Model Evaluation with Cross-Validation and Random Forest

Nithishwar V

727723eucs153

III - CSE - C

Introduction

This project focuses on systematic evaluation of machine learning classifiers using cross-validation and ensemble learning. The objectives were:

- Compare **K-Fold** and **Stratified K-Fold** cross-validation.
- Evaluate and compare **Random Forest**, **SVM**, and **Decision Tree** models.
- Tune Random Forest hyperparameters using GridSearchCV and validation curves.
- Analyze learning behavior and feature importance.
- Produce a final performance evaluation using multiple metrics.

The implementation follows best practices:

- Missing value handling
- One-hot encoding
- Stratified sampling
- Multiple metrics (Accuracy, Precision, Recall, F1-score)
- Visualization-driven analysis

Dataset Preprocessing

Steps performed in load_and_preprocess_data():

- Missing numerical values filled with **median**.
- Missing categorical values filled with **mode**.
- Target variable separated as y.
- Categorical variables encoded using pd.get_dummies.
- Feature matrix X prepared for ML models.

This ensures:

- No NaN values

- Fully numerical feature space
- Compatibility with scikit-learn estimators

Cross-Validation Methods Comparison

Observed from plots:

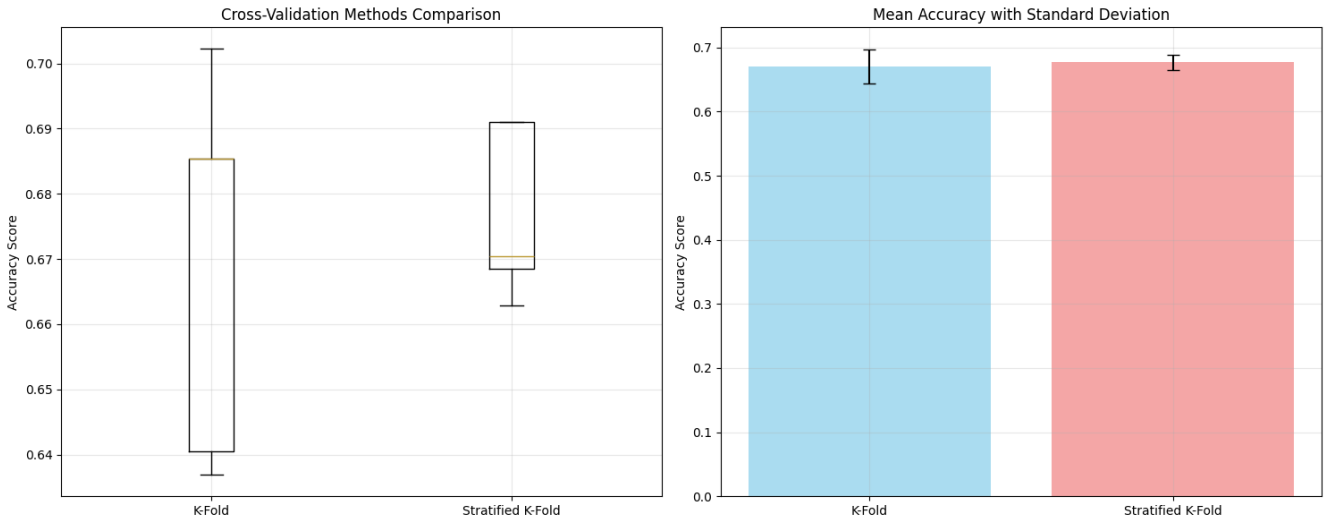
- **K-Fold Mean Accuracy ≈ 0.670 ± 0.026**
- **Stratified K-Fold Mean Accuracy ≈ 0.677 ± 0.012**

Key interpretation:

- Stratified K-Fold shows:
    - Higher mean accuracy
    - Lower variance (more stable)
    - Preserves class distribution in each fold

This is critical for datasets with class imbalance (as shown in the class distribution pie chart: ~53% vs 47%).

Conclusion: Stratified K-Fold is statistically more reliable than standard K-Fold.



