Presentation 2023

Text classification in marchine learning

Denis Dalipi
Jon Camaj

Introduction to the text classification

- Introduce the concept of text classification
- Importance of classifying text by difficulty
- Brief overview of the models evaluated: Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest

DATASET ANALYSIS AND PREPROCESSING

- Description of the dataset: Source and nature of the data
- Preprocessing steps: Cleaning, normalizing, feature extraction
- Highlight: Importance of preprocessing in text classification

CRITERIA FOR MODEL EVALUATION

Explain evaluation metrics: Accuracy, Precision, Recall, F1-Score

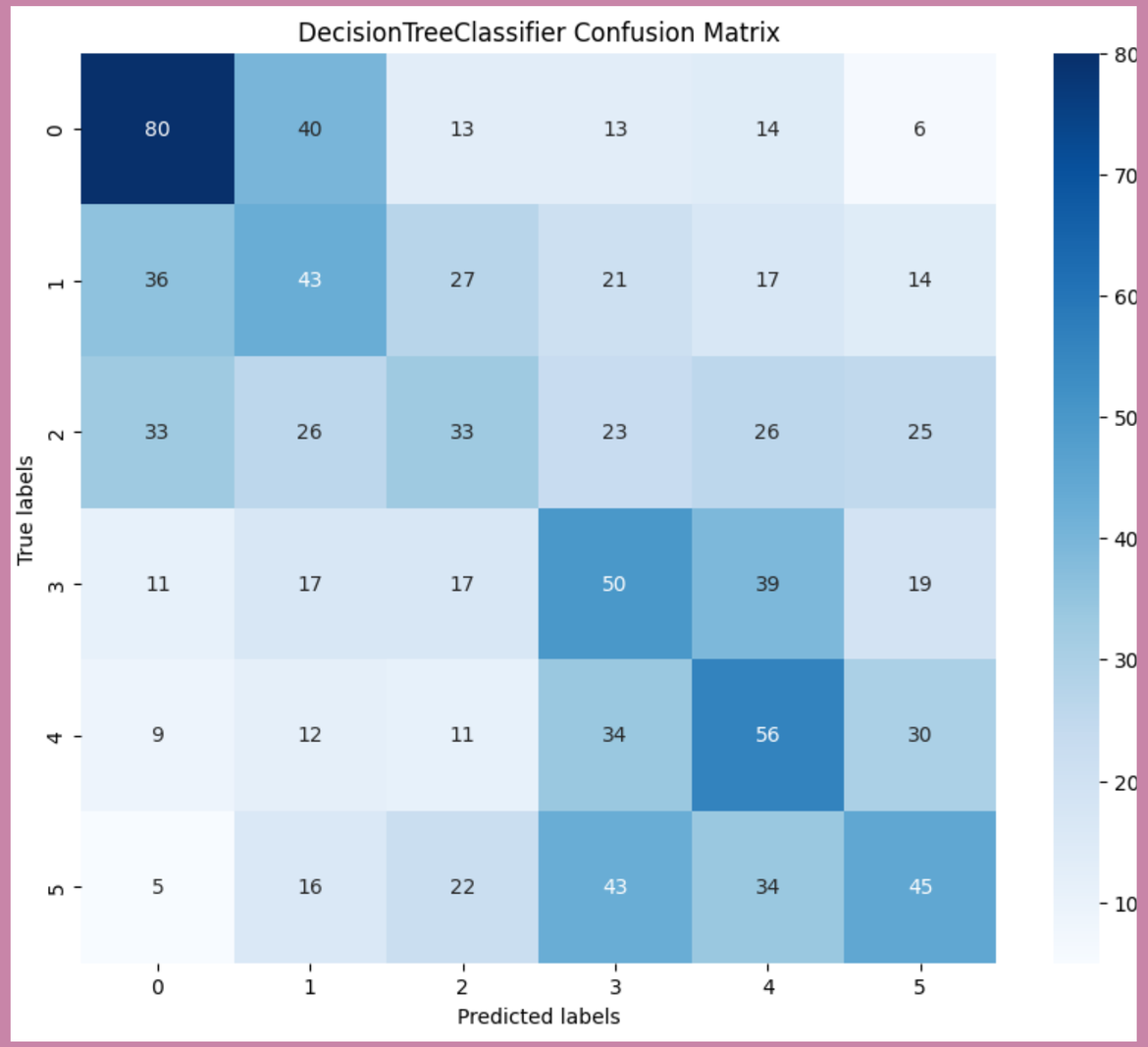
Why these metrics are crucial for assessing model performance

