CC4051 Machine Learning - FCUP

Model Comparison and Analysis

Camila Alves, Miguel Goncalves, Nina Lichtenberger

Faculty of Computer Science, University of Porto

24.04.2025

Introduction

- Comparative analysis of classification models:
  - Logistic Regression, LDA, QDA
  - Decision Trees (pruned/unpruned)
  - SVM (linear/RBF kernels)
- Key objectives:
  - Understand model assumptions and behaviors
  - Analyze bias-variance tradeoff
  - Evaluate ensemble methods (Bagging, Random Forest, Boosting)
- Methodology:
  - Artificial datasets highlighting model strengths/weaknesses
  - Real-world datasets for validation
  - Python implementation with cross-validation

Model Testing Framework

Dataset Generation

- Controlled experiments with:
  - Class distribution
  - Boundary complexity
  - Noise levels
  - Class overlap

Evaluation Metrics

- Cross-validated accuracy
- Confusion matrices
- Classification reports
- Bias-variance decomposition

Logistic Regression

Assumptions

- Linearity in log-odds
- Independent observations
- No multicollinearity

Artificial Dataset

- Linearly separable features
- Structured noise
- Minimal multicollinearity

Model	Accuracy
Logistic Regression	0.978
Linear SVM	0.977
LDA	0.965
QDA	0.953
RBF SVM	0.913
Decision Tree	0.843
Decision Tree (Max Depth = 2)	0.781

Logistic Regression: Performance

Model Comparison

- **LDA (96.5%), QDA (95.3%)** — strong, but assume Gaussian distribution
- **RBF SVM (91.3%)** — might overfit due to lack of non-linearity
- **Decision Trees (84.3%, 78.1%)** — sensitive to data structure, lowest accuracy

Influence of Data Properties

- Minimal effect: sample size, class imbalance, noise, class overlap, class distribution
- **Major effect**: class distribution
  - Multi-class → accuracy drops **from 98% to 80%**
- **Non-linear boundaries:** Logistic regression fails:
  - Only 52% on XOR dataset