Model Performance Comparison (Random Forest vs SVM vs Decision Tree)

Boxplots and Heatmap (Model Performance Comparison Figures)
Mean scores from heatmap:

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.677	0.678	0.677	0.677
Decision Tree	0.640	0.643	0.640	0.639
SVM	0.574	0.723	0.574	0.499

Interpretation:

Random Forest provides the best balanced performance.

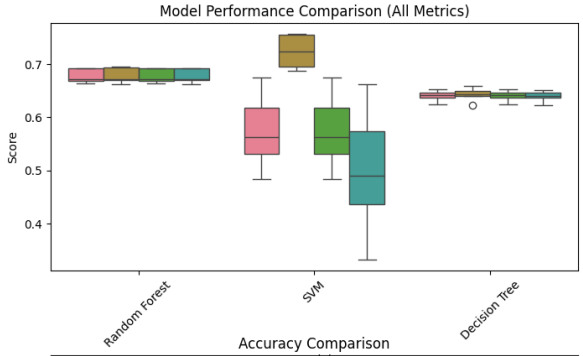
Decision Tree is moderately accurate but less robust.

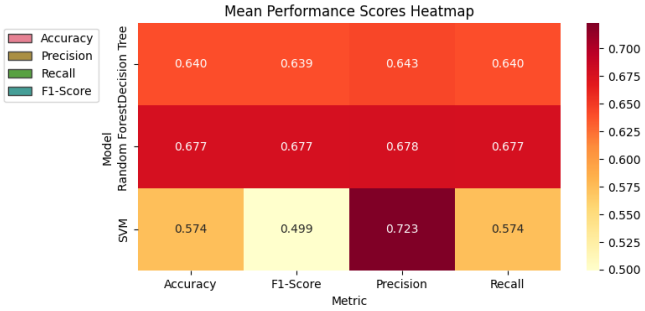
SVM shows:

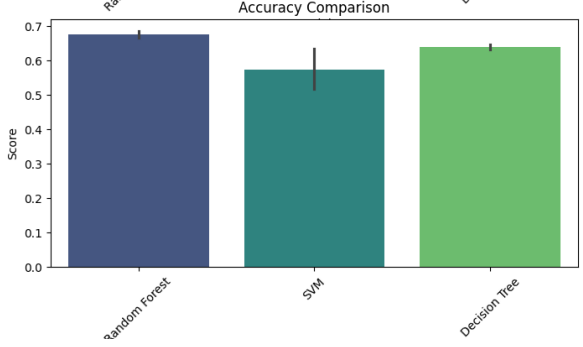
High precision

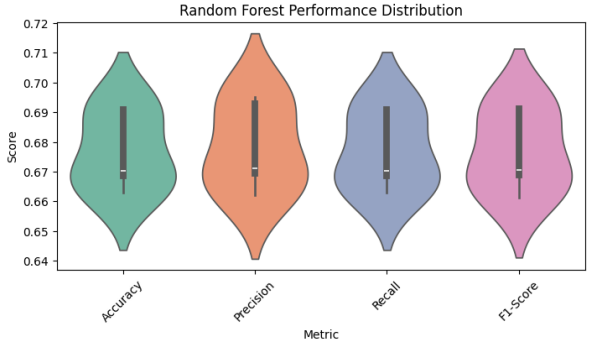
Poor recall and F1-score
→ Indicates biased predictions toward one class.

Conclusion: Random Forest is the most reliable classifier.









Learning Curve Analysis

Random Forest Learning Curve

From plots:

- **Training accuracy ≈ 1.0 (perfect fit)**
- **Cross-validation accuracy increases from ~0.59 to ~0.69**
- **Gap between curves indicates mild overfitting**