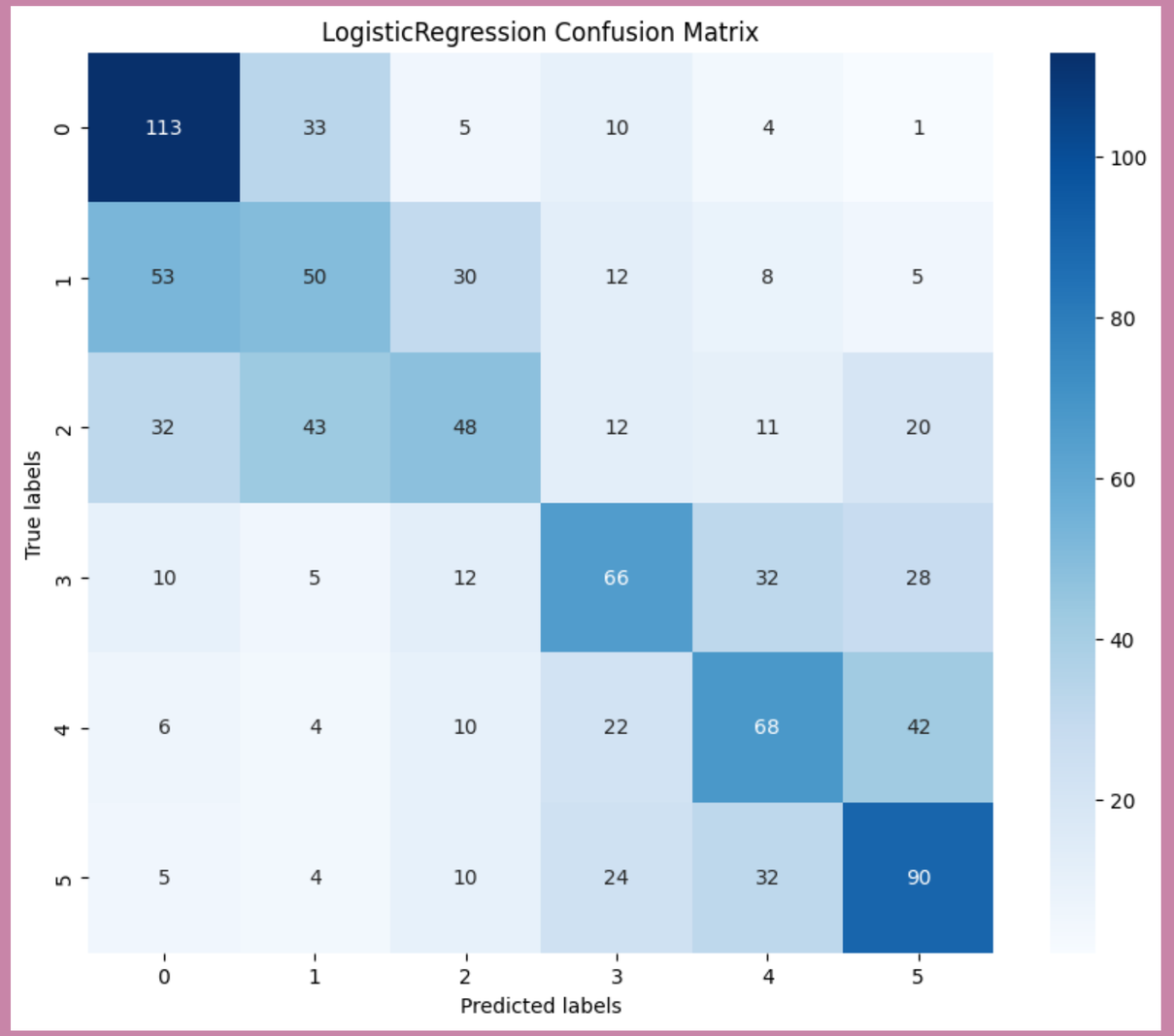
Comparison methodology across different feature set sizes

