- Examined feature distributions and identified potential outliers
- Checked for class imbalance (63.3% vs 36.7%)
- Verified data types and consistency

Final Dataset:

- After cleaning: 44,084 records with 25 features
- Created two datasets for model comparison:
  - Full feature set: All 29 features (including engineered features)
  - Reduced feature set: Top 15 most important features

OUTLIER DETECTION AND REMOVAL

Outlier Detection Method:

- Used Interquartile Range (IQR) method for robust outlier identification
- For each numerical feature:
  - Calculated Q1 (25th percentile) and Q3 (75th percentile)
  - Defined outlier boundary as Q1-1.5*IQR and Q3+1.5*IQR
  - Flagged values outside this range as potential outliers
- Conservative approach: only removed rows that were outliers in 5+ columns
- This preserved legitimate variability while removing truly anomalous records

Results:

- 468 outliers identified and removed (1.05% of data)
- Final dataset: 44,084 records
- Particularly affected features: triglyceride, ALT, AST, GTP (liver enzymes)

Why IQR Method?

- Resistant to extreme values (doesn't rely on mean)
- Uses quartiles for robust representation of data distribution
- 1.5 * IQR threshold is statistically sound and widely accepted
- Preserves natural biological variability while removing true anomalies

TARGET VARIABLE AND CLASS DISTRIBUTION

Target Variable Analysis:

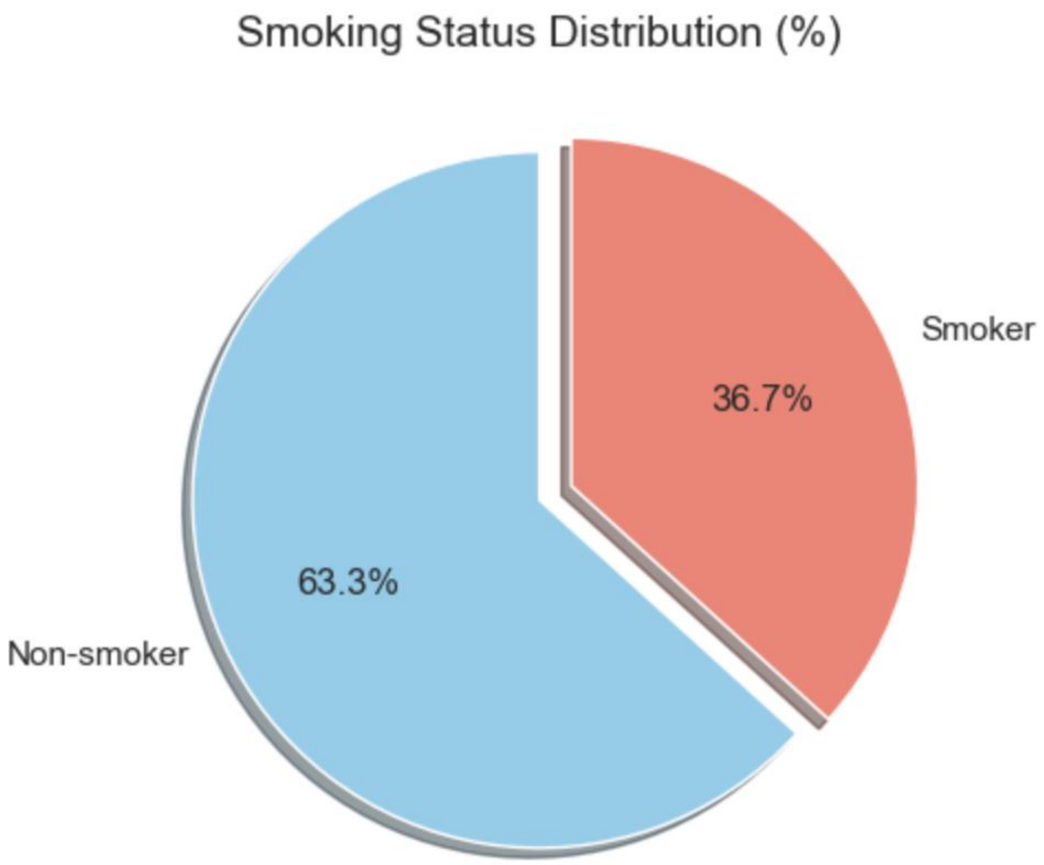
- Binary classification: Smoker (1) vs Non-smoker (0)
- After cleaning: 63.3% non-smokers (27,972), 36.7% smokers (16,112)
- Class ratios preserved in train/test splits using stratification

Class Imbalance Assessment:

- 63:37 ratio represents moderate imbalance
- Not severe enough to require advanced resampling techniques
- Comparable to real-world smoking prevalence in many populations
- According to class imbalance literature:
  - 60:40 or 70:30 is mild imbalance
  - 80:20 is moderate imbalance
  - 90:10 or more extreme is severe imbalance

Evaluation Strategy:

- 80% training set, 20% test set
- Stratified sampling to maintain class distribution
- Multiple metrics beyond accuracy:
  - Precision, Recall, F1-score, ROC-AUC



NUMERICAL FEATURE DISTRIBUTIONS

Numerical Feature Histograms:

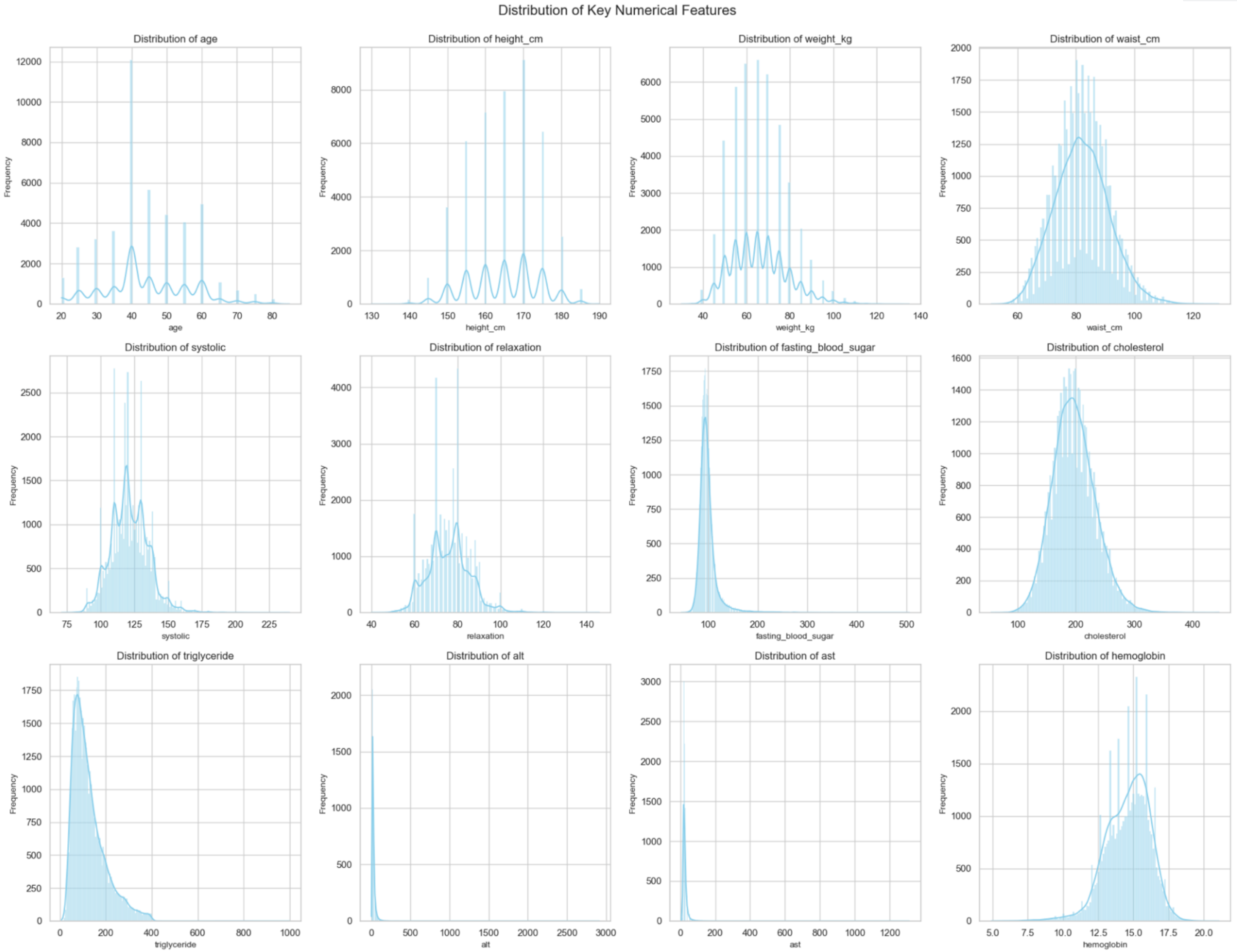
- Explored distributions of key health indicators to understand data characteristics
- Most features show approximately normal distributions with varying degrees of skewness
- Some features display clear bimodal patterns related to gender differences

Key Observations:

- Height and hemoglobin show bimodal distributions (gender effect)
- Liver enzymes (AST, ALT) are right-skewed with long tails
- Triglyceride levels show significant right skew with outliers
- Blood pressure and cholesterol metrics have approximately normal distributions
- Age distribution reflects sampling across different age groups

Implications for Modeling:

- Skewed distributions may affect some models more than others
- Bimodal distributions highlight the importance of gender as a factor
- Range and scale differences justify normalization for distance-based models
- Outlier handling was critical for heavily skewed features



Distribution of Key Numerical Features

Distribution of age

Distribution of height_cm

Distribution of weight_kg

Distribution of waist_cm

Distribution of systolic

Distribution of relaxation

Distribution of fasting_blood_sugar

Distribution of cholesterol

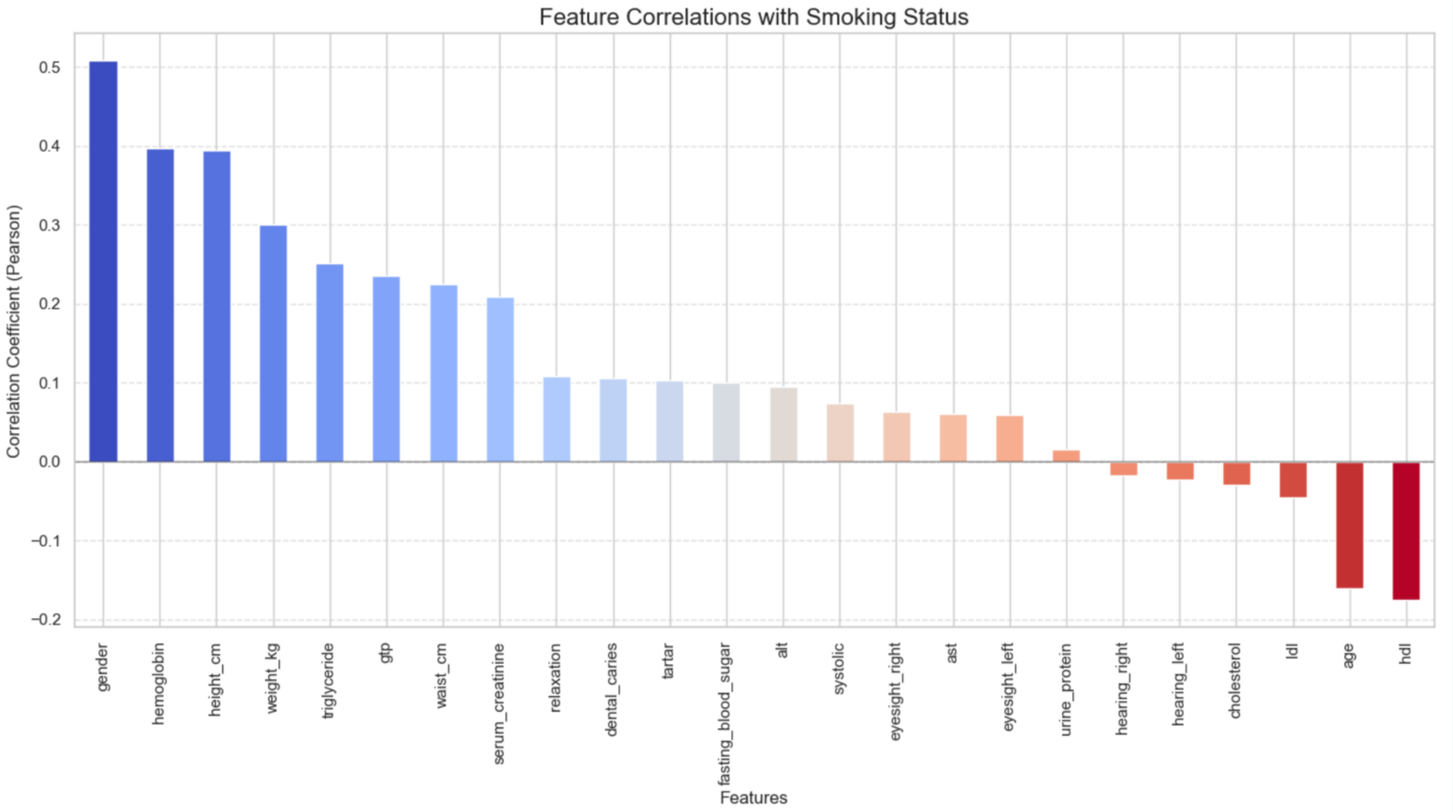
Distribution of triglyceride

Distribution of alt

Distribution of ast

Distribution of hemoglobin

CORRELATION ANALYSIS



Top Positive Correlations:

- Gender (0.508) - Significantly higher in males
- Hemoglobin (0.397) - Higher levels in smokers (CO effect)
- Height (0.394) - Taller individuals (gender-related)
- Weight (0.299) - Heavier individuals
- Triglyceride (0.251) - Higher levels in smokers
- GTP (0.247) - Elevated liver enzyme in smokers

Top Negative Correlations:

- HDL (-0.176) - Lower "good" cholesterol in smokers
- Age (-0.161) - Younger individuals more likely smokers
- LDL (-0.045) - Negative correlation with "bad" cholesterol
- Cholesterol (-0.030) - Slight negative correlation with total cholesterol
- Hearing status (-0.024) - Minor correlation with hearing