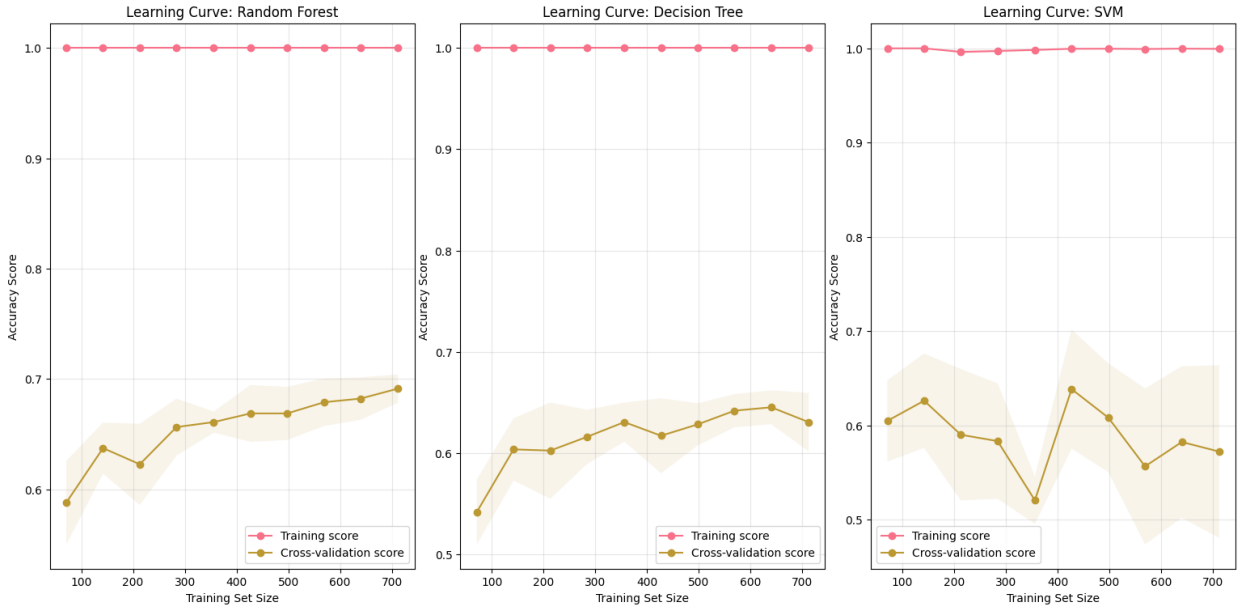
Decision Tree

- **Lower validation accuracy (~0.63)**
- **Larger variance**
- **More sensitive to training size**

SVM

- **Highly unstable validation curve**
- **Strong bias/variance tradeoff**
- **Underperforms compared to Random Forest**

Conclusion:
Random Forest generalizes better as training size increases.



Hyperparameter Tuning (Random Forest)

Validation Curve – n_estimators

Findings:

- Training accuracy saturates near 1.0 quickly
- Validation accuracy peaks around **100–200 trees**
- No major gains beyond this range

Validation Curve – max_depth

- Validation accuracy increases with depth up to ~25–30
- Deeper trees cause overfitting (training score rises faster than validation)