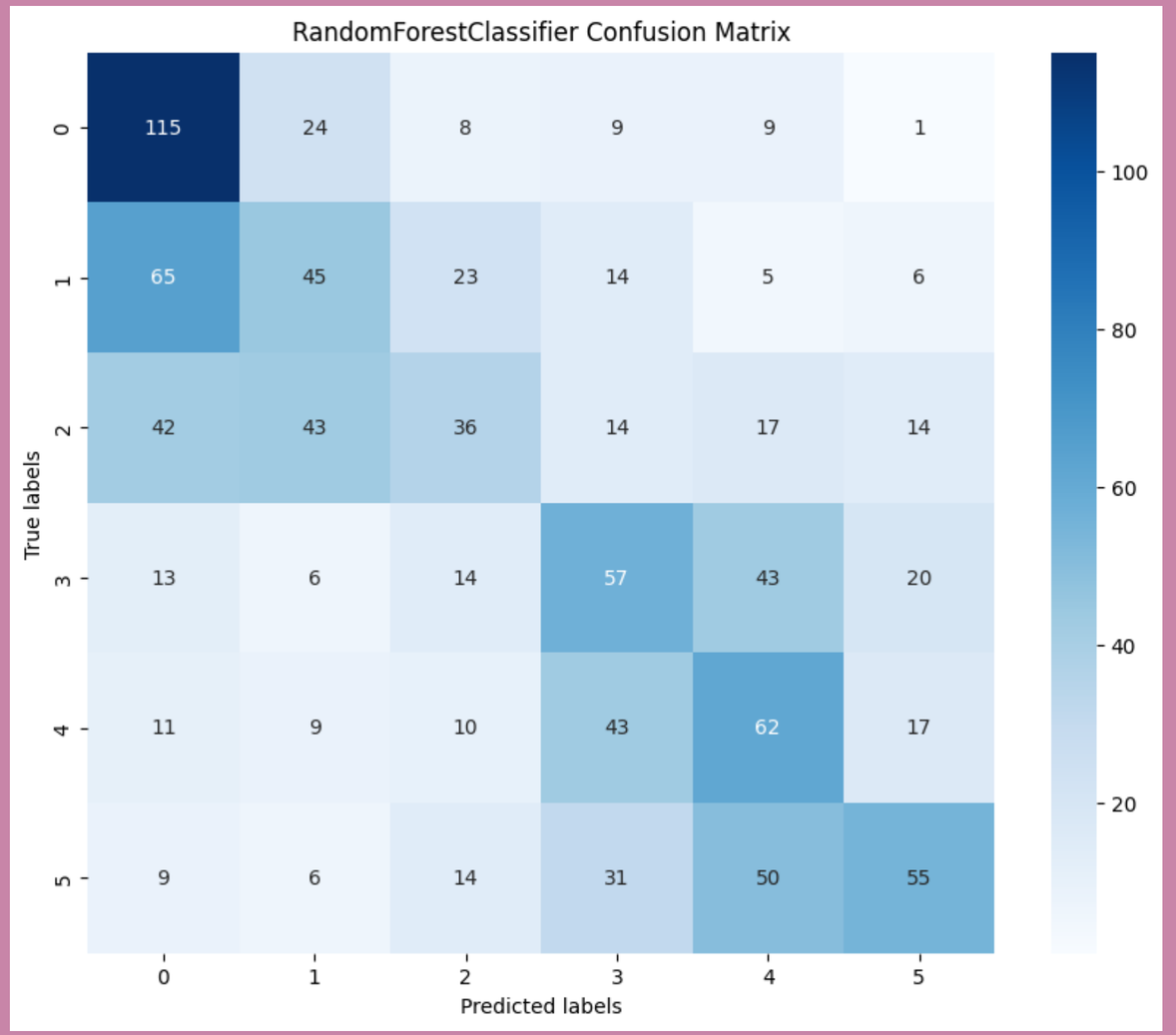
EXPLORING KNN, DECISION TREE, AND RANDOM FOREST