Physiological Significance:

- Smoking increases carbon monoxide in blood, leading to compensatory increase in hemoglobin
- Smoking negatively impacts HDL (good cholesterol) levels
- Smoking affects liver function (GTP, ALT)
- Strong gender effect suggests different smoking patterns between males and females
- Correlations align with established medical knowledge about smoking effects

FEATURE ENGINEERING

New Features Created:

- BMI (Body Mass Index) = weight_kg / (height_cm/100)²
  - Standard health metric used in medical assessment
- Blood Pressure Ratio = systolic / relaxation
  - Indicator of cardiovascular health
- Cholesterol Ratio = total_cholesterol / hdl
  - Key predictor of cardiovascular risk
- LDL/HDL Ratio = ldl / hdl
  - More specific lipid balance indicator
- Hypertension (binary) = 1 if (systolic ≥ 140 OR relaxation ≥ 90)
  - Clinical diagnosis indicator

Correlation with Smoking:

- Cholesterol Ratio: 0.145 (strongest new feature)
- LDL/HDL Ratio: 0.071
- BMI: 0.104
- BP Ratio: -0.061
- Hypertension: 0.010

Clinical Relevance:

- Engineered features capture important health relationships
- Cholesterol ratio is widely used in cardiovascular risk assessment
- BMI provides standardized weight-to-height relationship
- BP ratio reflects arterial stiffness and cardiovascular function
- These metrics are routinely used in clinical practice

FEATURE IMPORTANCE AND SELECTION

Feature Importance Method:

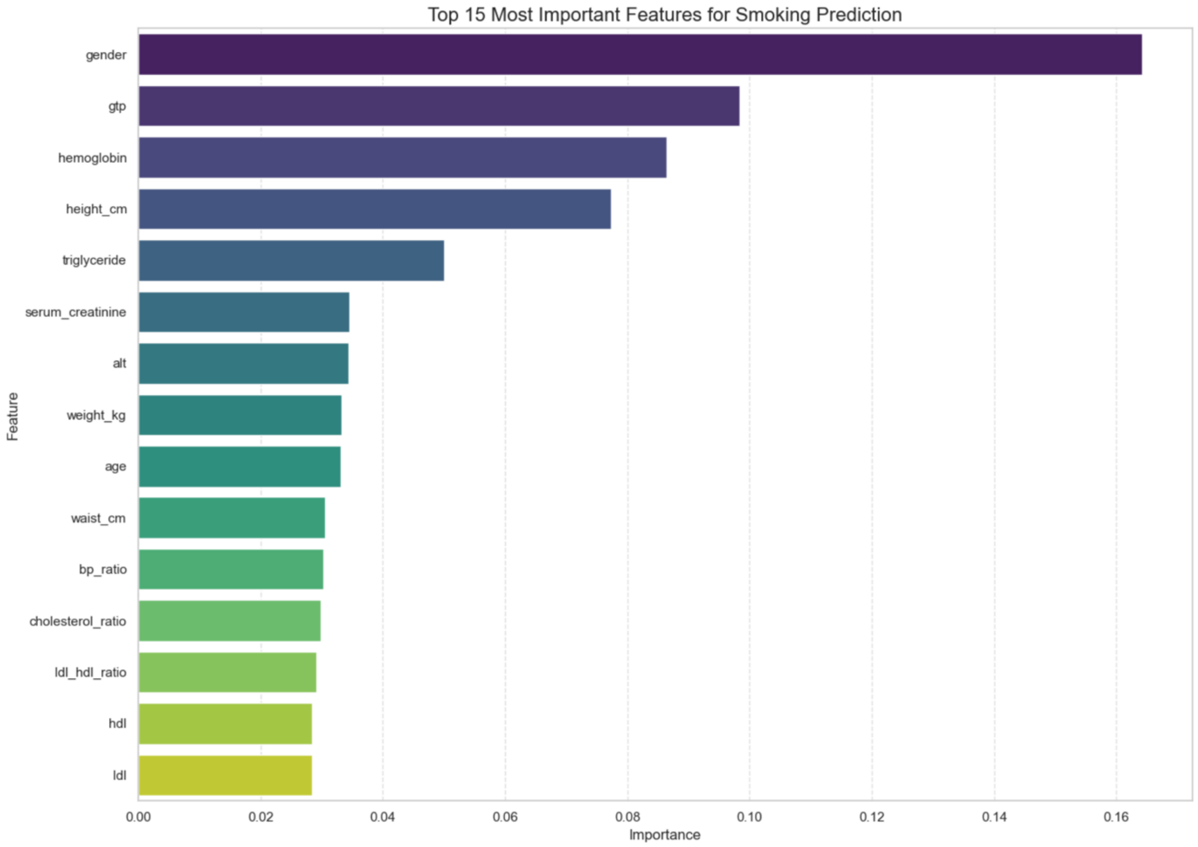
- Used Random Forest for feature ranking
- Benefits: Handles non-linear relationships, robust to outliers
- 100 trees with controlled depth to avoid overfitting
- Cross-validated to ensure stability and generalizability

Feature Selection Strategy:

- Created reduced dataset with top 15 features
- Will compare model performance: full vs. reduced set
- Potential benefits of reduced feature set:
  - Lower data collection costs
  - Faster computation
  - Reduced overfitting risk
  - More interpretable models

Top 15 Important Features:

1. Gender (0.164)
2. GTP (0.098)
3. Hemoglobin (0.086)
4. Height (0.077)
5. Triglyceride (0.050)
6. Serum creatinine (0.035)
7. ALT (0.034)
8. Weight (0.033)
9. Age (0.033)
10. Waist circumference (0.031)
11. Blood Pressure ratio (0.030)
12. Cholesterol ratio (0.029)
13. LDL/HDL ratio (0.029)
14. HDL (0.028)
15. LDL (0.028)



MODEL: KNN

KNN Approach:

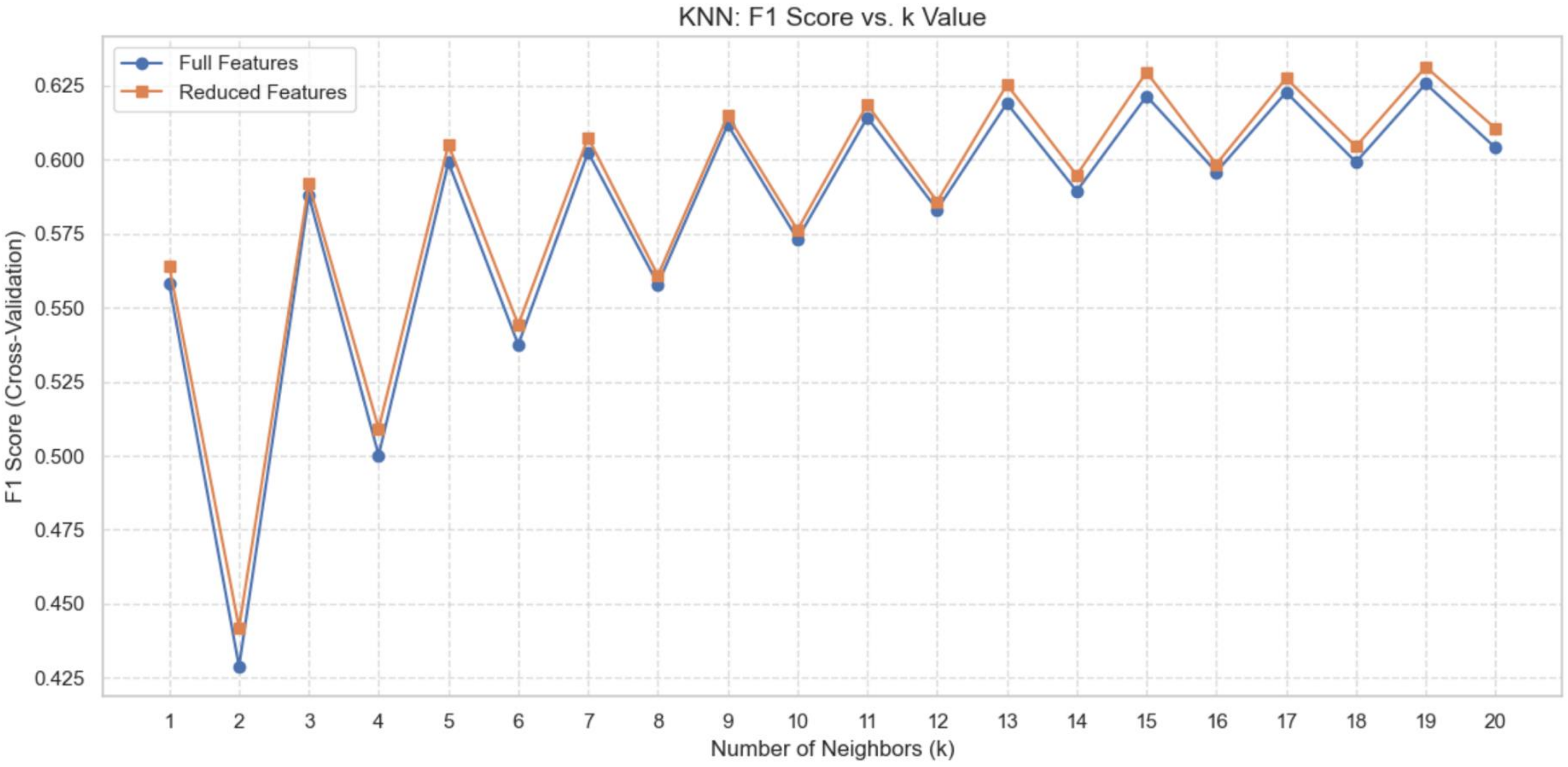
- K-Nearest Neighbors classifies based on similarity to training examples
- Distance-based method requiring careful feature scaling
- Applied MinMaxScaler to scale all features to [0,1] range
- Euclidean distance used for proximity calculation

Optimization Process:

- Tested K values from 1 to 20 using 5-fold cross-validation
- Used F1-score as optimization metric to balance precision and recall
- Evaluated both full and reduced feature sets

Results:

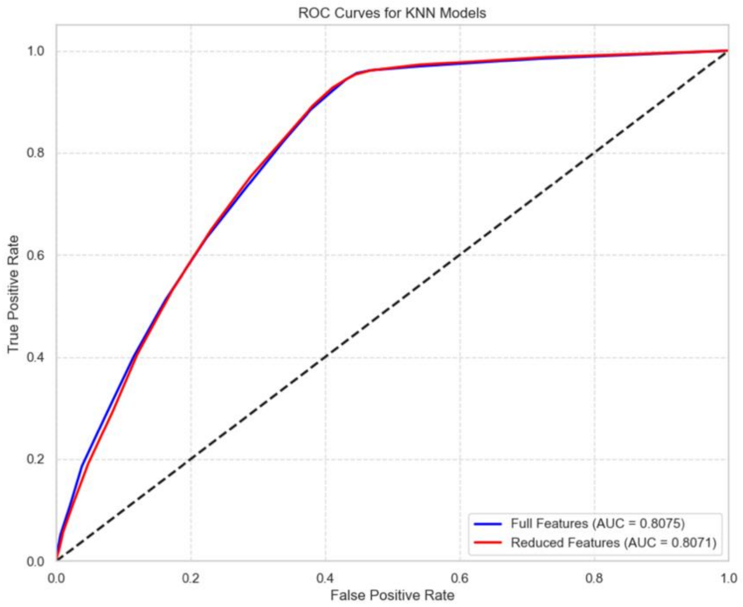
- Optimal K for both datasets: K = 19
- Larger K indicates smoother decision boundaries needed
- Similar optimal K for both feature sets suggests consistent neighborhood structure
- K=19 provides noise resistance while maintaining predictive power

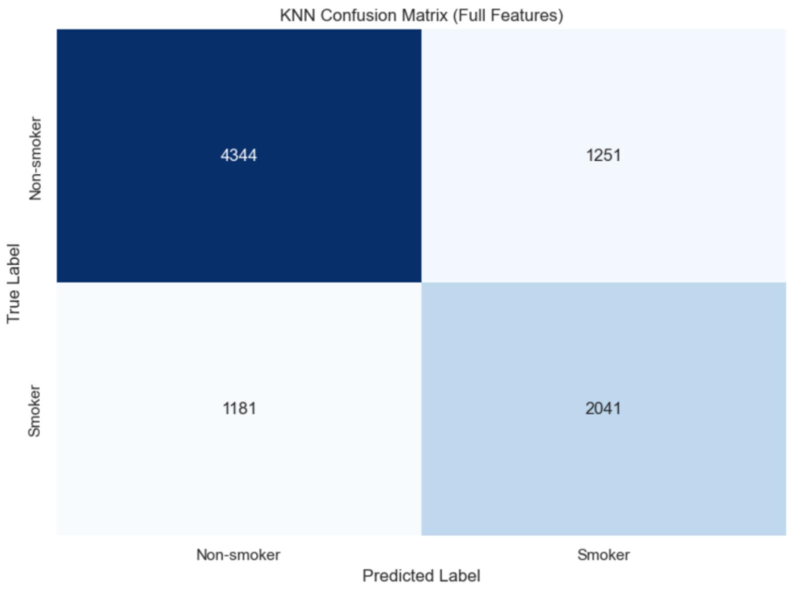


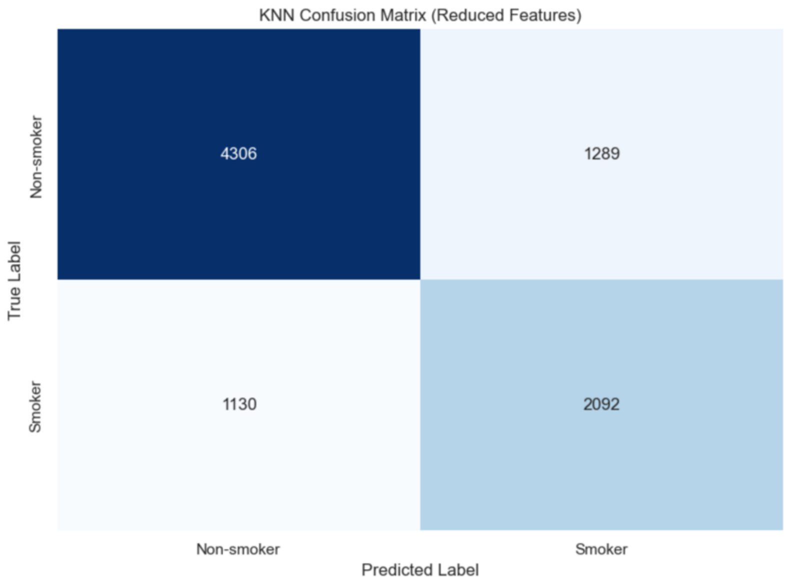
MODEL: KNN

Performance Metrics:

Metric	Full Features	Reduced Features
Accuracy	0.724	0.726
Precision	0.620	0.619
Recall	0.633	0.649
F1 Score	0.627	0.634
ROC AUC	0.808	0.807







Key Observations:

- Reduced feature set performs slightly better than full set
- Improved recall with reduced features (+1.6%)
- Feature reduction simplifies model without performance loss
- Good balance between precision and recall
- Provides strong baseline with simple, interpretable algorithm
- ROC AUC of ~0.81 indicates good discrimination ability

MODEL: DECISION TREE (CART)

CART Implementation:

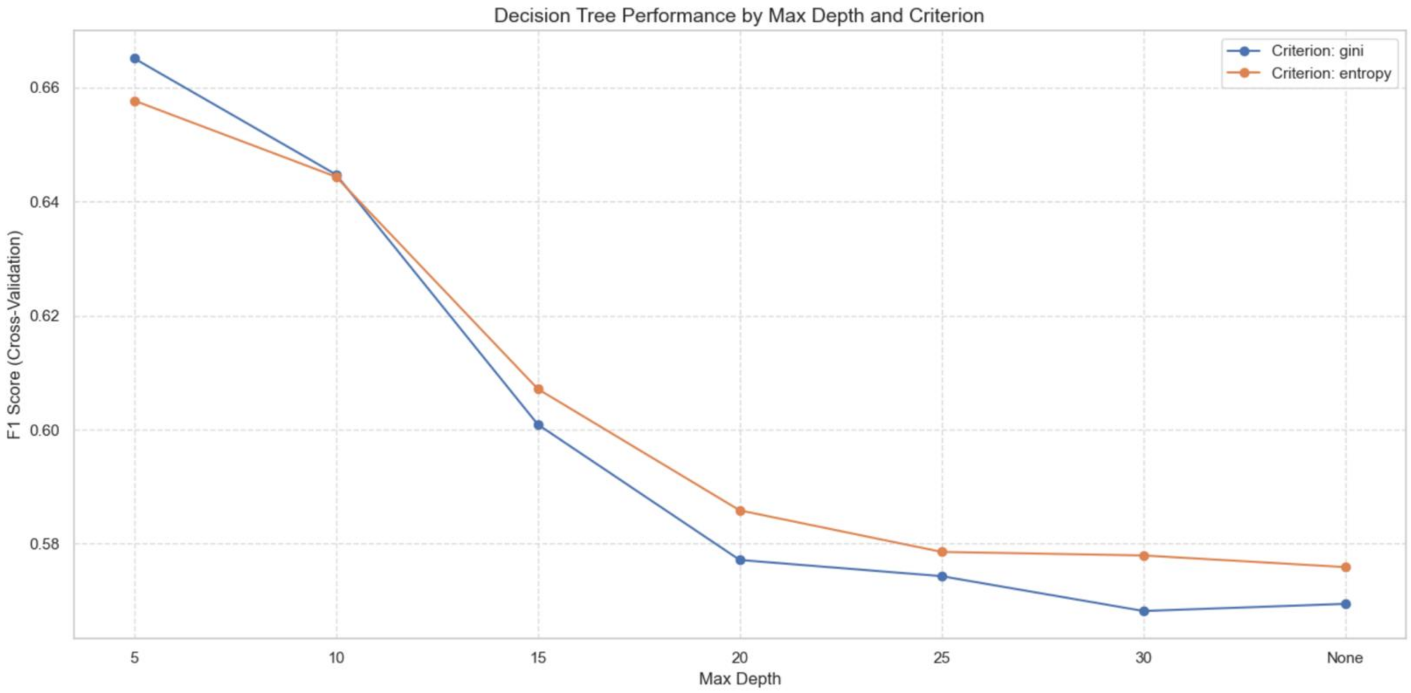
- Classification and Regression Trees (CART) recursively partition data
- Creates a flowchart-like structure of binary decisions
- No feature scaling required (threshold-based decisions)
- Naturally handles both numerical and categorical features

Optimization Approach:

- Tested different split criteria:
  - Gini impurity: Measures probability of misclassification
  - Entropy: Measures information gain
- Explored max_depth values: 5, 10, 15, 20, 25, 30, None
- Evaluated 50 different random states for tree initialization
- Used cross-validation F1-score for parameter selection

Best Parameters:

- Criterion: gini
- Max Depth: 5
- Random State: 1
- Best Cross-Validation F1 Score: 0.67



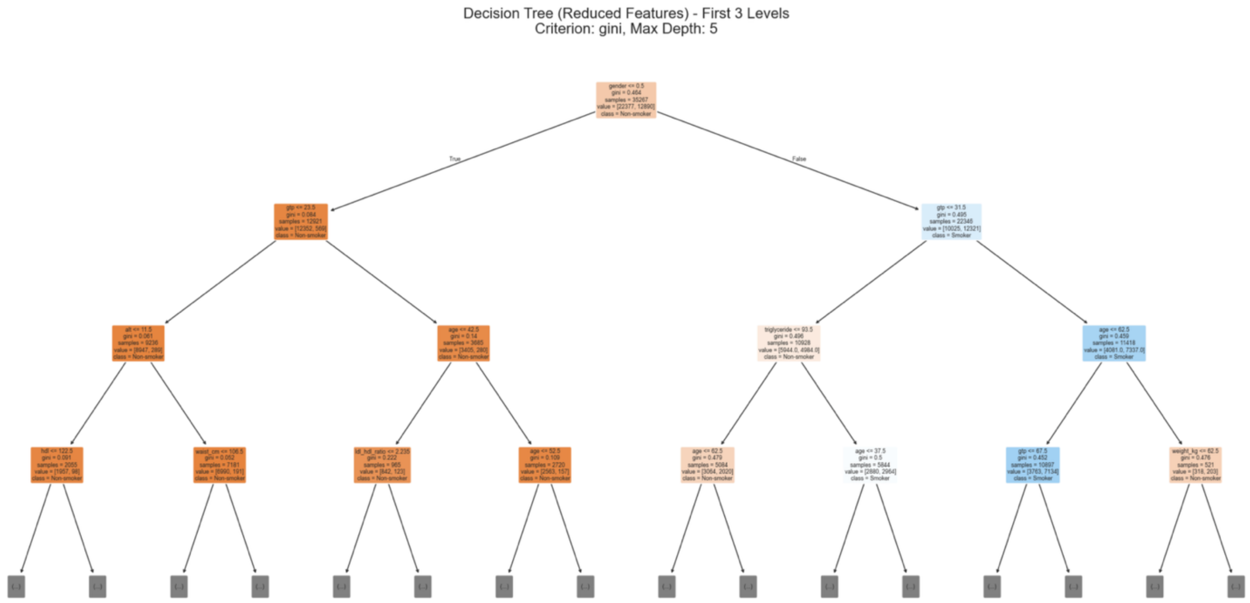
MODEL: DECISION TREE (CART)

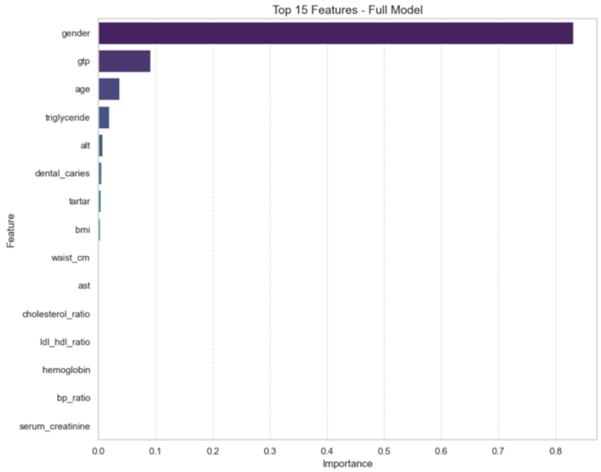
Tree Structure Insights:

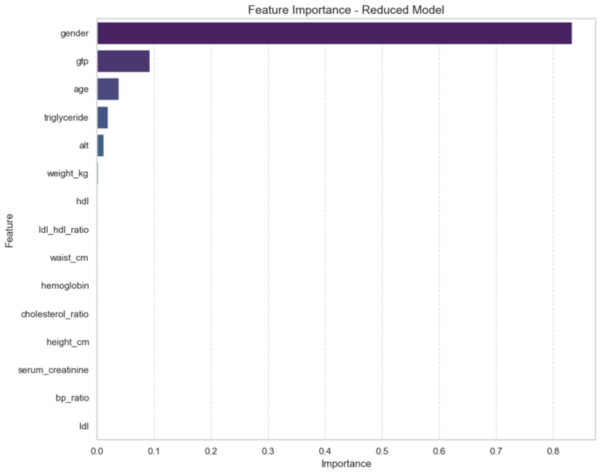
- Root node splits on gender (most discriminative feature)
- For males (gender=1):
  - GTP is the next most important split
  - Age and hemoglobin form subsequent decision points
- For females (gender=0):
  - Different feature pathway follows
  - Hemoglobin and triglyceride are key factors

Top Important Features:

1. Gender (0.83) - Dominates the decision process
2. GTP (0.09) - Liver enzyme elevated in smokers
3. Age (0.04) - Age-related smoking patterns
4. Triglyceride (0.02) - Lipid marker affected by smoking
5. ALT (0.01) - Liver function indicator



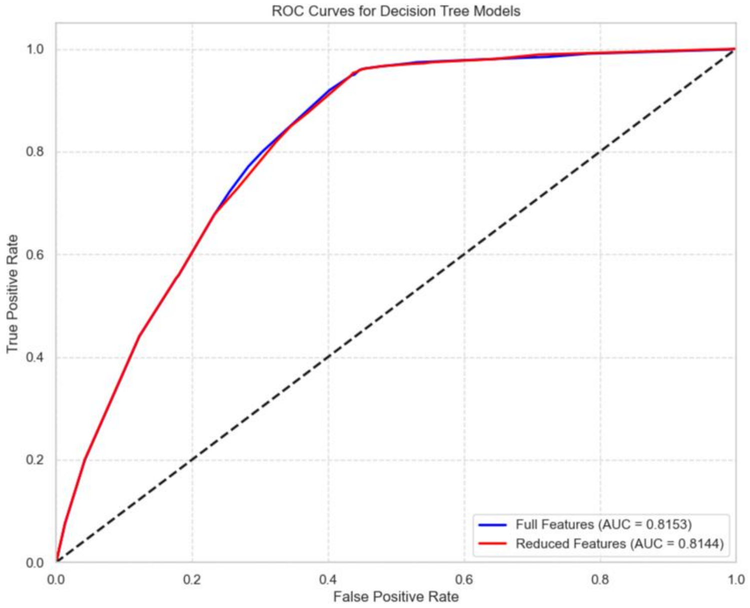


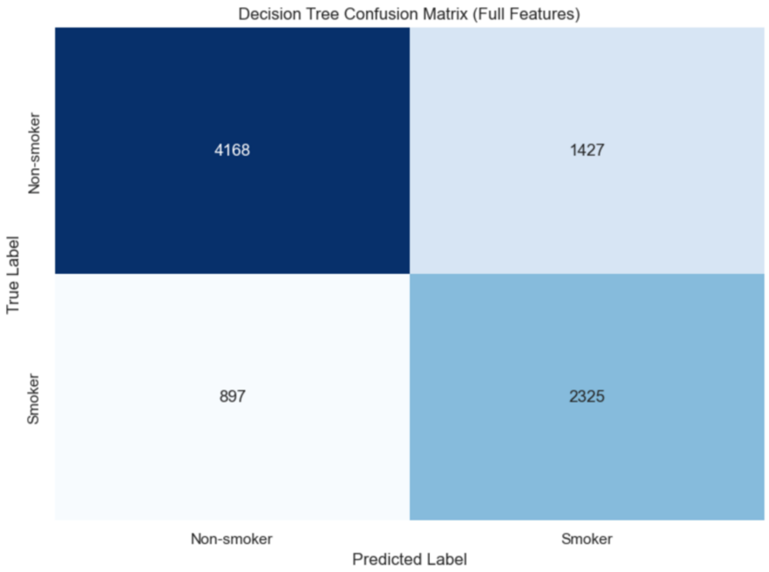


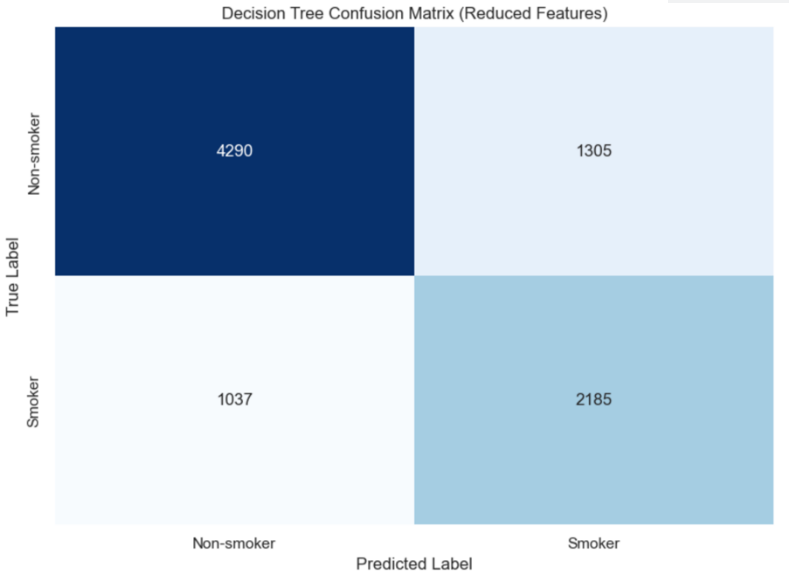
MODEL: DECISION TREE (CART)