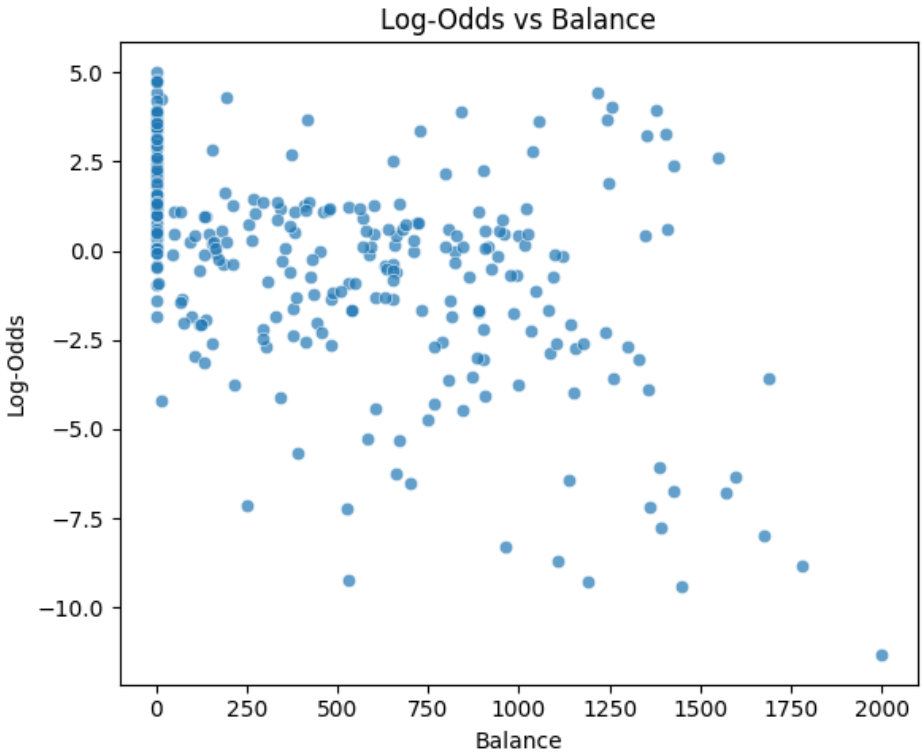
Logistic Regression Implementation

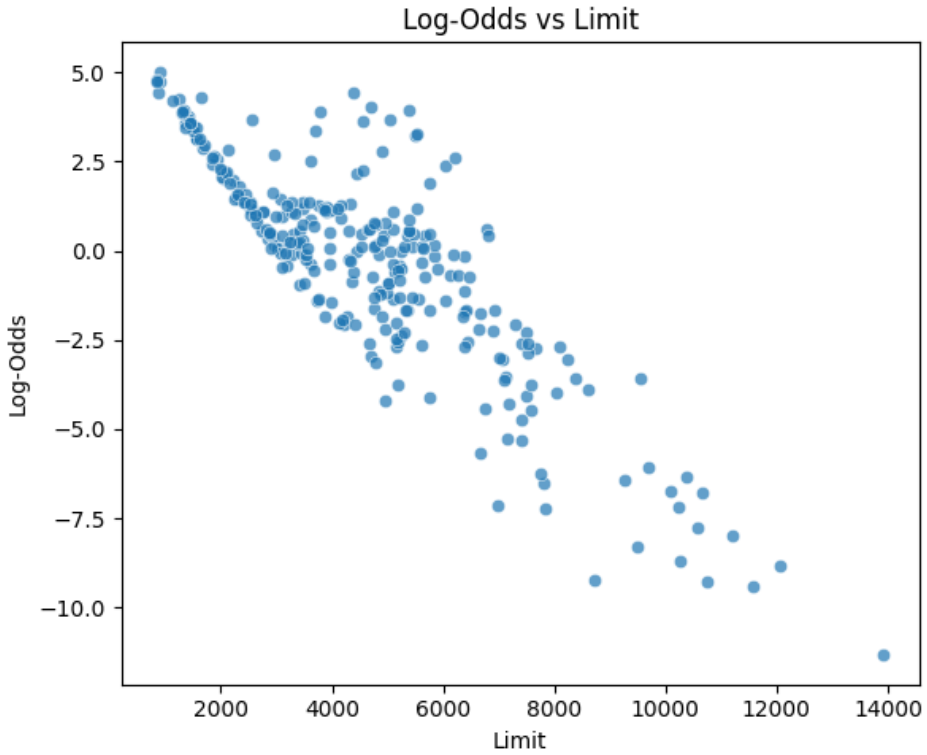
```
# Generate linearly separable data
true_coefficients = np.array([3, -2, 1, 2.5, -3])
linear_combination = X @ true_coefficients
probabilities = 1 / (1 + np.exp(-linear_combination))
y = (probabilities > 0.5).astype(int)

# Add structured noise
correlated_noise = 0.8 * X + np.random.normal(0, 0.05, X.shape)
irrelevant_features = np.random.normal(0, 1, (n_samples, 5))
X_noisy = np.hstack((X, correlated_noise, irrelevant_features))

# Evaluate models
models = {
    "Logistic Regression": LogisticRegression(),
    "LDA": LDA(),
    "QDA": QDA(),
    "Linear SVM": SVC(kernel="linear")
}
cv_results = {}
for name, model in models.items():
    cv_scores = cross_val_score(model, X_noisy, y, cv=5, scoring='
    accuracy')
    cv_results[name] = cv_scores.mean()
```

Logistic Regression: Real Dataset

- Credit Card Balance Dataset
  - information on 10,000 credit card customers (demographic and credit card use)
  - target variable: income class
  - dataset was used for Logistic Regression, LDA, as well as QDA due to similar assumptions
- Performance:
  - Accuracy of 86.25% - Balanced precision, recall, and F1-score
  - outperformed by linear SVM, LDA, and QDA
  - hints at similar use cases of logistic regression, LDA, and QDA

Logistic Regression: Real Dataset Assumptions





→ Multicollinearity can be checked with VIF values

→ Balance and Limit showed some correlation with VIFs around 4

Linear Discriminant Analysis (LDA)

Assumptions

- Normality of predictors
- Homogeneous covariances
- Well-separated classes

Artificial Dataset

- Gaussian class distributions
- Shared covariance matrix
- Class-dependent noise to disrupt QDA
- Mild multicollinearity

Model	Accuracy
LDA	0.953
Logistic Regression	0.951
Linear SVM	0.948
RBF SVM	0.943
QDA	0.937
Decision Tree	0.868
Decision Tree (Max Depth = 2)	0.805

LDA: Performance

Model Comparison

- **Logistic Regression (95.1%)**, **Linear SVM (94.8%)** — very close, similar assumptions
- **QDA (93.7%)** — suffered due to class-dependent noise in covariance
- **Decision Trees (86.8%, 80.5%)** — sensitive to noise, lowest accuracy