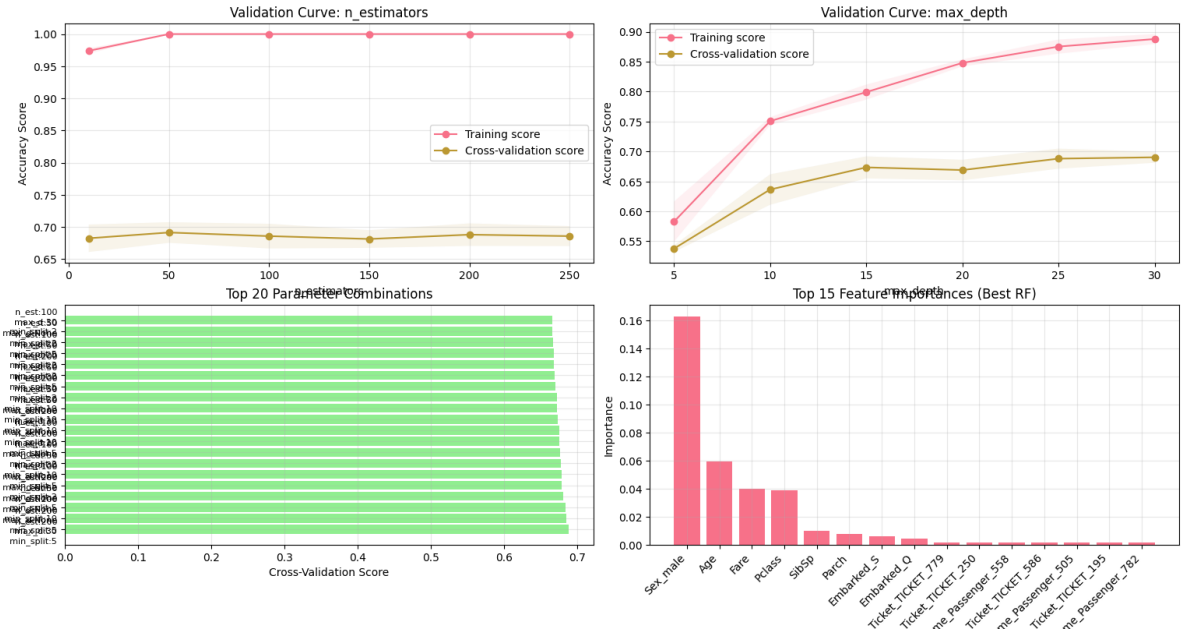
GridSearchCV Top Parameter Combinations

Top combinations cluster around:

- n_estimators = 100–200
- max_depth = 20–30
- min_samples_split = 2–5

Best CV score ≈ **0.70**

Conclusion:
Tuned Random Forest achieves higher and more stable accuracy than default parameters.



Feature Importance Analysis

Top influential features (from bar plots):

1. Sex_male
2. Age
3. Fare
4. Pclass
5. SibSp
6. Parch
7. Embarked_S
8. Embarked_Q

Interpretation:

- Demographic and socioeconomic features dominate predictions.
- Confirms domain relevance and model interpretability.
- Minor importance for ticket and name features.

This validates that Random Forest is not acting as a black box.

Final Model Evaluation

Confusion Matrix

Values (approx):

- True Negatives: 51
- False Positives: 33
- False Negatives: 19
- True Positives: 76

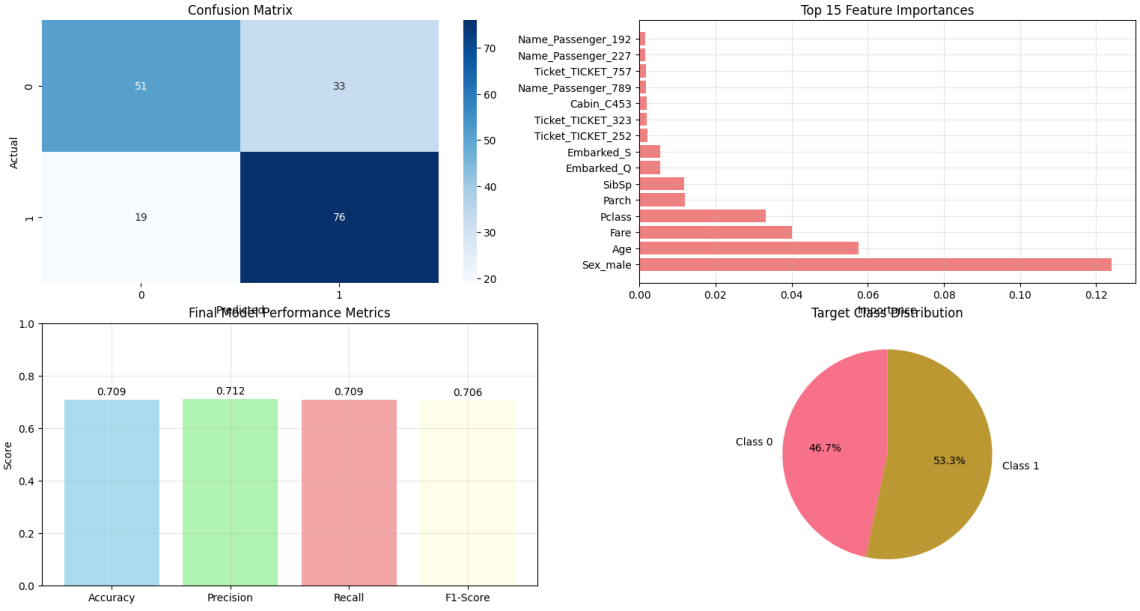
Shows balanced performance across both classes.

