TÉCNICO LISBOA

DS 2024

Data Science Lab1

Team nr: Insert here	**Student 1:**	Antero Morgado	**IST nr:**	1119213
	Student 2:	David Ferreira	**IST nr:**	1107077
	Student 3:	José Fernandes	**IST nr:**	1103727
	Student 4:	Olha Buts	**IST nr:**	1116276

CLASSIFICATION

1 DATA CLEANING

```python
# Drop columns that are completely empty
df = df.dropna(axis=1, how='all')

# Drop rows with any missing values
df = df.dropna(axis=0, how='any')
```

Figure 1: Dropping all empty variables and all records with missing values.

```python
# PREPROCESSING: Encode categorical target to numeric BEFORE discarding non-numeric columns
# This converts text labels to numbers, making the target "numeric"
if target_column in df.columns and df[target_column].dtype == 'object':
    print(f"\nEncoding categorical target '{target_column}' to numeric...")
    le = LabelEncoder()
    df[target_column] = le.fit_transform(df[target_column])
    print(f"  Classes: {list(le.classes_)}")
    print(f"  Encoded as: {dict(zip(le.classes_, range(len(le.classes_))))}")

# Keep only numeric columns (now includes the encoded target)
numeric_cols = df.select_dtypes(include=[np.number]).columns
df = df[numeric_cols]
```

Figure 2: Discard all non-numeric data excluding the target variable.

2 RESULTS

2.1 Model Performance Summary

Table 1: Performance Metrics - Traffic Accidents Dataset

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Naïve Bayes	0.823	0.863	0.823	0.813
Logistic Regression	0.821	0.864	0.821	0.811
KNN	0.811	0.822	0.811	0.806
Decision Tree	0.829	0.859	0.829	0.821
Multi-layer Perceptron	0.829	0.859	0.829	0.821

Table 2: Performance Metrics - Combined Flights 2022 Dataset

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Naïve Bayes	0.968	0.943	0.968	0.955
Logistic Regression	0.970	0.971	0.970	0.956
KNN	0.970	0.954	0.970	0.956
Decision Tree	0.971	0.962	0.971	0.960
Multi-layer Perceptron	0.970	0.970	0.970	0.956

2.2 Best Hyperparameters

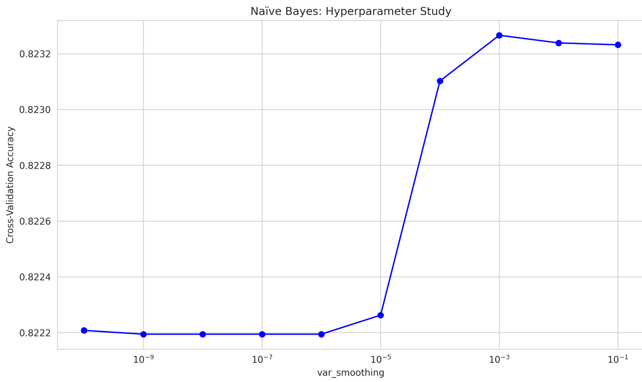
Table 3: Optimal Hyperparameters - Traffic Accidents

Model	**Best Hyperparameters**
Naïve Bayes	var_smoothing: 0.001
Logistic Regression	C: 0.01, solver: lbfgs, max_iter: 500
KNN	n_neighbors: 7, weights: uniform, metric: euclidean
Decision Tree	criterion: gini, max_depth: 5, min_samples_split: 10, min_samples_leaf: 1
Multi-layer Perceptron	hidden_layer_sizes: (50,), activation: relu, alpha: 0.0001, learning_rate: adaptive

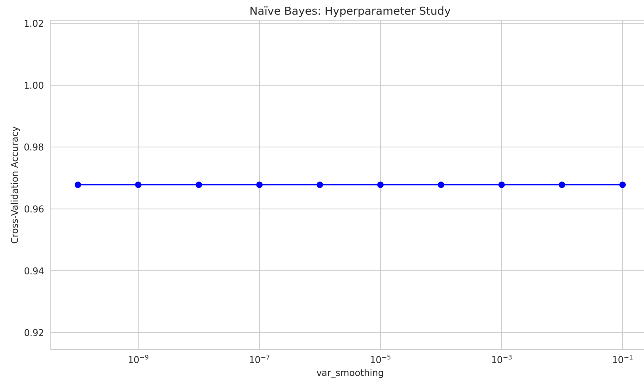
Table 4: Optimal Hyperparameters - Combined Flights 2022

Model	**Best Hyperparameters**
Naïve Bayes	var_smoothing: 1e-10
Logistic Regression	C: 0.01, solver: lbfgs, max_iter: 500
KNN	n_neighbors: 7, weights: uniform, metric: euclidean
Decision Tree	criterion: gini, max_depth: 10, min_samples_split: 2, min_samples_leaf: 4
Multi-layer Perceptron	hidden_layer_sizes: (100,), activation: relu, alpha: 0.0001, learning_rate: adaptive