Influence of Data Properties

- Minimal effect: sample size, class overlap and noise
- **Major effect**:
  - multi-class: for 3 classes → accuracy drops **from 95% to 78%**
  - data distribution: for shifted data distribution → accuracy drops **from 95% to 75%**
  - non linear boundaries: Only 52% on XOR dataset
  - class imbalance → F1-score **0.00**

LDA: Credit Card Dataset

- Shapiro-Wilk Test shows that features are normally distributed
- covariance matrices are not identical, but Levene's Test suggests that the differences are not significant
  $\rightarrow$ assumptions might be fulfilled - LDA and QDA could both be the fitting model
- accuracy of 88.50%
- only outperformed by the linear SVM

Class 0 Covariance Matrix

$$\begin{bmatrix} 1.0966 & -0.0364 & 0.0264 & 0.3135 & 0.2206 \\ -0.0364 & 1.2358 & -0.0349 & 0.1483 & -0.0144 \\ 0.0264 & -0.0349 & 0.8992 & 0.0117 & 0.1574 \\ 0.3135 & 0.1483 & 0.0117 & 1.0027 & 0.1076 \\ 0.2206 & -0.0144 & 0.1574 & 0.1076 & 0.9404 \end{bmatrix}$$

Class 1 Covariance Matrix

$$\begin{bmatrix} 0.8868 & 0.0436 & 0.0926 & -0.1331 & -0.0400 \\ 0.0436 & 1.0331 & 0.0050 & -0.0469 & -0.0273 \\ 0.0926 & 0.0050 & 0.9399 & 0.1028 & 0.0522 \\ -0.1331 & -0.0469 & 0.1028 & 1.0027 & -0.0883 \\ -0.0400 & -0.0273 & 0.0522 & -0.0883 & 0.7887 \end{bmatrix}$$

Quadratic Discriminant Analysis (QDA)

Assumptions

- Normality of predictors
- Class-specific covariances
- Non-linear boundaries

Artificial Dataset

- Normally distributed features
- Class-specific covariance
- Non-linear decision boundaries
- Multicollinearity to challenge other models

Model	Accuracy
QDA	0.933
RBF SVM	0.927
Decision Tree	0.896
Linear SVM	0.867
LDA	0.860
Logistic Regression	0.859
Decision Tree (Max Depth = 2)	0.882