Performance Metrics:

Metric	Full Features	Reduced Features
Accuracy	0.736	0.734
Precision	0.620	0.626
Recall	0.722	0.678
F1 Score	0.667	0.651
ROC AUC	0.815	0.814







Observations:

- Full feature set provides better recall (+4.3%)
- Reduced feature set gives slightly better precision
- Only 2.4% reduction in F1-score with reduced features
- Minimal impact on ROC AUC (0.001 difference)
- Simple tree structure provides excellent interpretability
- Outperforms KNN on recall, accuracy, and ROC AUC

MODEL: RANDOM FOREST

Random Forest Approach:

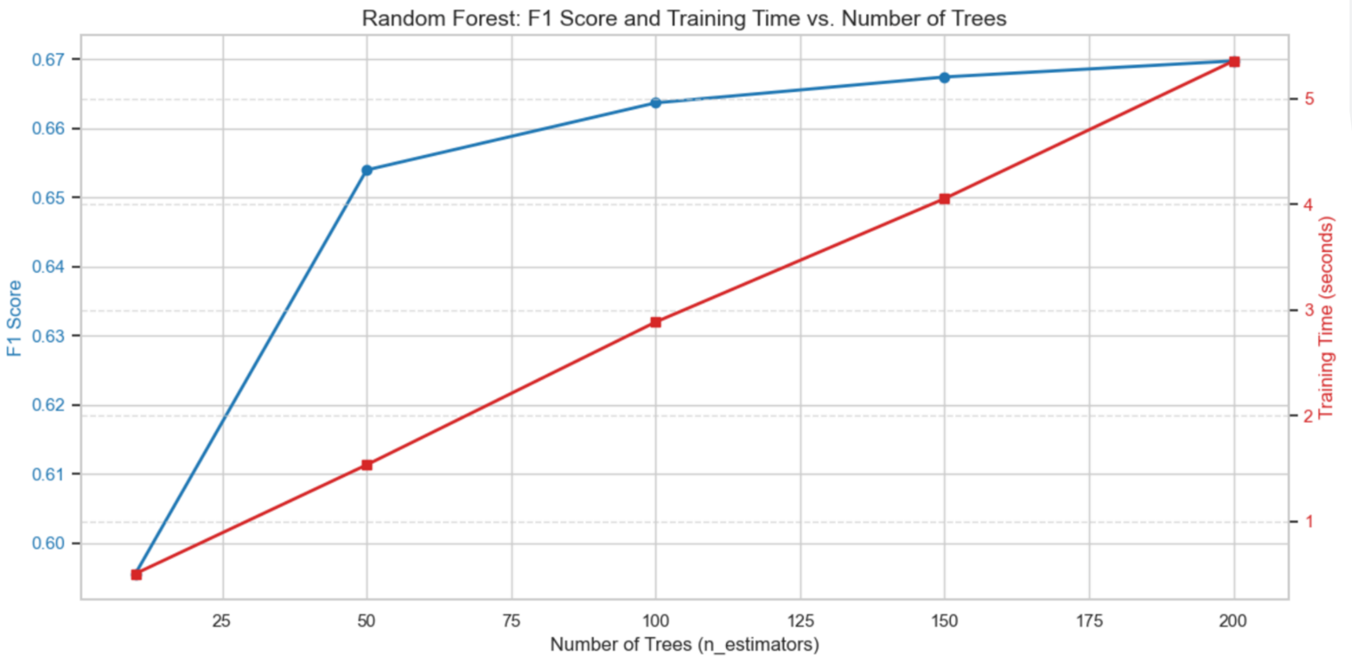
- Ensemble of decision trees, each trained on random data subsets
- Each tree considers random feature subset when splitting
- Final prediction by majority voting across all trees
- Naturally handles feature interactions and non-linearities
- More stable than single decision trees

Parameter Optimization Process:

- Two-stage optimization:
  - First optimized n_estimators (number of trees): 10, 50, 100, 150, 200
  - Then optimized max_depth using best n_estimators: 5, 10, 15, 20, 25, 30, None
- Measured both performance (F1-score) and training time

Results:

- Best n_estimators: 200
- Best max_depth: 10
- F1-score: 0.68 (cross-validation)
- Performance plateaus after 150 trees



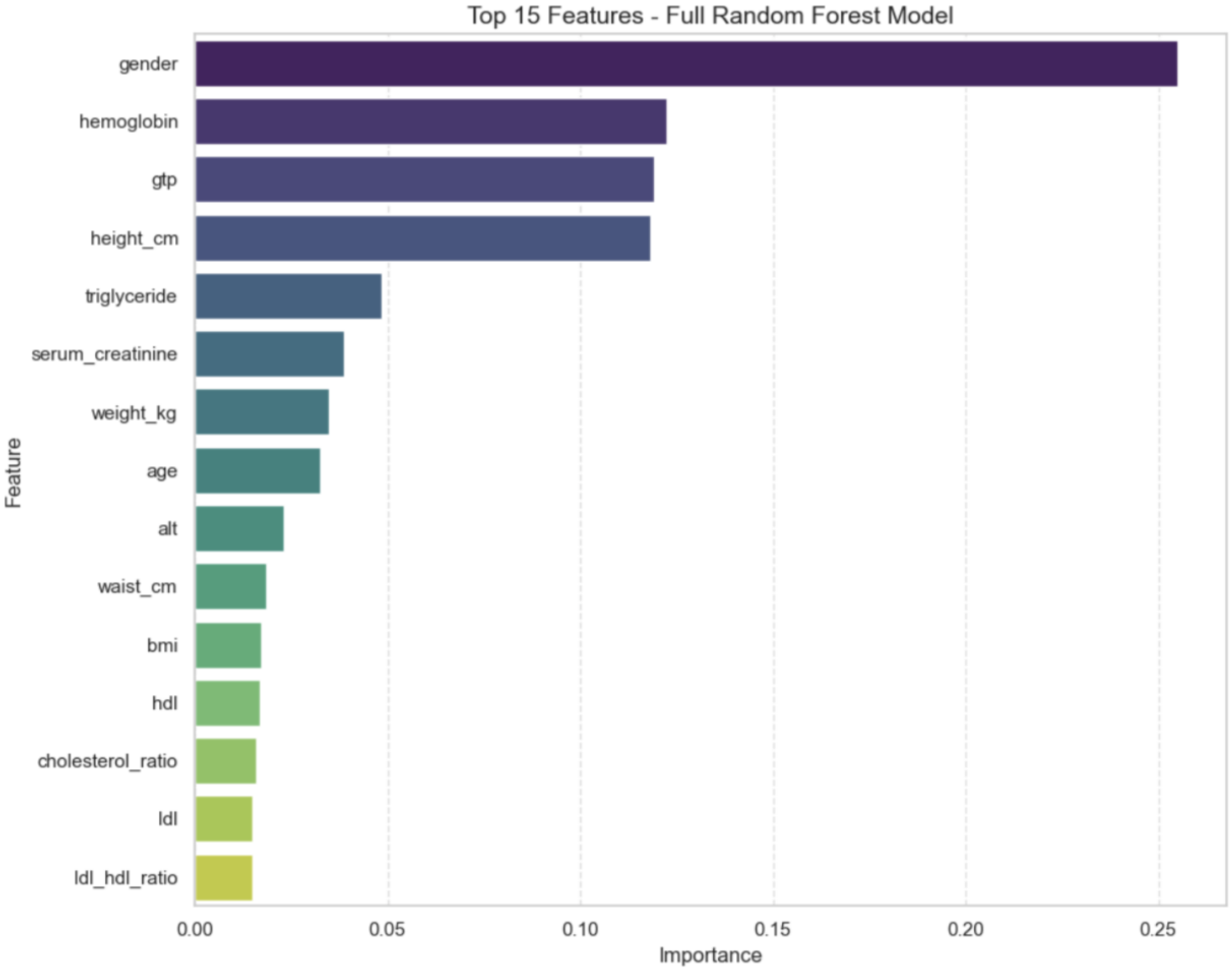
MODEL: RANDOM FOREST

Feature Importance Analysis:

- More balanced feature importance distribution than single tree
- Less dominated by gender (25% vs. 83% in single tree)
- Considers complex feature interactions

Top Important Features:

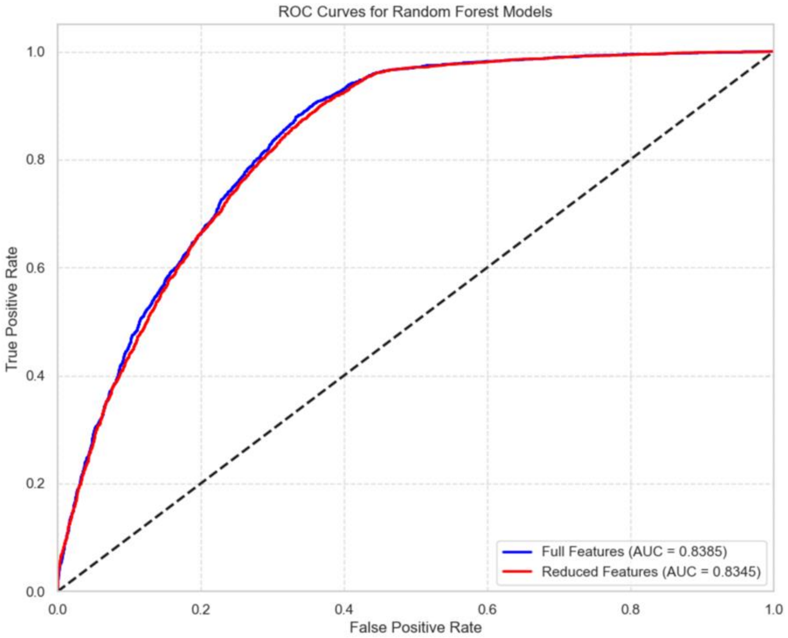
1. Gender (0.25) - Still most important, but more balanced
2. Hemoglobin (0.12) - Affected by carbon monoxide from smoking
3. GTP (0.12) - Liver enzyme elevated in smokers
4. Height (0.12) - Correlates with gender differences
5. Triglyceride (0.05) - Lipid affected by smoking

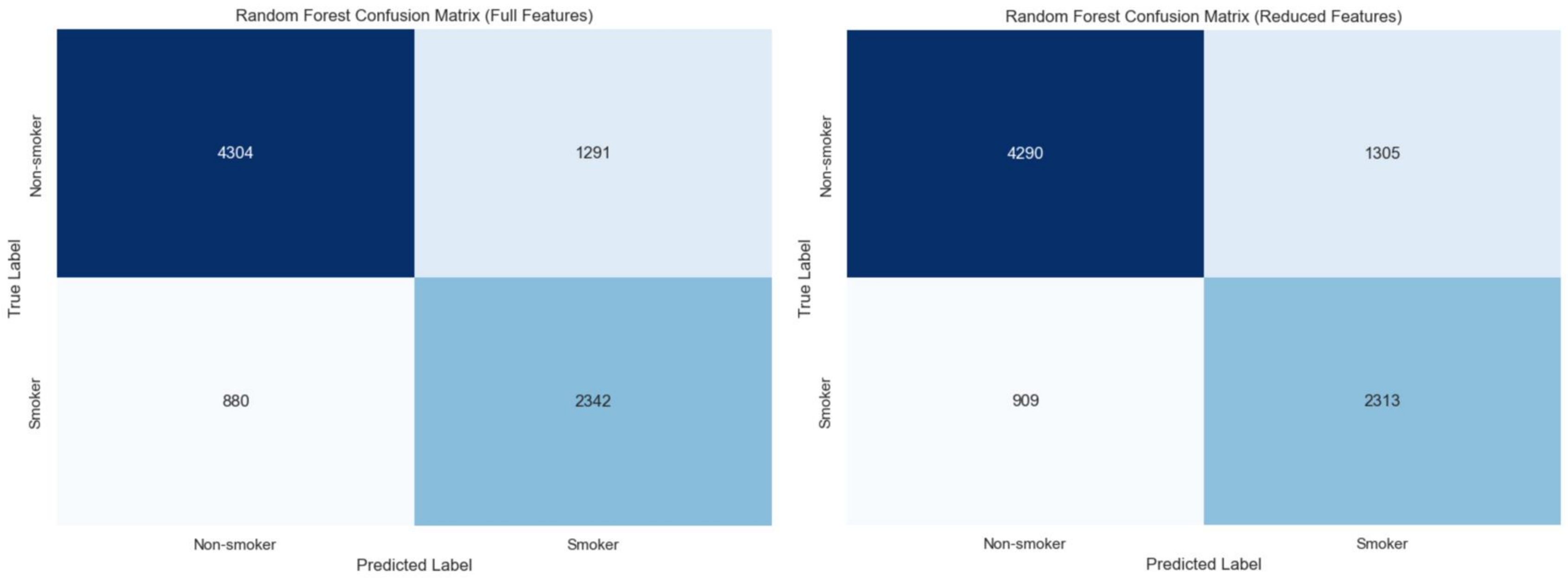


MODEL: RANDOM FOREST

Performance Metrics:

Metric	Full Features	Reduced Features
Accuracy	0.754	0.749
Precision	0.645	0.639
Recall	0.727	0.718
F1 Score	0.683	0.676
ROC AUC	0.838	0.835





Key Observations:

- Highest accuracy among all models tested
- Best performing tree-based model across all metrics
- Strong balance between precision and recall
- Only 1% reduction in F1-score with reduced features
- Highest ROC AUC (0.838) indicates excellent discrimination

MODEL: SVM

SVM Implementation Details:

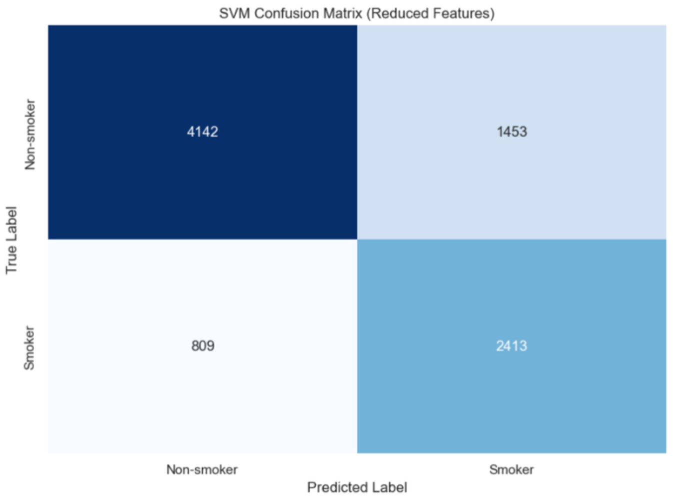
- Support Vector Machine finds optimal hyperplane separating classes
- Used Radial Basis Function (RBF) kernel for non-linear boundaries
- Feature scaling with MinMaxScaler crucial for SVM performance
- Probability estimation enabled for ROC AUC calculation
- Default parameters used: C=1.0, gamma='scale

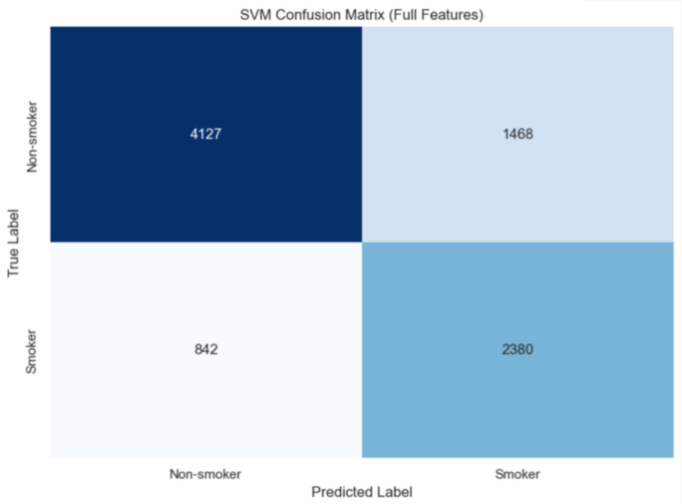
Performance Metrics:

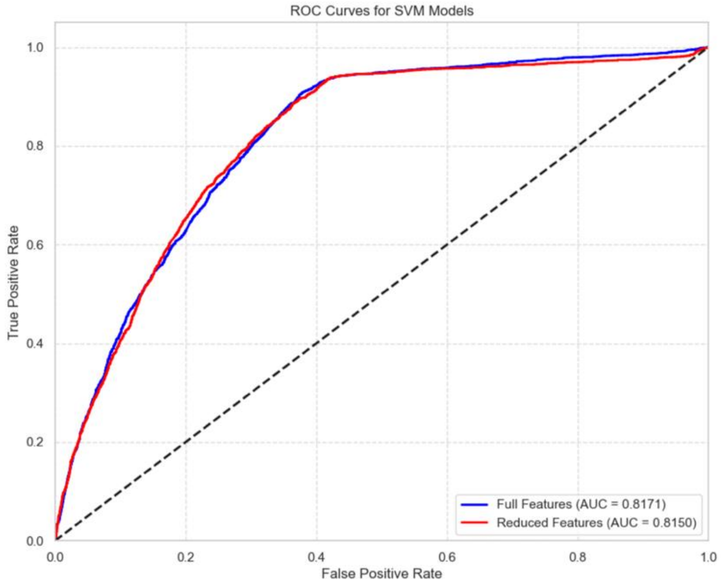
Metric	Full Features	Reduced Features
Accuracy	0.738	0.743
Precision	0.619	0.624
Recall	0.739	0.749
F1 Score	0.673	0.681
ROC AUC	0.817	0.815

Key Observations:

- Reduced feature set performs better than full set (+1.2% F1-score)
- Higher recall and precision with fewer features
- Suggests SVM handles dimension reduction well
- Improved accuracy, precision, and recall with reduced features
- Good balance of precision and recall makes it versatile


SVM Confusion Matrix (Reduced Features)


SVM Confusion Matrix (Full Features)


ROC Curves for SVM Models

MODEL: NAIVE BAYES

Naive Bayes Approach:

- Probabilistic classifier based on Bayes' theorem
- "Naive" assumption: features are conditionally independent
- Different variants handle different data distributions
- Fast training and prediction, even with large datasets

Variants Tested:

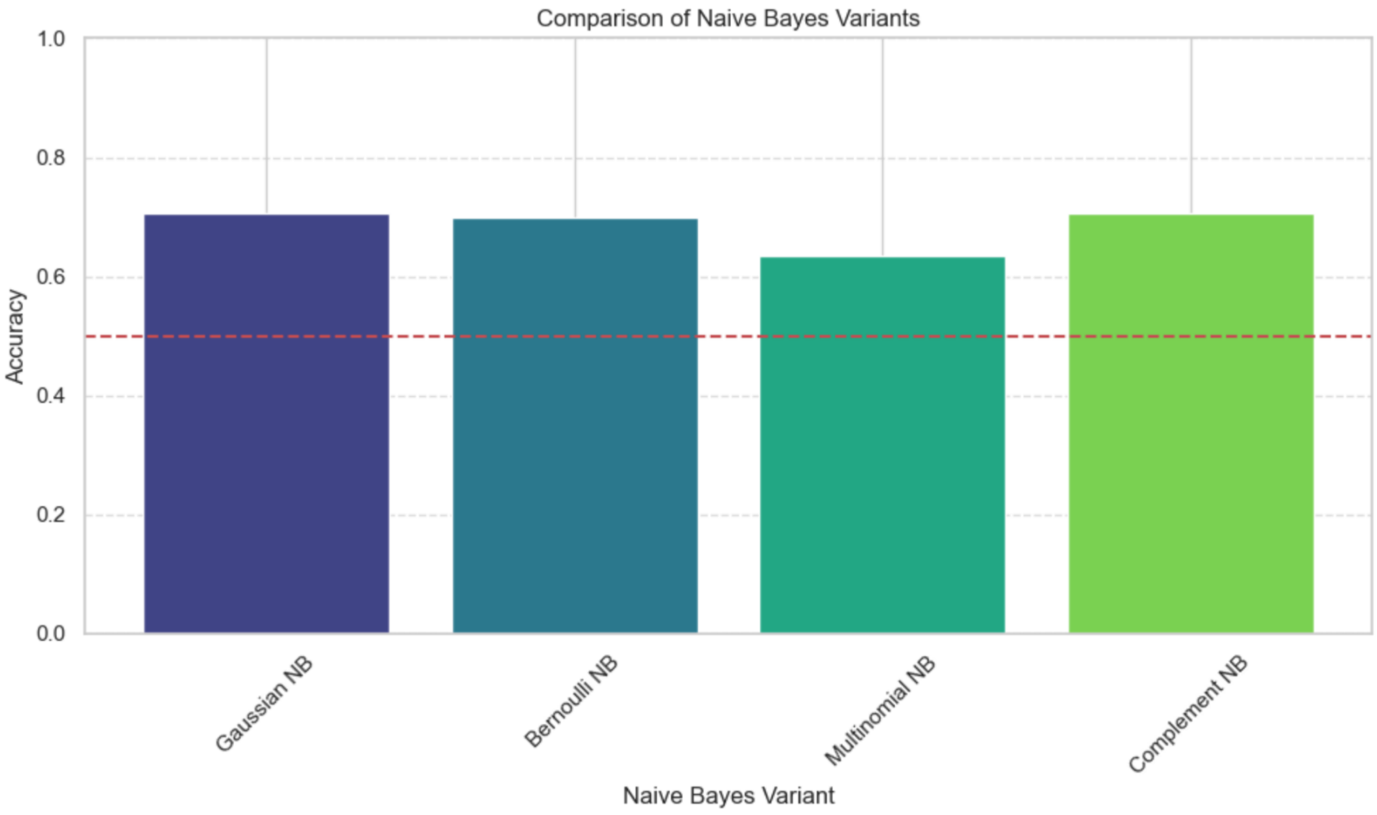
- Gaussian NB: Assumes features follow normal distribution
- Bernoulli NB: For binary/boolean features after binarization
- Multinomial NB: For discrete count data (non-negative features)
- Complement NB: For imbalanced datasets (non-negative features)

Variant Performance:

- Gaussian NB: Accuracy = 0.706
- Bernoulli NB: Accuracy = 0.700
- Multinomial NB: Accuracy = 0.635
- Complement NB: Accuracy = 0.706

Selection Rationale:

- Gaussian NB performs best, matching our continuous health indicators
- Selected Gaussian NB for full evaluation on both feature sets



MODEL: NAIVE BAYES

Implementation Details:

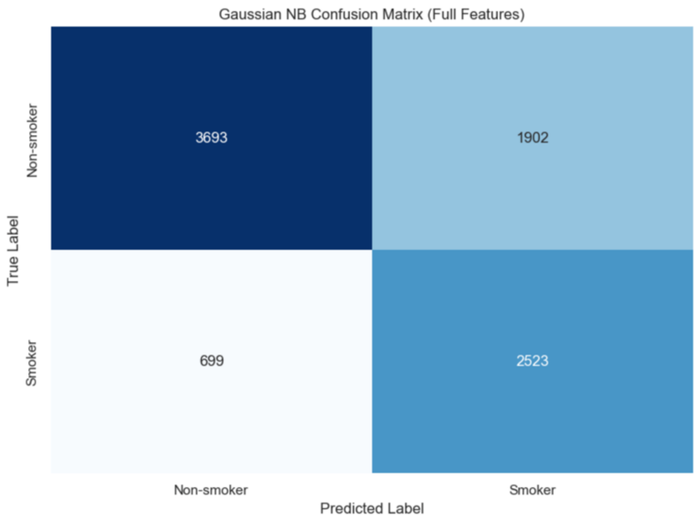
- Applied Gaussian NB to both feature sets
- No feature scaling required (algorithm accounts for feature distributions)
- Default parameters used (no hyperparameter tuning needed)
- Priors calculated automatically from class frequencies
- Fastest training time of all models tested

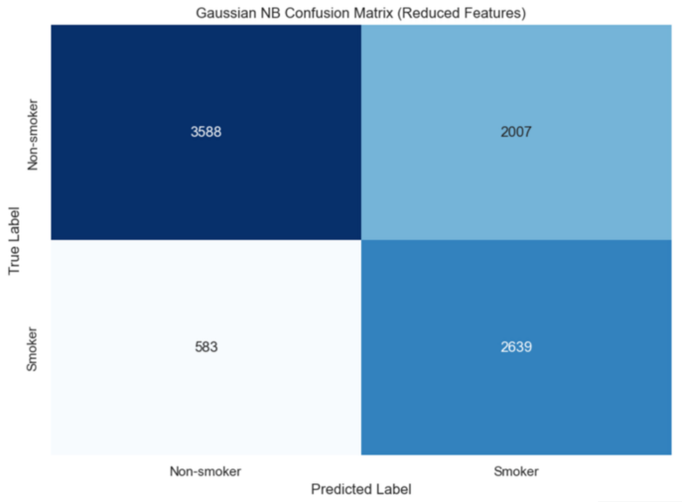
Performance Metrics:

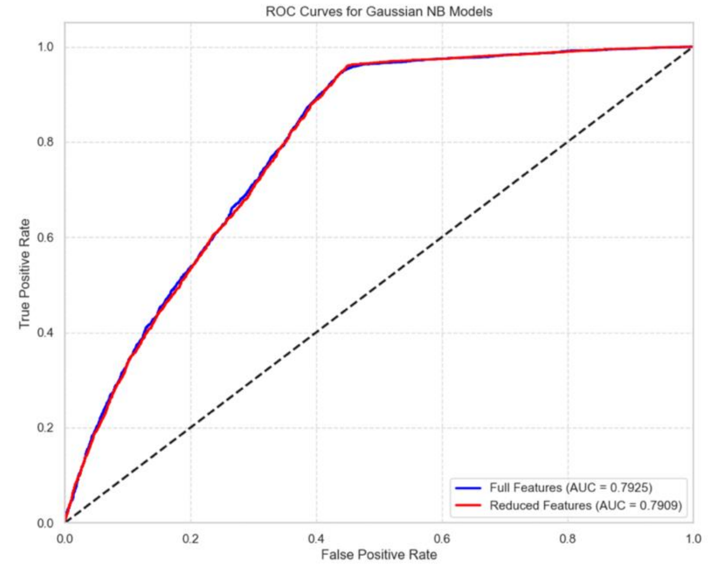
Metric	Full Features	Reduced Features
Accuracy	0.705	0.706
Precision	0.570	0.568
Recall	0.783	0.819
F1 Score	0.660	0.671
ROC AUC	0.793	0.791

Key Observations:

- Highest recall among all models (up to 82%)
- Particularly good at identifying true smokers (low false negatives)
- Lower precision indicates more false positives
- Reduced feature set improves recall by 3.6%
- Independence assumption may explain lower overall accuracy







MODEL: LOGISTIC REGRESSION

About Logistic Regression:

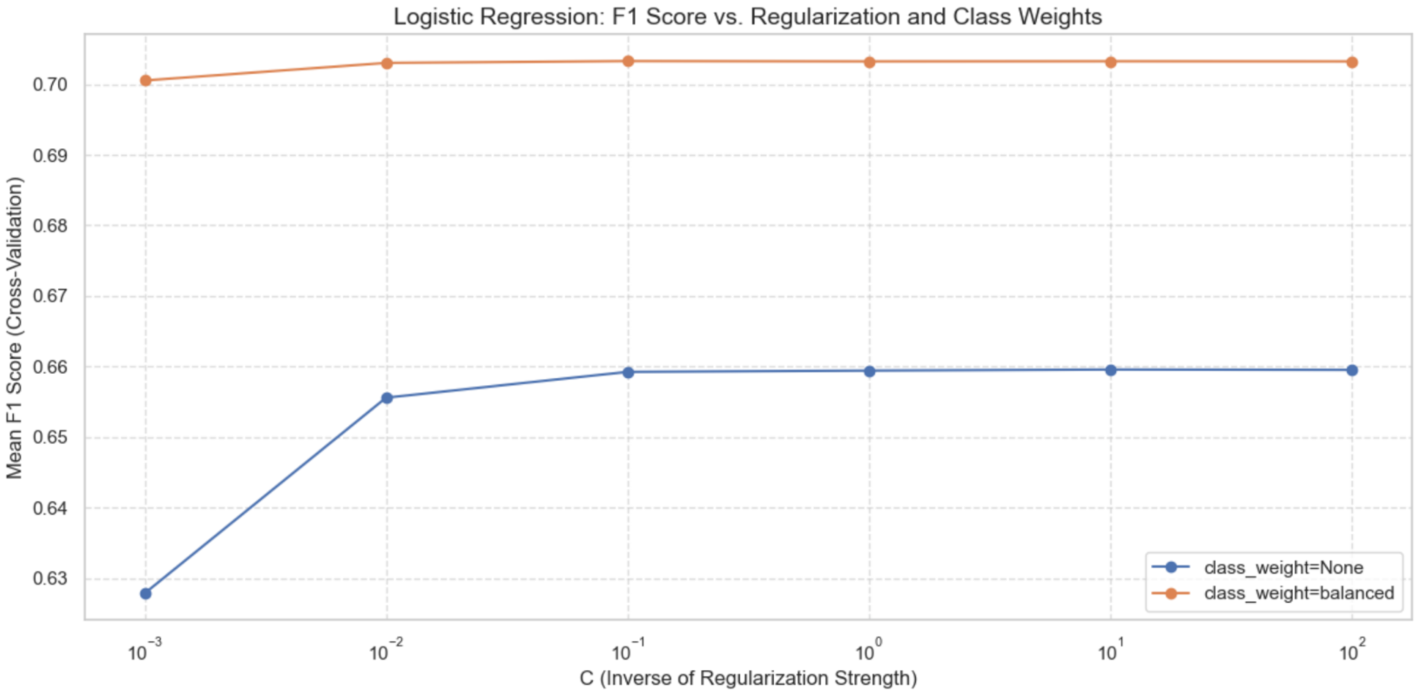
- Linear model for binary classification
- Predicts probability using logistic function transformation
- Models log-odds of positive class as linear combination of features
- Coefficients directly indicate feature importance and direction
- Widely used in medical and epidemiological research

Optimization Approach:

- Tested regularization strengths (C values: 0.001 to 100)
  - C controls inverse of regularization strength
  - Smaller C = stronger regularization = simpler model
- Explored class weight options:
  - None: Equal importance to all classes
  - 'balanced': Weights inversely proportional to class frequencies

Best Parameters:

- C = 0.1 (moderate regularization)
- class_weight = 'balanced'
- Best Cross-Validation F1 Score: 0.703



MODEL: LOGISTIC REGRESSION

Coefficient Interpretation:

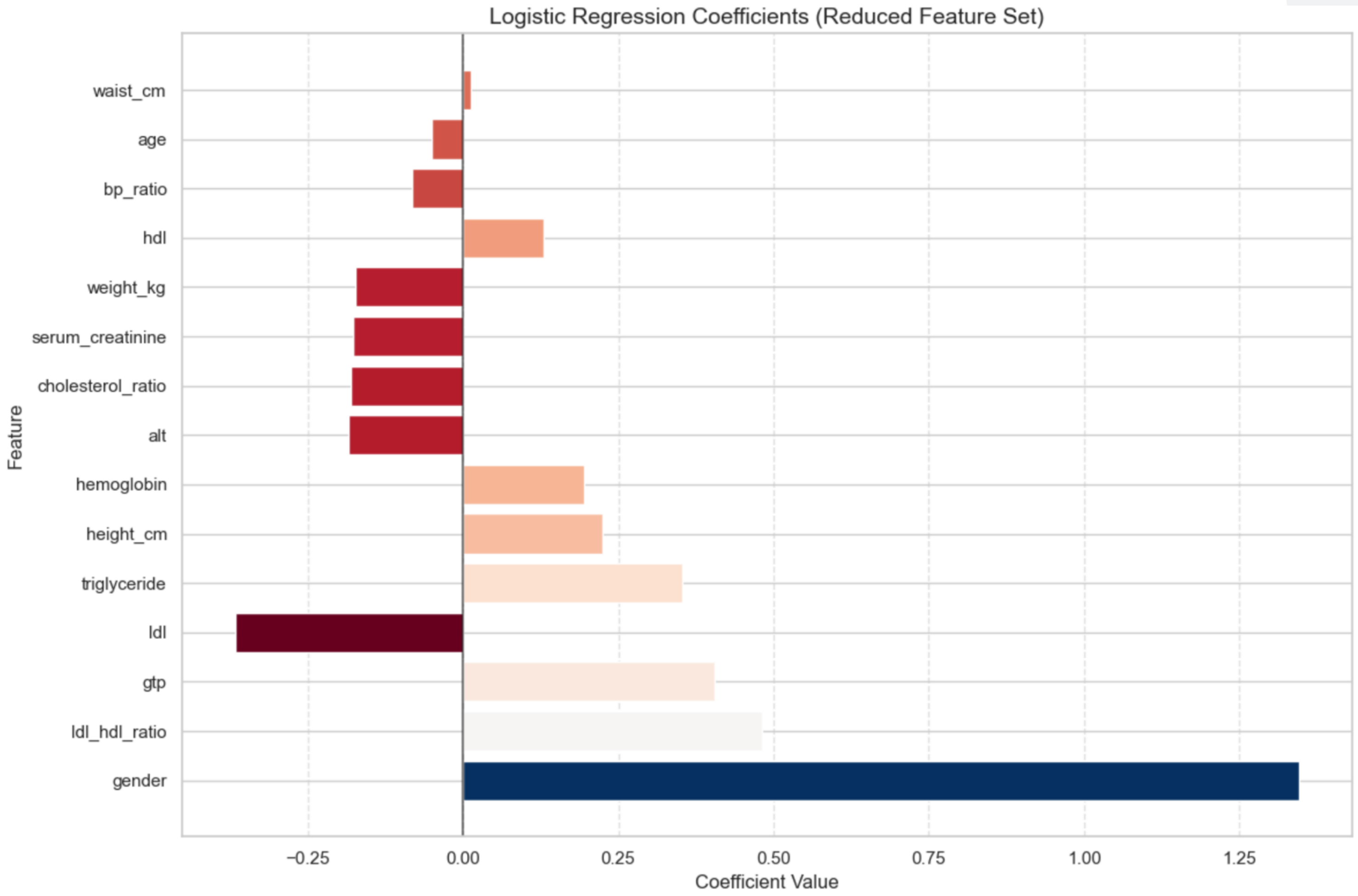
- Each coefficient represents log-odds change per unit feature increase
- Positive coefficient: feature increases smoking probability
- Negative coefficient: feature decreases smoking probability
- Magnitude indicates strength of effect

Top Positive Coefficients:

1. Gender: 1.346 (Males much more likely to be smokers)
2. LDL/HDL ratio: 0.481 (Higher ratio associated with smoking)
3. GTP: 0.405 (Elevated liver enzyme in smokers)
4. Triglyceride: 0.354 (Higher levels in smokers)
5. Height: 0.225 (Taller individuals more likely smokers)

Top Negative Coefficients:

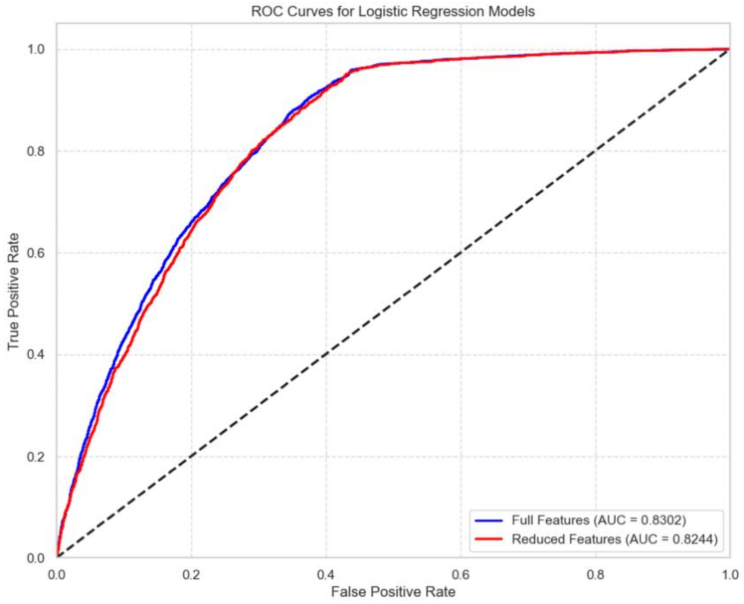
1. LDL: -0.367 (Lower "bad" cholesterol in smokers)
2. ALT: -0.184 (Liver Enzyme)
3. Cholesterol ratio: -0.181 (Complex lipid interactions)
4. Serum creatinine: -0.178 (Kidney function marker)
5. Weight: -0.174 (Complex relationship with smoking)

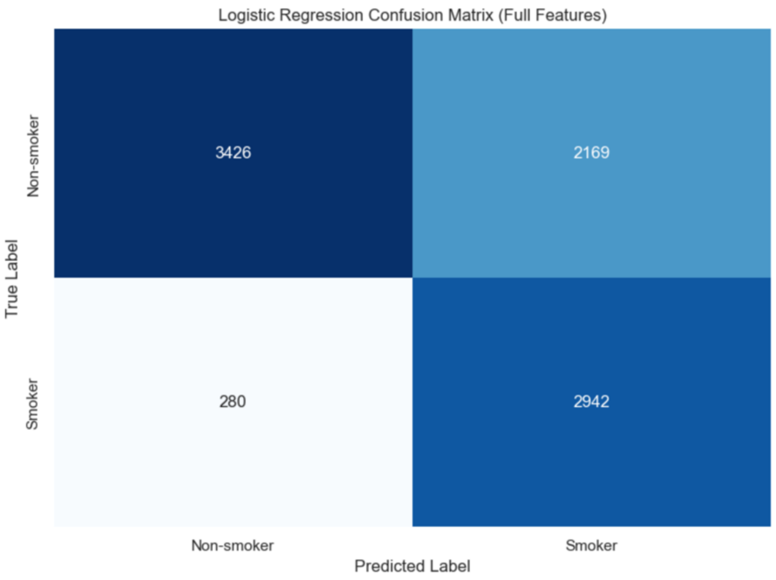


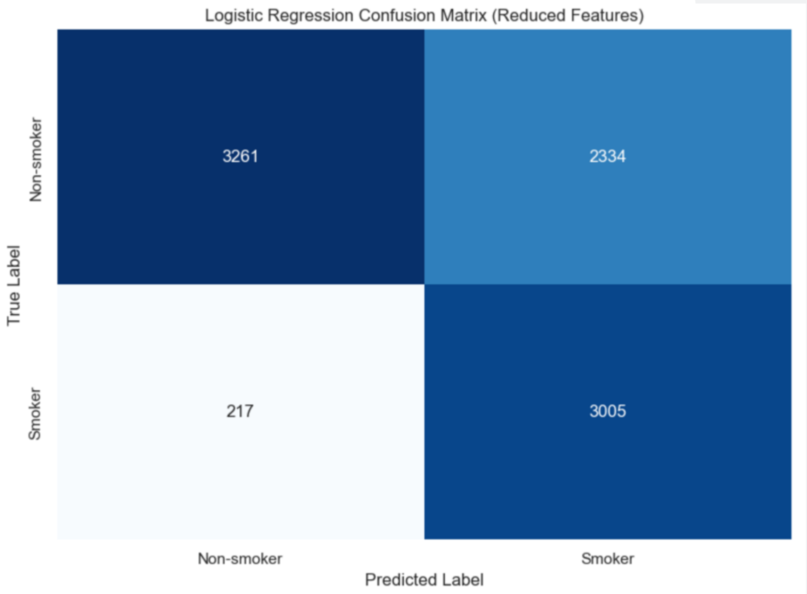
MODEL: LOGISTIC REGRESSION

Performance Metrics:

Metric	Full Features	Reduced Features
Accuracy	0.722	0.711
Precision	0.576	0.563
Recall	0.913	0.933
F1 Score	0.706	0.702
ROC AUC	0.830	0.824







Key Observations:

- Highest F1 score among all models
- Outstanding recall performance (>90%)
- Lower precision due to higher false positive rate
- Excellent at identifying smokers (misses <10%)
- Balanced class weights help address moderate class imbalance
- Maintains strong performance with reduced features
- Good balance between interpretability and performance