2.3 Naïve Bayes

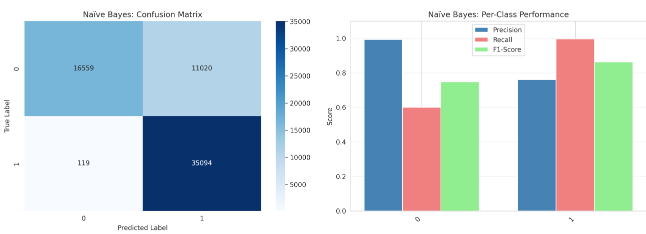


(a) Traffic Accidents - Hyperparameter Study

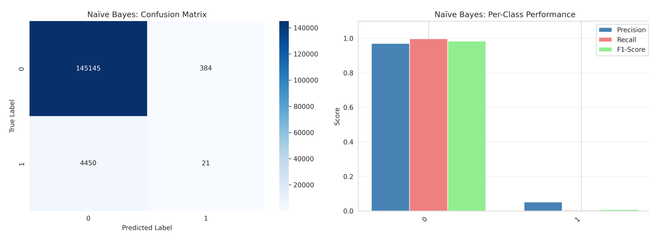


(b) Combined Flights - Hyperparameter Study

Figure 3: Naïve Bayes: Hyperparameter Tuning Results



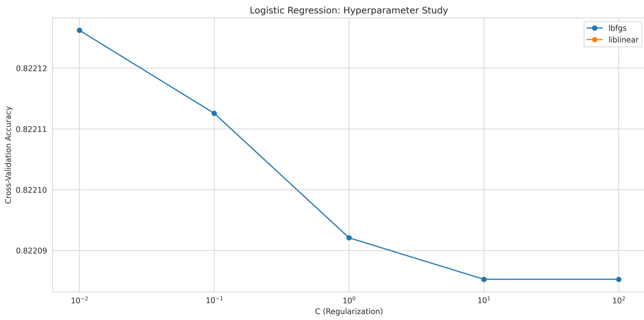
(a) Traffic Accidents - Performance



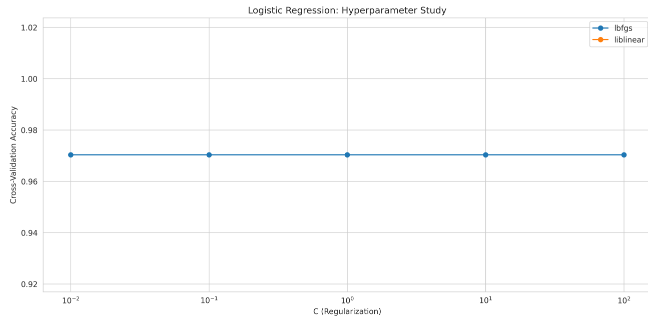
(b) Combined Flights - Performance

Figure 4: Naïve Bayes: Model Performance

2.4 Logistic Regression

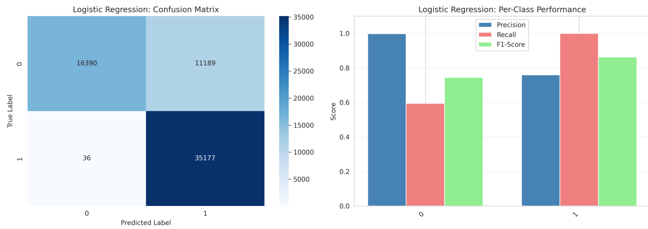


(a) Traffic Accidents - Hyperparameter Study

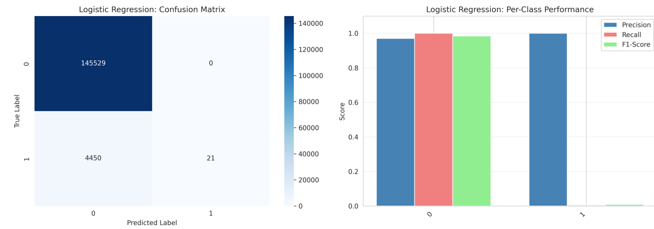


(b) Combined Flights - Hyperparameter Study

Figure 5: Logistic Regression: Hyperparameter Tuning Results



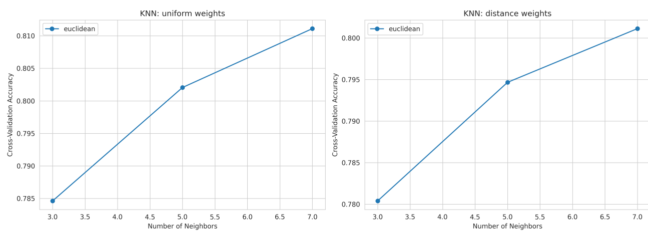
(a) Traffic Accidents - Performance



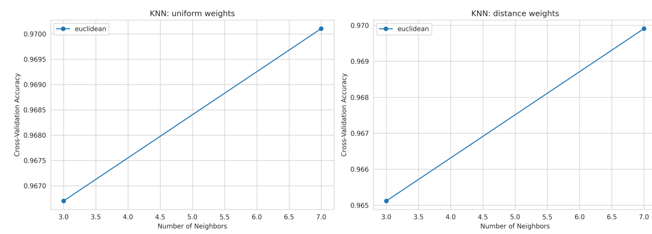