QDA: Performance

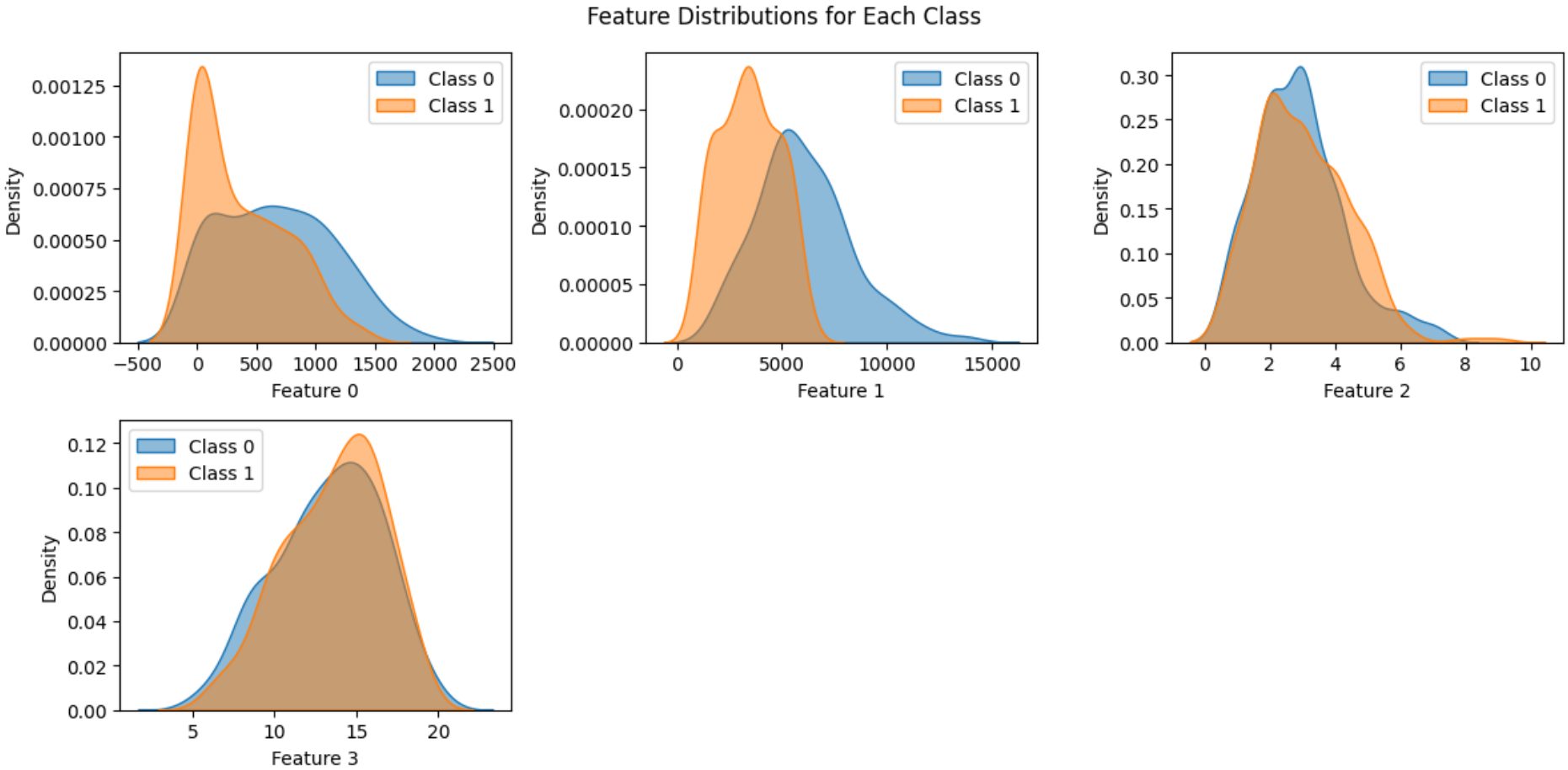
Model Comparison

- **RBF SVM (92.7%)**, **Decision Tree (89.6%)** — ability to capture non-linearity
- **SVMs** — performance was decreased by adverse correlations in the two classes and outiers

Influence of Data Properties

- Minimal effect: sample size
- **Major effect**:
  - Class overlap → accuracy from **93% to 76%**
  - High noise level → accuracy from **93% to 73%**
  - Class imbalance → F1-score **0.00**
  - Multi-class setting → accuracy from **93% to 74%**
  - Data distribution shift → accuracy from **93% to 58%**

QDA: Credit Card Dataset



- Covariance differences were not significant according to Levene's test - but high Frobenius norm and plots indicate class-specific covariance
- similar performance of QDA and LDA (accuracy of 88.5% vs 88.7%).

Decision Trees

Assumptions

no specific assumptions like other models

Artificial Dataset

- Hierarchical feature interactions
- Non-linear boundaries
- Irrelevant features
- for DT with depth = 2: only one hierarchical rule

Model Without Pruning	Accuracy
Decision Tree	0.915
RBF SVM	0.805
QDA	0.793
Logistic Regression	0.745
LDA	0.739
Linear SVM	0.737
Decision Tree (Max Depth = 2)	0.727

Model With Pruning	Accuracy
Decision Tree (Max Depth = 2)	0.882
RBF SVM	0.825
Linear SVM	0.822
LDA	0.900
Logistic Regression	0.813
QDA	0.798
Decision Tree	0.759