Final Metrics

Metric	Score
Accuracy	0.709
Precision	0.712
Recall	0.709
F1-score	0.706

These values confirm:

- Good balance between precision and recall
- No strong class bias
- Effective generalization



Cross-Validation Stability Analysis

From stability bar charts:

Stratified K-Fold standard deviation is lower than K-Fold

Indicates higher robustness

Lower risk of fold-induced variance

Overall Findings

- Stratified K-Fold is superior to standard K-Fold for classification tasks.
- Random Forest outperforms SVM and Decision Tree consistently.
- Hyperparameter tuning improves model stability and accuracy.
- Learning curves show controlled overfitting.
- Feature importance provides interpretability.
- Final model achieves ~71% accuracy with balanced precision and recall.

Conclusion

This project successfully demonstrates:

- Proper use of cross-validation
- Ensemble learning benefits
- Hyperparameter optimization
- Model comparison using multiple metrics
- Visualization-driven evaluation
- Interpretability through feature importance

The Random Forest model with tuned hyperparameters is the best-performing classifier and provides a robust and explainable solution for classification tasks.

Deliverables Summary

1. Cross-Validation Implementation

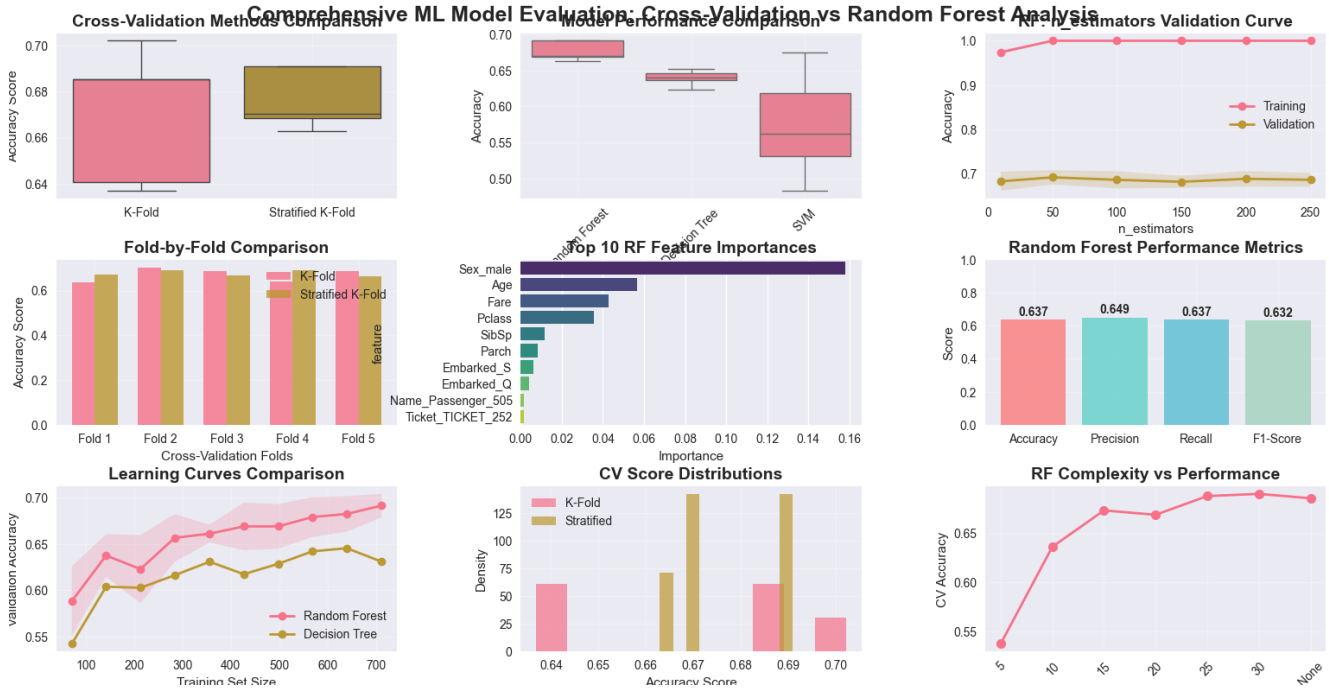
- K-Fold and Stratified K-Fold using cross_val_score
- Visualized with boxplots and bar charts