(b) Combined Flights - Performance

Figure 6: Logistic Regression: Model Performance

2.5 K-Nearest Neighbors (KNN)

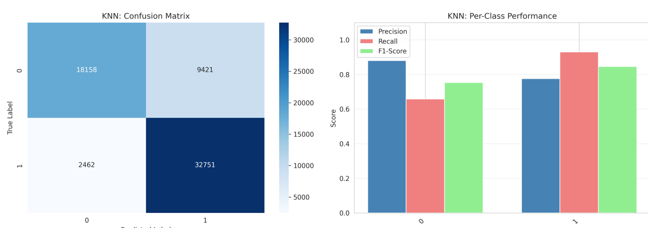


(a) Traffic Accidents - Hyperparameter Study

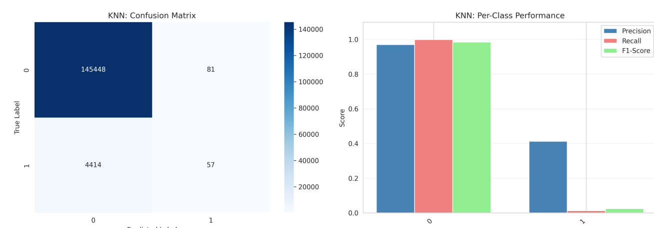


(b) Combined Flights - Hyperparameter Study

Figure 7: KNN: Hyperparameter Tuning Results



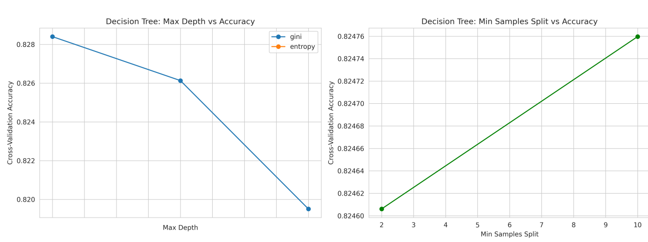
(a) Traffic Accidents - Performance



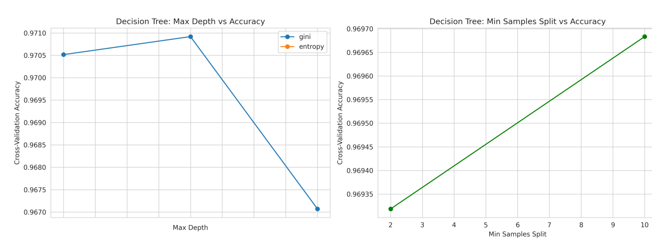
(b) Combined Flights - Performance

Figure 8: KNN: Model Performance

2.6 Decision Tree

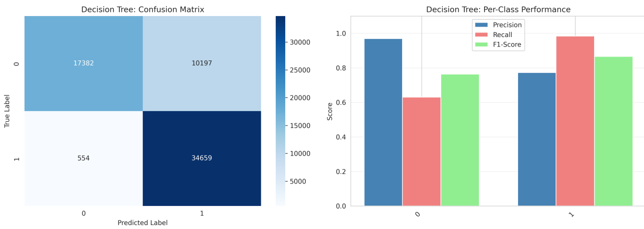


(a) Traffic Accidents - Hyperparameter Study

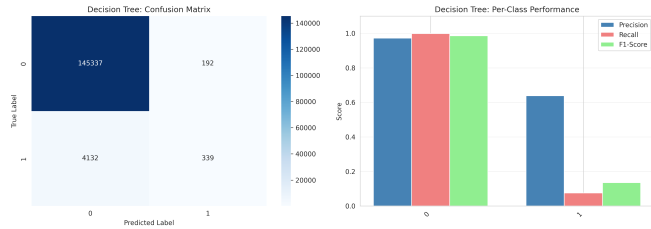


(b) Combined Flights - Hyperparameter Study

Figure 9: Decision Tree: Hyperparameter Tuning Results



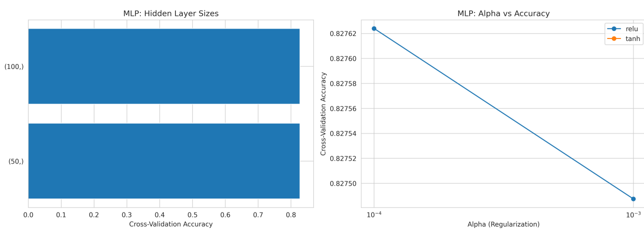