Decision Trees: Performance

Model Comparison

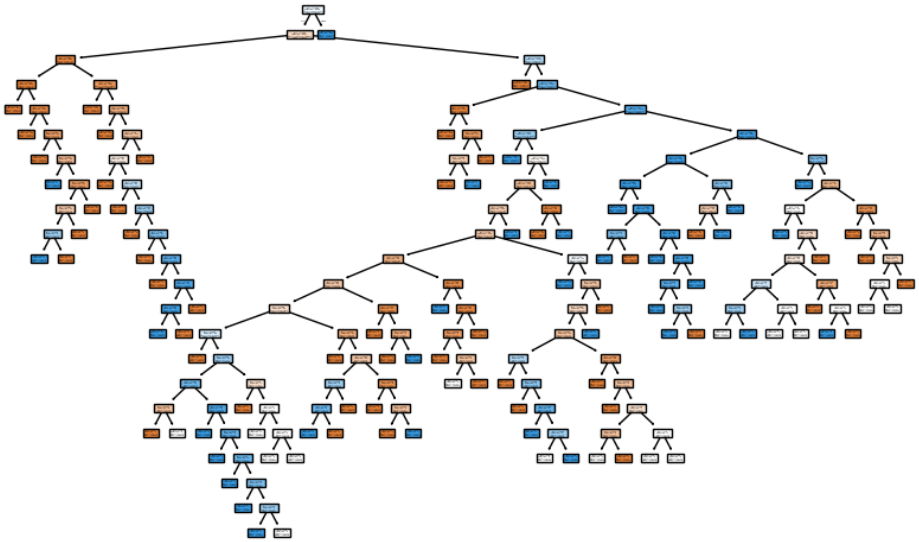
- **RBF SVM is main competitor** due to non-linearity, suffers from irrelevant features
- **Linear Models** perform significantly better on the dataset with one hierarchical rule
- **Unpruned DT** is worst-performer for one rule - overfitting!

Influence of Data Properties

- Noise $\rightarrow$ influences **unpruned DT** more than other models - overfitting!
- Less hierarchical rules $\rightarrow$ **improves performance of the other models**
- Irrelevant features $\rightarrow$ DTs were better at **feature selection** than the other models

Unpruned Decision Trees: Real Dataset

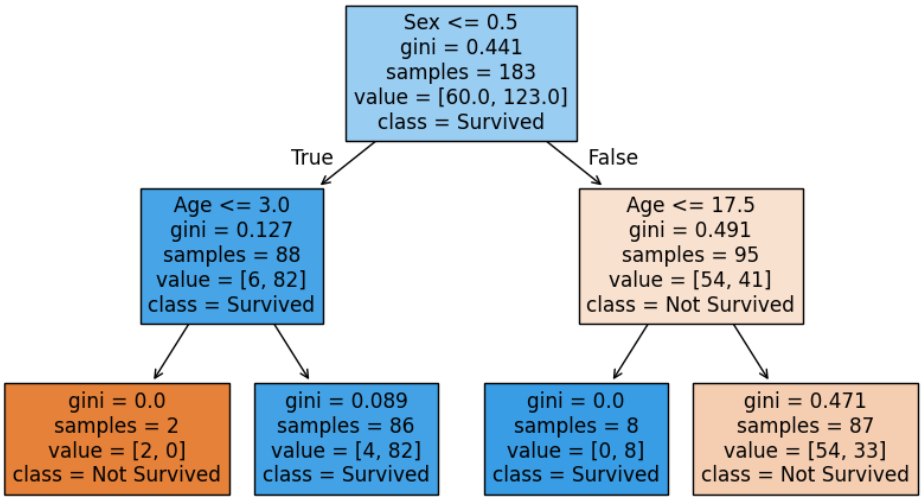
- Chess (King-Rook vs. King-Pawn) Dataset: typical decision tree scenario due to the following properties:
  - Rule-Based Nature
  - Nonlinear Relationships
  - Hierarchical Decision-Making



- Very large depth
- Does not seem to overfit
- Accuracy of 95.5%
- Outperforms all other models (followed by SVMs)

Pruned Decision Tree: Real Dataset

- Titanic Survival Dataset: ideal for a shallow Decision Tree due to the following properties:
  - Strong Hierarchical Structure
  - Nonlinear Relationships
  - Simple yet Effective Splits: "Women and Children First Policy"



- slightly outperforms the other models (only minor performance differences)
- only few real world cases in which relationships are that easy

Support Vector Machines (Linear and RBF)

