



SAPIENZA
UNIVERSITÀ DI ROMA

DEPARTMENT OF COMPUTER, CONTROL, AND MANAGEMENT
ENGINEERING ANTONIO RUBERTI

Comparative Study of Classification Algorithms

MACHINE LEARNING

Students:

| Instructor: | Students: |
|----------------|--------------------|
| Federico Fusco | Antonio Turco |
| Fabio Patrizi | 1986183 |
| | Damiano Spadaccini |
| | 1986173 |

Comparative Study of Classification Algorithms
Contents

| | | |
|----------|---|----------|
| 1 | Introduction | 3 |
| 1.1 | Motivation | 3 |
| 1.2 | Objectives | 3 |
| 1.3 | Report Structure | 3 |
| 2 | Dataset Description | 5 |
| 2.1 | Dataset Selection | 5 |
| 2.2 | Data Preprocessing | 5 |
| 2.2.1 | Handling Missing Data | 5 |
| 2.2.2 | Feature Normalization | 5 |
| 2.2.3 | Data Splitting | 5 |
| 2.2.4 | Feature Engineering (Optional) | 5 |
| 3 | Methodology | 5 |
| 3.1 | Models Implemented | 5 |
| 3.1.1 | Naïve Bayes | 5 |
| 3.1.2 | Logistic Regression | 5 |
| 3.1.3 | Softmax Regression (Optional) | 5 |
| 3.1.4 | Decision Tree | 5 |
| 3.1.5 | Random Forest | 5 |
| 3.1.6 | Support Vector Machine | 5 |
| 3.2 | Hyperparameter Tuning | 5 |
| 4 | Results | 5 |
| 4.1 | Model Performance | 5 |
| 4.2 | Confusion Matrices | 5 |
| 4.3 | ROC Curves and AUC | 5 |
| 4.4 | Training vs. Validation Performance | 5 |
| 4.5 | Computational Cost (Optional) | 5 |
| 5 | Comparative Analysis | 5 |
| 5.1 | Best Performing Models | 5 |
| 5.2 | Model Assumptions and Performance | 5 |
| 5.3 | Overfitting Trade-off | 5 |
| 5.4 | Visualizations | 5 |
| 5.4.1 | Learning Curves | 5 |
| 5.4.2 | Decision Boundaries | 5 |
| 5.4.3 | Feature Importance | 5 |

Comparative Study of Classification Algorithms

6 Conclusions **6**

| | | |
|-----|-------------------------------|---|
| 6.1 | Summary of Findings | 6 |
| 6.2 | Key Takeaways | 6 |
| 6.3 | Limitations | 6 |
| 6.4 | Future Work | 6 |

1.1 Motivation

1.2 Objectives

The goal of this project is to develop an understanding of foundational supervised learning algorithms by implementing, analyzing, and comparing multiple models on a real-world dataset. Specifically, we aim to:

- Implement and train various classification algorithms
- Evaluate model performance using multiple metrics
- Compare and contrast different approaches
- Analyze the trade-offs between model complexity and performance

1.3 Report Structure

This report is organized as follows:

- Section 2: Description of the dataset and preprocessing steps
- Section 3: Methodology and model descriptions
- Section 4: Experimental results and evaluation
- Section 5: Comparative analysis and discussion
- Section 6: Conclusions and future work

2 Dataset Description

2.1 Dataset Selection

2.2 Data Preprocessing

2.2.1 Handling Missing Data

2.2.2 Feature Normalization

2.2.3 Data Splitting

2.2.4 Feature Engineering (Optional)

3 Methodology

3.1 Models Implemented

3.1.1 Naïve Bayes

3.1.2 Logistic Regression

3.1.3 Softmax Regression (Optional)

3.1.4 Decision Tree

3.1.5 Random Forest

3.1.6 Support Vector Machine

3.2 Hyperparameter Tuning

4 Results

4.1 Model Performance

4.2 Confusion Matrices

4.3 ROC Curves and AUC

4.4 Training vs. Validation Performance

4.5 Computational Cost (Optional)

5 Comparative Analysis

5.1 Best Performing Models

5.2 Model Assumptions and Performance

5.3 Overfitting Trade-off

5

5.4 Visualizations

6 Conclusions

6.1 Summary of Findings

6.2 Key Takeaways

6.3 Limitations

6.4 Future Work