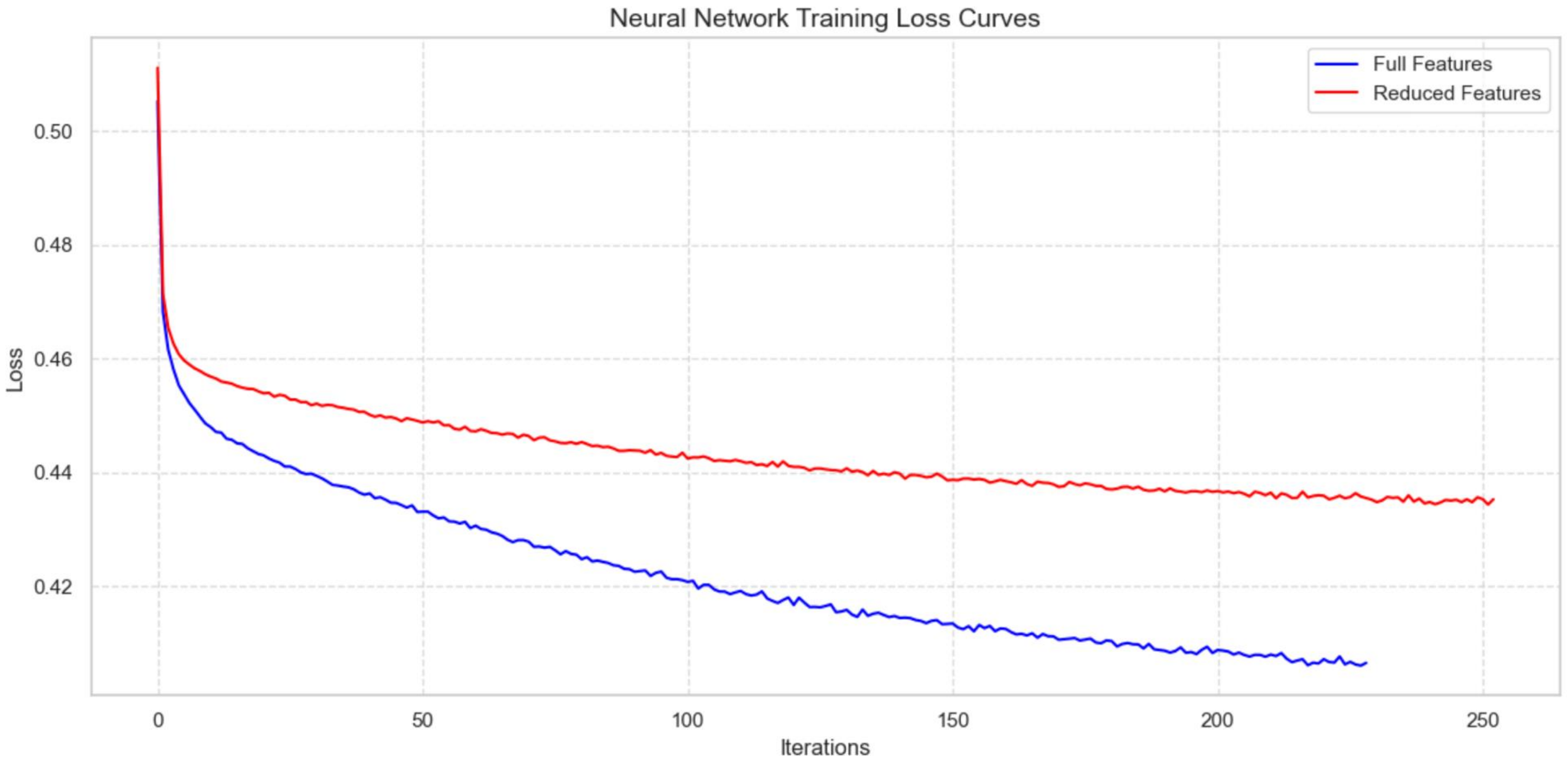
MODEL: ARTIFICIAL NEURAL NETWORK

Network Architecture:

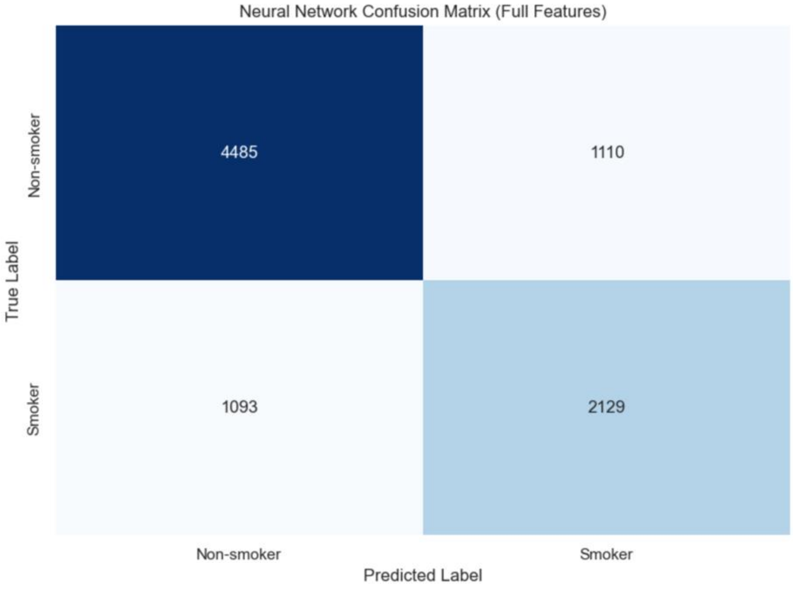
- Multi-Layer Perceptron (MLP) with scikit-learn implementation
- Single hidden layer with 100 neurons
- ReLU activation function for non-linearity
- Adam optimizer for efficient training
- Output layer with softmax activation for probabilities

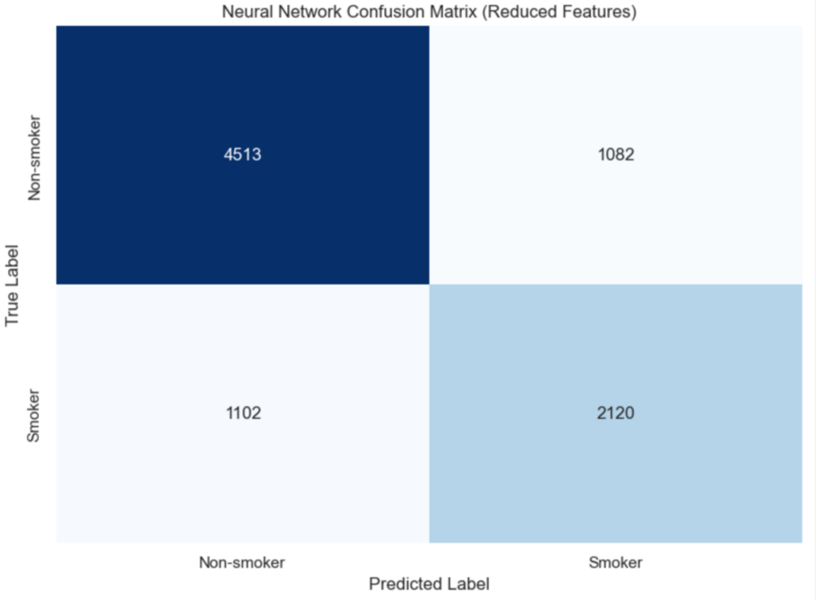
Implementation Details:

- Feature standardization with StandardScaler (critical for neural networks)
- alpha=0.0001 for L2 regularization
- batch_size='auto' for efficient training
- max_iter=300 iterations for convergence
- learning_rate_init=0.001



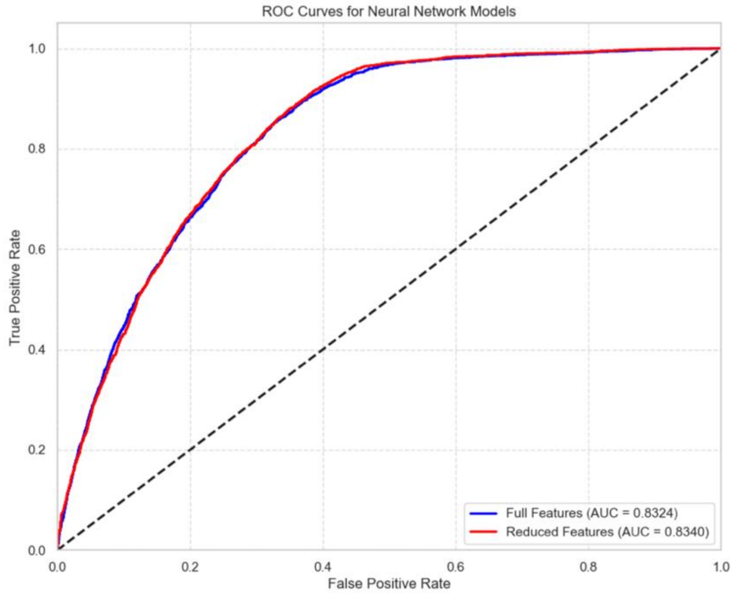
MODEL: ARTIFICIAL NEURAL NETWORK


Neural Network Confusion Matrix (Full Features)


Neural Network Confusion Matrix (Reduced Features)

Performance Metrics:

Metric	Full Features	Reduced Features
Accuracy	0.750	0.752
Precision	0.657	0.662
Recall	0.661	0.658
F1 Score	0.659	0.660
ROC AUC	0.832	0.834


ROC Curves for Neural Network Models

Key Observations:

- Highest precision among all models
- Second-highest ROC AUC (0.834)
- Most balanced between precision and recall
- Reduced feature set performs slightly better
- Strong overall discrimination ability
- Less interpretable than other models
- Good choice when precision is critical

MODEL COMPARISON - ACCURACY & PRECISION

Accuracy Comparison:

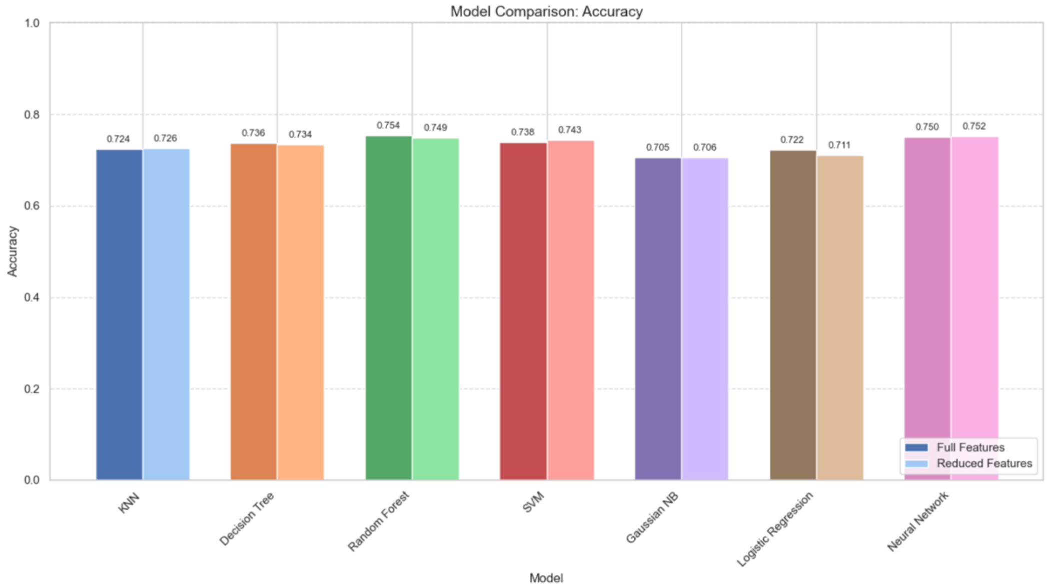
- Random Forest (Full): 0.754 [Highest]
- Neural Network (Reduced): 0.752
- Random Forest (Reduced): 0.749
- SVM (Reduced): 0.743
- Decision Tree (Full): 0.736

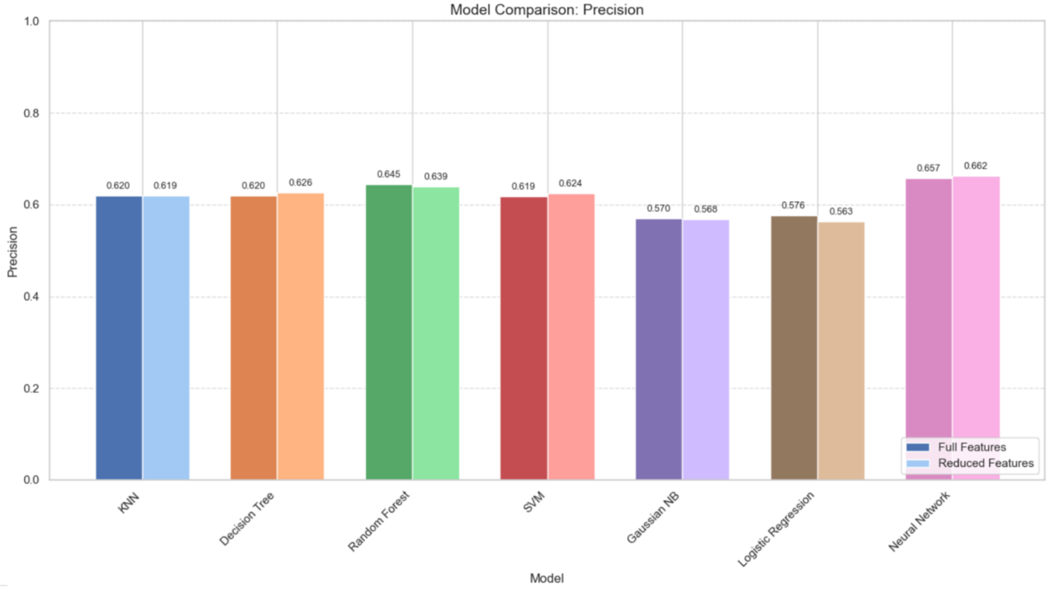
Precision Comparison:

- Neural Network (Reduced): 0.662 [Highest]
- Neural Network (Full): 0.657
- Random Forest (Full): 0.645
- Random Forest (Reduced): 0.639
- SVM (Reduced): 0.624

Observations:

- Tree-based models and neural networks lead in accuracy
- Random Forest edges out Neural Network by small margin
- Most models maintain accuracy within ~2% with reduced features
- 75% accuracy is strong given the challenging nature of the prediction task
- Neural Network provides most precise predictions
- Feature reduction has minimal impact on precision for most models





MODEL COMPARISON - RECALL & F1 SCORE

Recall Comparison:

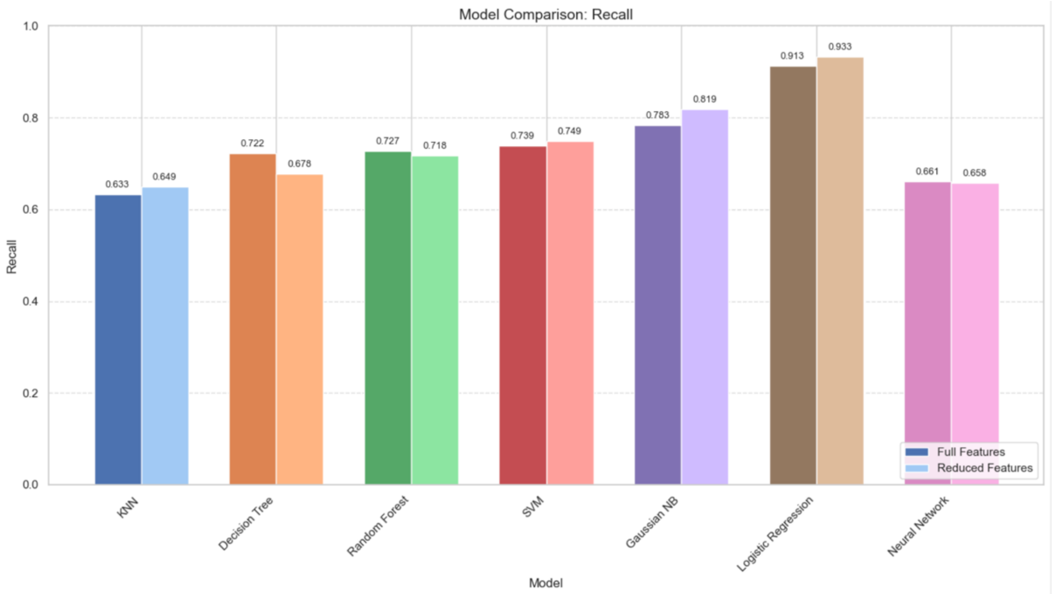
- Logistic Regression (Reduced): 0.933 [Highest]
- Logistic Regression (Full): 0.913
- Naive Bayes (Reduced): 0.819
- Naive Bayes (Full): 0.783
- SVM (Reduced): 0.749