Assumptions

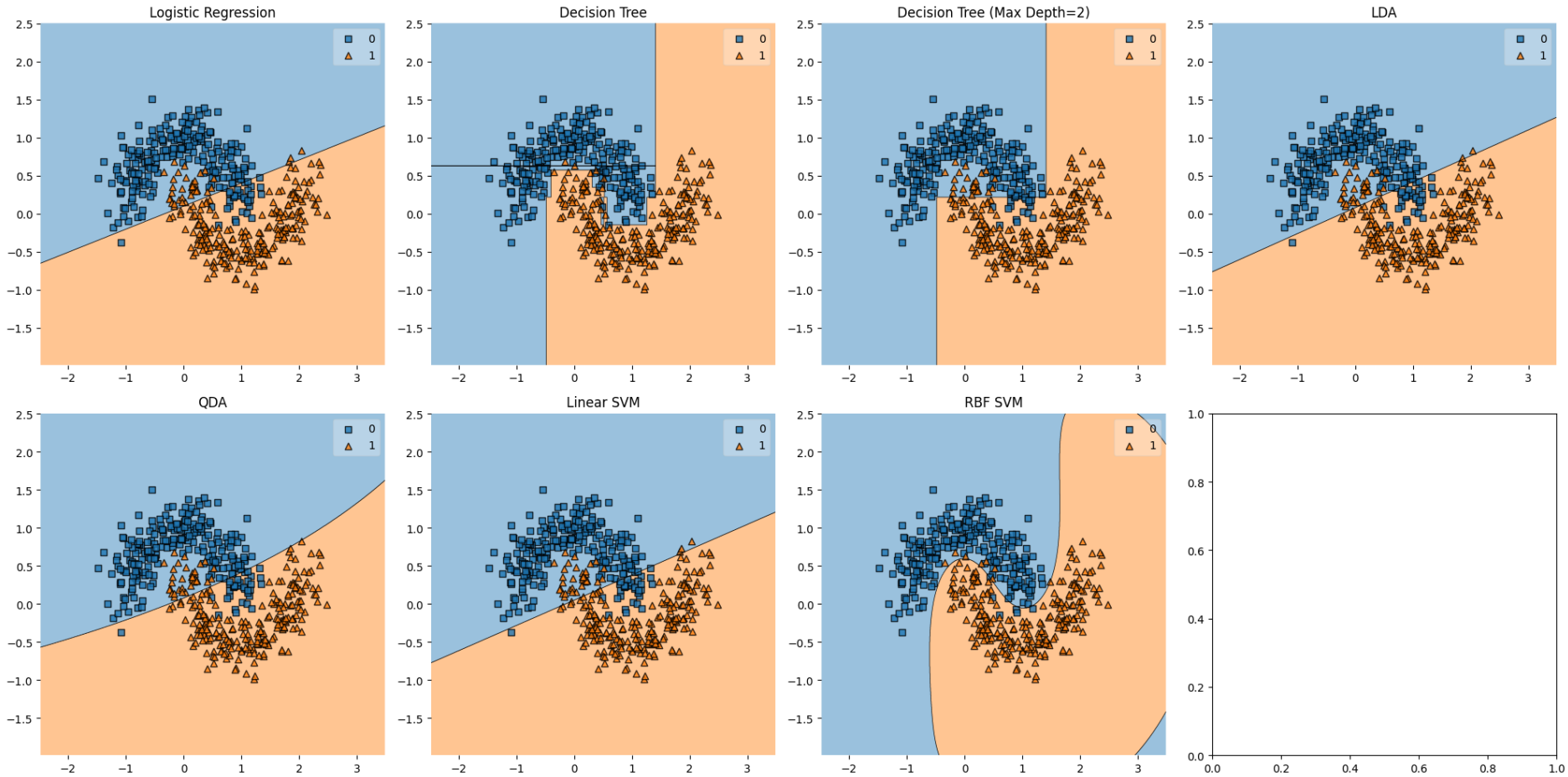
- Linearly separable (RBF:in higher dimensional space)
- Margin maximization

Artificial Dataset

- Linear class boundaries (linear) - non-linear boundaries (RBF)
- High feature informativeness
- Minimal noise and clear classes
- Feature redundancy
- Feature standardization (RBF)

Model	**Accuracy**
Linear SVM	0.945
Logistic Regression	0.935
RBF SVM	0.935
Decision Tree (Max Depth = 2)	0.925
Decision Tree	0.915
LDA	0.900
QDA	0.890

RBF SVM Dataset: Decision Boundary Comparison



- With **very high noise**, other models outperform the RBF SVM despite the non-linear data structure
- RBF SVM handles **complexity** and **feature selection** better

Linear SVM: Real Dataset

- Breast Cancer Wisconsin Dataset
- classes are well separated by few features
- linear separability assumption is reasonable (benign tumors tend to low values, malignant ones to higher values)

Model	Accuracy
RBF SVM	0.970
Linear SVM	0.967
LDA	0.957
QDA	0.955
Decision Tree (Max Depth = 2)	0.940
Decision Tree	0.925

Conclusions

- Outperformed by RBF SVM
  $\rightarrow$ no perfect linear separability - performance profits from transformation into higher-dimensional space
- Separability appears to be close to linear considering the accuracy of 96.7%