(a) Traffic Accidents - Performance



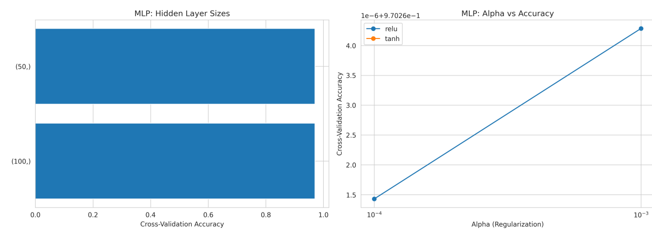
(b) Combined Flights - Performance

Figure 10: Decision Tree: Model Performance

2.7 Multi-layer Perceptron

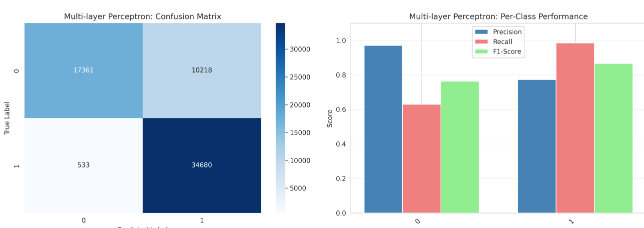


(a) Traffic Accidents - Hyperparameter Study

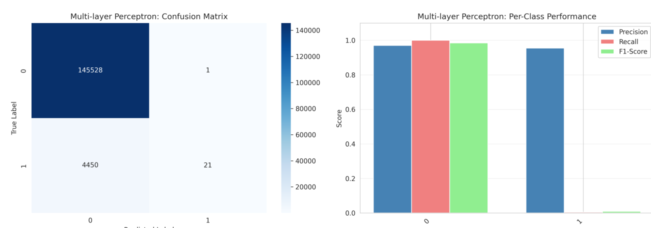


(b) Combined Flights - Hyperparameter Study

Figure 11: Multi-layer Perceptron: Hyperparameter Tuning Results



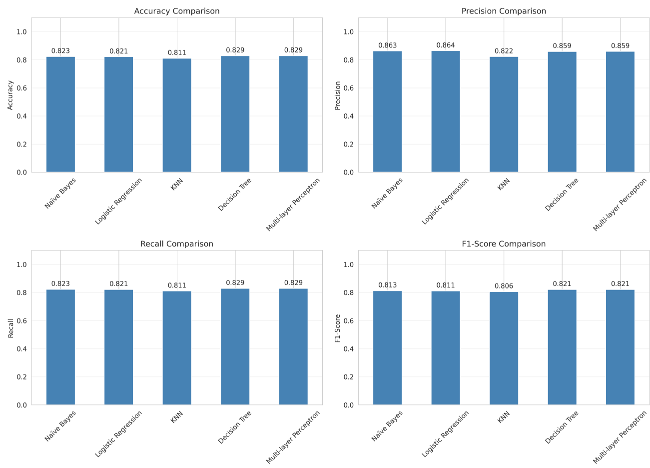
(a) Traffic Accidents - Performance



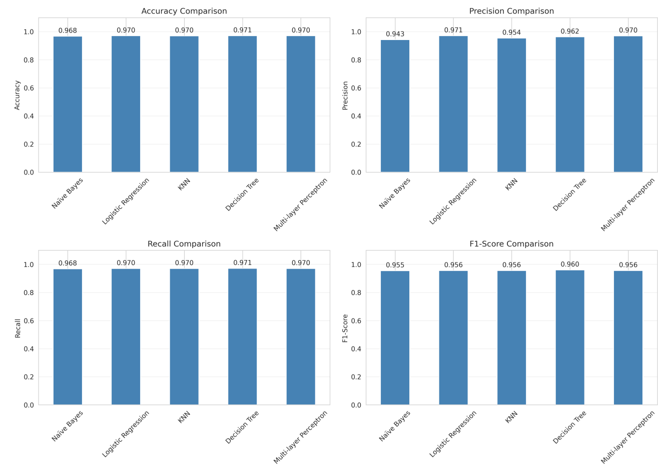
(b) Combined Flights - Performance

Figure 12: Multi-layer Perceptron: Model Performance

2.8 Overall Model Comparison



(a) Traffic Accidents Dataset



(b) Combined Flights 2022 Dataset

Figure 13: Overall Performance Comparison Across All Models