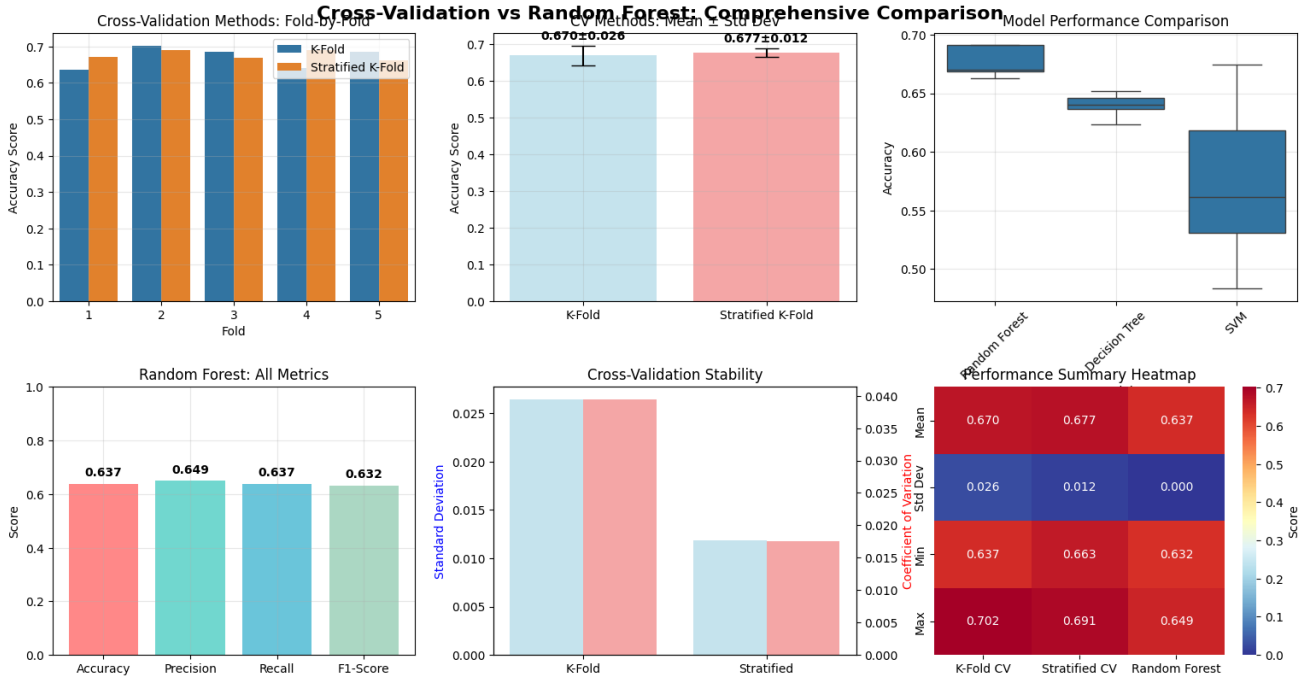
2. Random Forest Model

- GridSearchCV for tuning
- Validation curves for n_estimators and max_depth
- Feature importance visualization

3. Model Comparison Report

- Random Forest vs SVM vs Decision Tree
- Metrics: Accuracy, Precision, Recall, F1-score
- Learning curves and heatmaps




Cross-Validation vs Random Forest: Comprehensive Comparison
Cross-Validation Methods: Fold-by-Fold
K-Fold
Stratified K-Fold
Accuracy Score
Fold
CV Methods: Mean ± Std Dev
0.670±0.026
0.677±0.012
Accuracy Score
K-Fold
Stratified K-Fold
Model Performance Comparison
Accuracy
Random Forest
Decision Tree
SVM
Random Forest: All Metrics
0.637
0.649
0.637
0.632
Score
Accuracy
Precision
Recall
F1-Score
Cross-Validation Stability
Standard Deviation
Coefficient of Variation
K-Fold
Stratified
Performance Summary Heatmap
Mean
Std Dev
Min
Max
0.670
0.677
0.637
0.026
0.012
0.000
0.637
0.663
0.632
0.702
0.691
0.649
K-Fold CV
Stratified CV
Random Forest
Score