RBF SVM: Real Dataset

- Banknote Authentication Dataset
- Assumption: some fake banknotes are very similar to real ones, others are not — requires high flexibility

Model	Accuracy
RBF SVM	0.999
Linear SVM	0.985
LDA	0.976
QDA	0.985
Decision Tree (Max Depth = 2)	0.917
Decision Tree	0.984
Logistic Regression	0.981

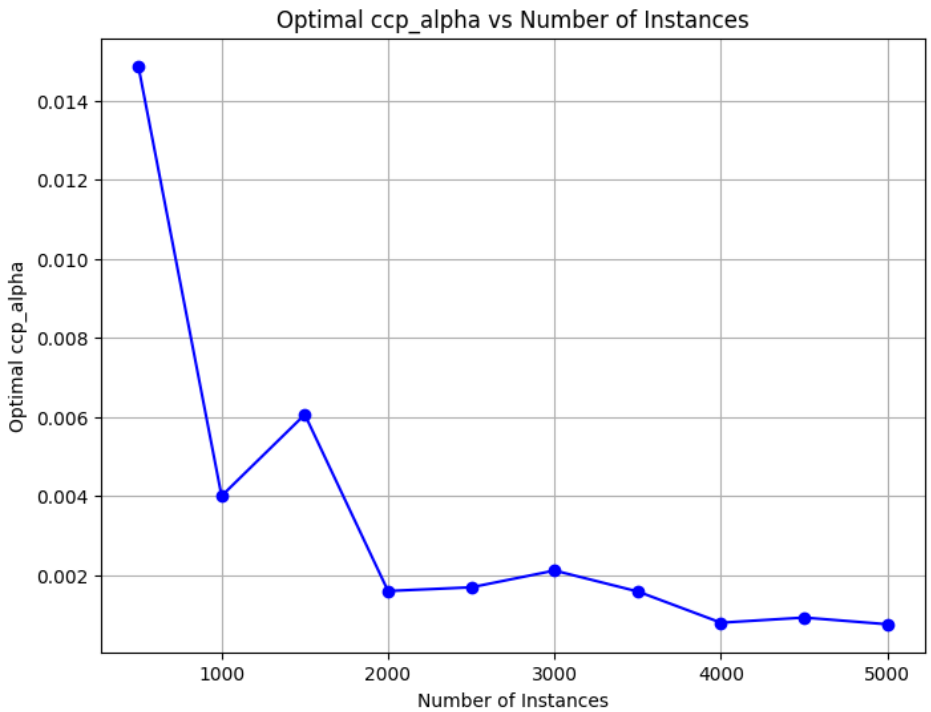
Conclusions

- RBF SVM achieves near-perfect accuracy (0.999)
- Most other models also show strong performance — hints at linear separability

Bias-Variance Tradeoff: Influence of Sample Size

Decision Tree Pruning

- Sample size was parameter with highest influence on ccp_alpha
- Optimal ccp_alpha decreases exponentially with dataset size
- Bigger datasets allow more complexity

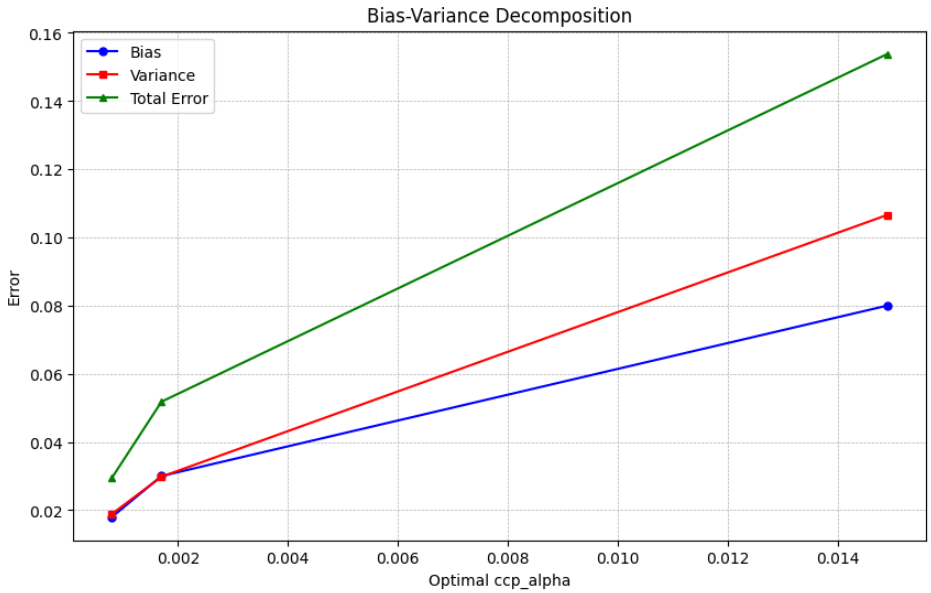


Samples	ccp_alpha	Bias	Variance
500	0.0149	0.080	0.1066
2500	0.0017	0.030	0.0298
5000	0.0008	0.018	0.0189