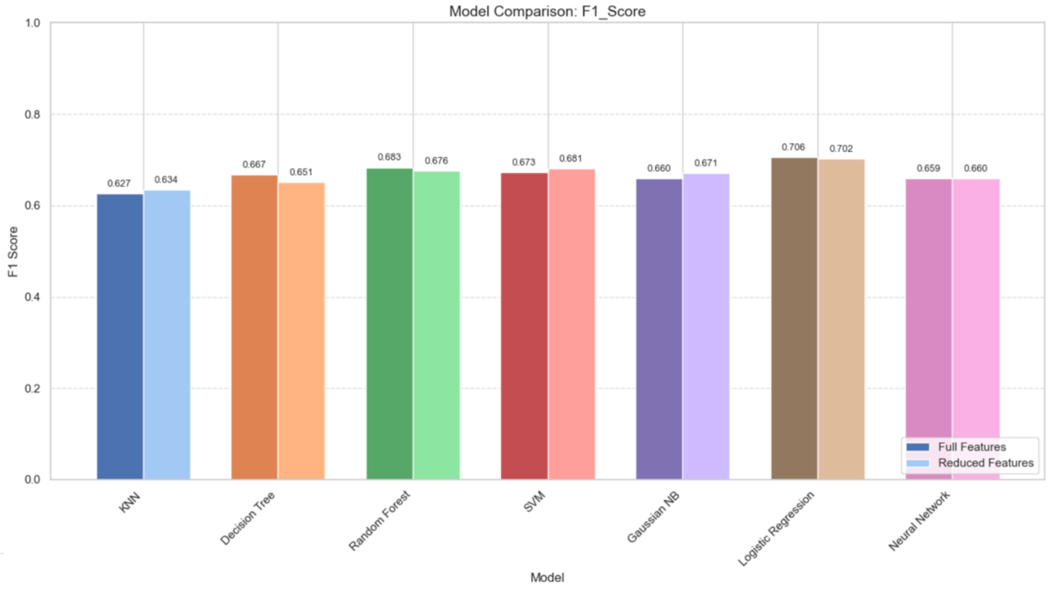
F1 Score Comparison:

- Logistic Regression (Full): 0.706 [Highest]
- Logistic Regression (Reduced): 0.702
- Random Forest (Full): 0.683
- SVM (Reduced): 0.681
- Random Forest (Reduced): 0.676

Observations:

- Logistic Regression and Naive Bayes excel at finding smokers
- Logistic Regression achieves best overall balance (F1)
- Different models have different strengths for specific metrics
- Model choice should depend on whether false positives or false negatives are more concerning
- All models show strong F1 scores (>0.65) indicating good overall performance

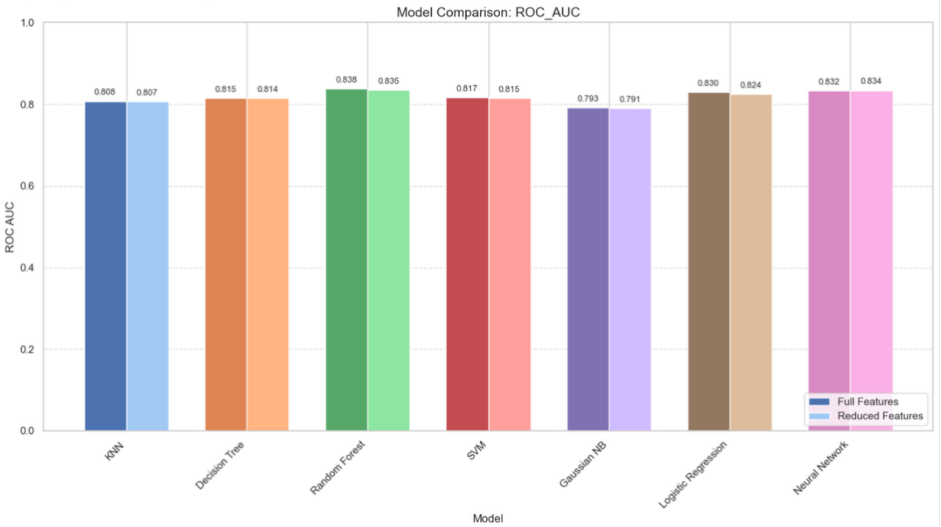




MODEL COMPARISON - ROC AUC & FEATURE IMPACT

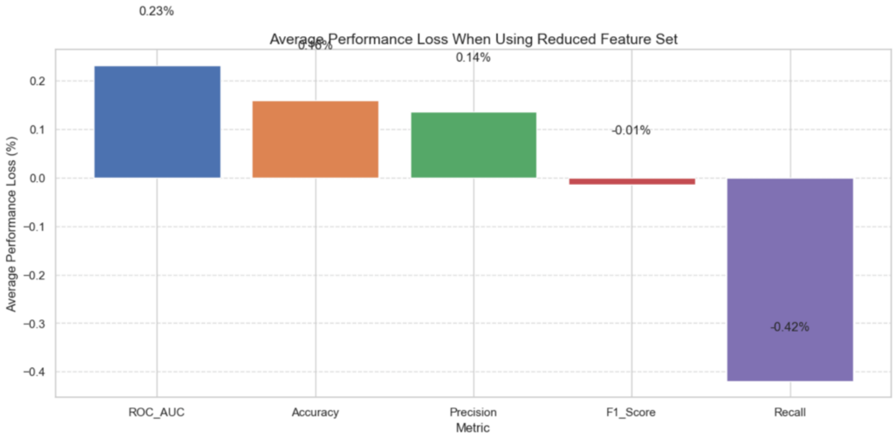
ROC AUC Comparison:

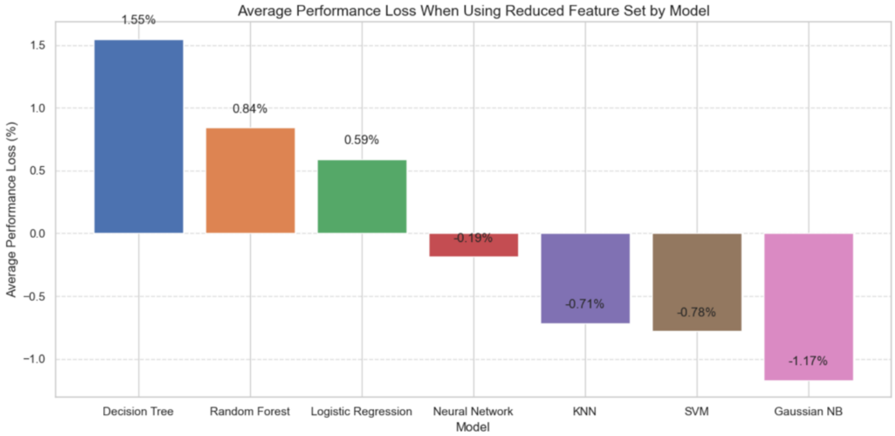
- Random Forest (Full): 0.838 [Highest]
- Neural Network (Reduced): 0.834
- Neural Network (Full): 0.832
- Logistic Regression (Full): 0.830
- Logistic Regression (Reduced): 0.824



Average Performance Change with Reduced Features:

- Decision Tree: -2.35%
- Logistic Regression: -0.58%
- Random Forest: -0.47%
- KNN: +0.31%
- Neural Network: +0.15%
- SVM: +1.13%
- Naive Bayes: +1.66%





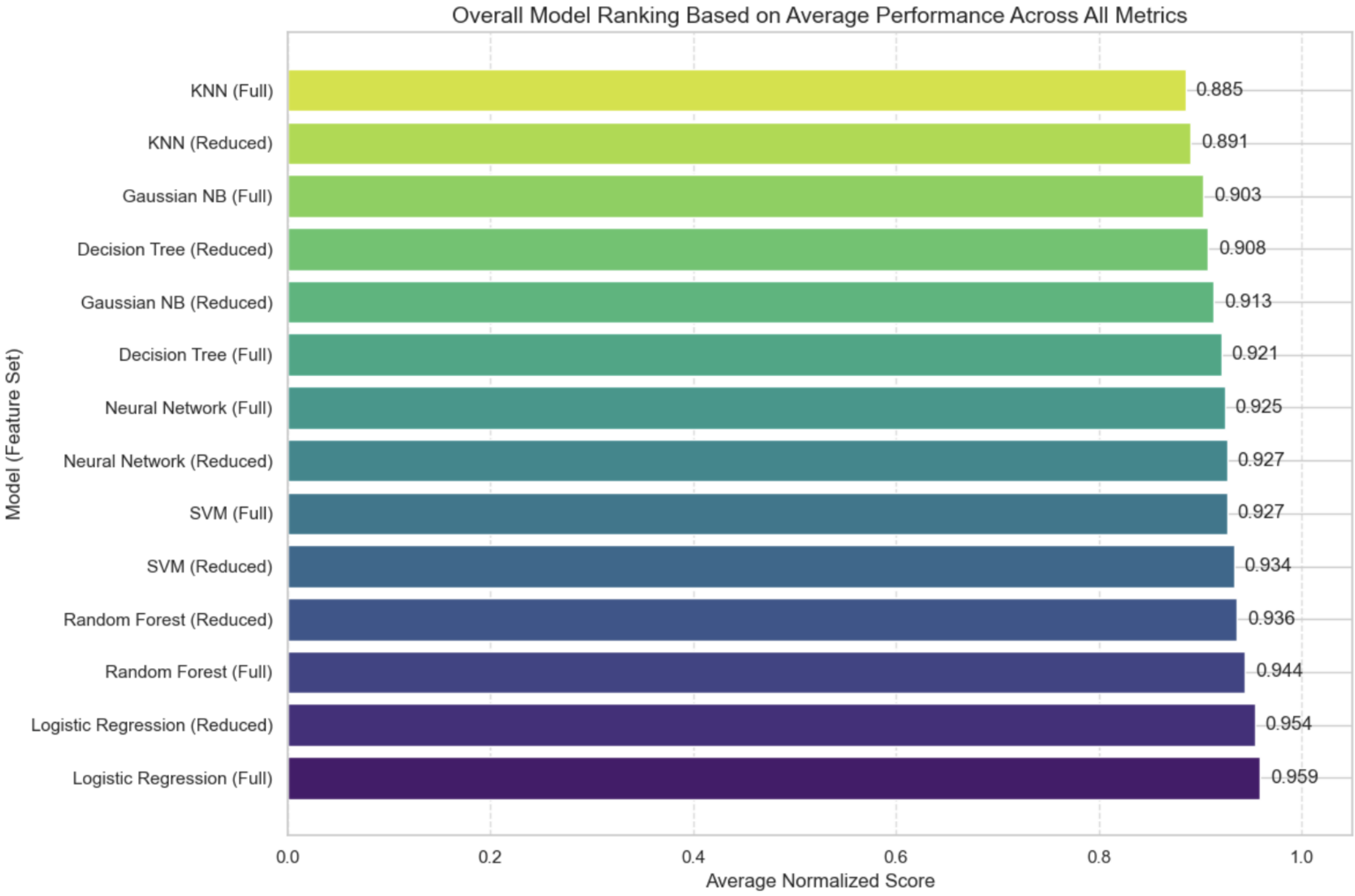
OVERALL MODEL RANKING

Normalized Score Ranking (Across All Metrics):

1. Logistic Regression (Full): 0.959
2. Logistic Regression (Reduced): 0.954
3. Random Forest (Full): 0.944
4. Random Forest (Reduced): 0.936
5. SVM (Reduced): 0.934
6. SVM (Full): 0.927
7. Neural Network (Reduced): 0.927
8. Neural Network (Full): 0.925

Key Observations:

- Logistic Regression provides best overall performance
- Random Forest ranked second overall
- Top models maintain strong performance with reduced features
- Feature reduction viable with minimal performance loss
- Simple models (Logistic Regression) outperform more complex ones



CONCLUSIONS AND RECOMMENDATIONS

Key Findings:

- Logistic Regression provides best overall performance (highest F1-score and recall)
- Neural Networks offer highest precision
- Random Forest achieves highest accuracy and ROC AUC
- 15 key features capture most predictive information

Most Important Predictors:

- Gender
- Liver enzymes (GTP, ALT)
- Hemoglobin
- Height and triglyceride levels

Recommendations:

- For balanced performance: Logistic Regression
- For minimizing false positives: Neural Network
- For highest discrimination ability: Random Forest
- Reduced feature set (15 features) viable for most applications

FUTURE WORK

- Explore ensemble methods combining strengths of different models
- Implement advanced hyperparameter tuning techniques
- Test more complex neural network architectures
- Validate on external datasets to ensure generalizability
- Develop a clinically-applicable prediction tool
- Investigate additional biological indicators potentially linked to smoking
- Explore interpretability techniques for complex models
- Extend model to predict smoking intensity (e.g., light vs. heavy smokers)

REFERENCES

- Body Signal of Smoking Dataset: https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking/data
- Scikit-learn Documentation: https://scikit-learn.org/stable/
- Pandas Documentation: https://pandas.pydata.org/docs/
- Matplotlib & Seaborn: https://matplotlib.org/ & https://seaborn.pydata.org/
- NumPy Documentation: https://numpy.org/doc/
- ydata-profiling: https://github.com/ydataai/ydata-profiling

STEVENS
INSTITUTE OF TECHNOLOGY
1870

THANK **YOU**

Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030

STEVENS
INSTITUTE OF TECHNOLOGY
1870

ANY QUESTIONS?

Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030