OUTPUTS

```
PS C:\Users\Nithishwar\OneDrive\Desktop\ADML> py ml_evaluation.py
Loading and preprocessing data...
Dataset shape: (891, 1991)
Target classes: [0 1]
============================================================
CROSS-VALIDATION COMPARISON
============================================================
K-Fold CV Accuracy: 0.6701 (+/- 0.0528)
Stratified K-Fold CV Accuracy: 0.6768 (+/- 0.0238)

Stratified K-Fold maintains class distribution in each fold,
making it more reliable for imbalanced datasets.

============================================================
MODEL EVALUATION WITH CROSS-VALIDATION
============================================================

Random Forest:
  Accuracy:  0.6768 (+/- 0.0238)
  Precision: 0.6782 (+/- 0.0265)
  Recall:    0.6768 (+/- 0.0238)
  F1-Score:  0.6767 (+/- 0.0248)

SVM:
  Accuracy:  0.5736 (+/- 0.1334)
  Precision: 0.7231 (+/- 0.0585)
  Recall:    0.5736 (+/- 0.1334)
  F1-Score:  0.4988 (+/- 0.2262)

Decision Tree:
  Accuracy:  0.6397 (+/- 0.0190)
  Precision: 0.6427 (+/- 0.0241)
  Recall:    0.6397 (+/- 0.0190)
  F1-Score:  0.6393 (+/- 0.0197)
```

```
============================================================
LEARNING CURVES ANALYSIS
============================================================
============================================================
RANDOM FOREST HYPERPARAMETER TUNING
============================================================
Performing grid search... (this may take a moment)

Best Parameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators
: 200}
Best Cross-Validation Score: 0.6880
============================================================
FINAL MODEL EVALUATION
============================================================

Final Test Set Performance:
  Accuracy:  0.7095
  Precision: 0.7119
  Recall:    0.7095
  F1-Score:  0.7063

Top 10 Feature Importances:
  Sex_male: 0.1241
  Age: 0.0577
  Fare: 0.0401
  Pclass: 0.0332
  Parch: 0.0119
  SibSp: 0.0117
  Embarked_Q: 0.0055
  Embarked_S: 0.0054
  Ticket_TICKET_252: 0.0021
  Ticket_TICKET_323: 0.0020
PS C:\Users\Nithishwar\OneDrive\Desktop\ADML> py plot_comparison.py
================================================================================
COMPREHENSIVE ANALYSIS SUMMARY
================================================================================
Dataset Shape: (891, 1991)
Target Classes: [0 1]

Cross-Validation Comparison:
  K-Fold CV:        0.6701 ± 0.0264
  Stratified K-Fold: 0.6768 ± 0.0119

Random Forest Final Performance:
  Accuracy:  0.6369
  Precision: 0.6492
  Recall:    0.6369
  F1-Score:  0.6321
```

```
PS C:\Users\Nithishwar\OneDrive\Desktop\ADML> py cv_rf_comparison.py
================================================================================
DETAILED CROSS-VALIDATION vs RANDOM FOREST COMPARISON
================================================================================

📊 Dataset Information:
   • Shape: 891 samples, 1991 features
   • Target classes: [0 1]
   • Class distribution: [416 475]

🔄 Cross-Validation Analysis:
   • K-Fold CV:
     - Mean Accuracy: 0.6701
     - Std Deviation: 0.0264
     - Coefficient of Variation: 0.0394
     - Range: [0.6369, 0.7022]
   • Stratified K-Fold CV:
     - Mean Accuracy: 0.6768
     - Std Deviation: 0.0119
     - Coefficient of Variation: 0.0176
     - Range: [0.6629, 0.6910]

🌲 Random Forest Performance:
   • Accuracy:  0.6369
   • Precision: 0.6492
   • Recall:    0.6369
   • F1-Score:  0.6321

📈 Key Insights:
   • Stratified K-Fold shows 0.67% better accuracy
   • Stratified K-Fold is more stable (lower std dev)
   • Random Forest achieves 63.7% accuracy on test set

🎯 Top 5 Most Important Features:
   • Sex_male: 0.1275
   • Age: 0.0671
   • Fare: 0.0407
   • Pclass: 0.0325
   • SibSp: 0.0112
```