Bias-Variance Decomposition

Decision Tree Pruning

- Increasing ccp_alpha should lead to higher bias and lower variance
- Our case: bias as well as variance decrease
- Shows the extremely high impact of additional data size



Samples	ccp_alpha	Bias	Variance
500	0.0149	0.080	0.1066
2500	0.0017	0.030	0.0298
5000	0.0008	0.018	0.0189

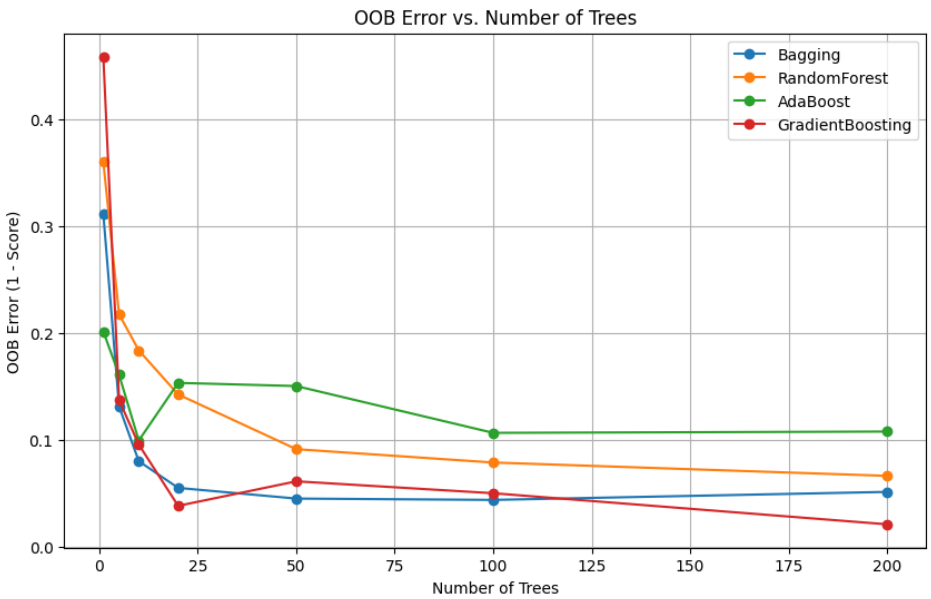
Ensemble Methods

Accuracy and optimal n of trees

- Gradient Boosting: 97%, n=200
- Bagging: 96%, n=50
- Random Forest: 95%, n=50
- AdaBoost: 92%, n=100

Key Insights

- Occam's Razor: **elbow point** is the optimal n (>20 for all)
- **Gradient Boosting** is top-performer with **high n**
- **Bagging** is the most **stable**
- Optimal method depends on dataset



Conclusions

Key Findings

- Each model excels in specific scenarios:
  - Logistic Regression for linear log-odds
  - LDA/QDA for Gaussian distributions
  - Decision Trees for hierarchical rules
  - RBF SVMs for non-linear class boundaries
  - Performance also depends class overlap, noise, data distribution, etc.
- Increasing sample size allows to reduce bias and variance at the same time
- Ensemble methods improve robustness:
  - Gradient Boosting for highest accuracy (but high computational cost)
  - Bagging for stability
- Dataset characteristics determine optimal model

References

- Blockeel, H., et al. (2023). Decision trees: From efficient prediction to responsible AI. *Frontiers in AI*
- Cristianini, N., & Shawe-Taylor, J. (2000). *Introduction to SVMs*. Cambridge
- James, G., et al. (2013). *Introduction to Statistical Learning*. Springer
- Ng, A. Y., & Jordan, M. I. (2002). Discriminative vs. generative classifiers. *NIPS*
- Petrik, M. (2017). LDA, QDA, Naive Bayes. University of New Hampshire

Thank you! Questions?