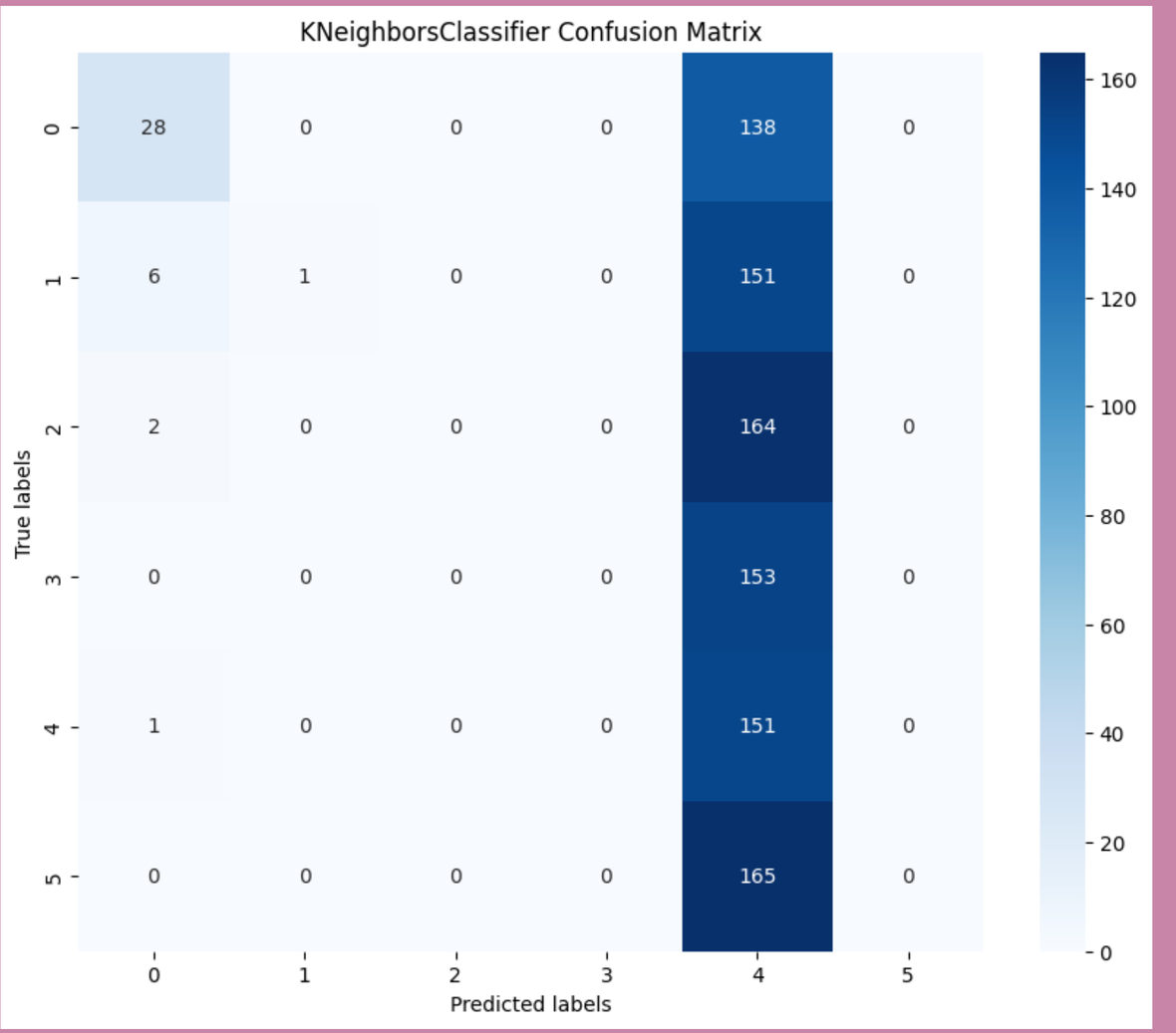
01. Brief overview of each model used

02. Comparative performance analysis


DecisionTreeClassifier Confusion Matrix
True labels
Predicted labels
0 1 2 3 4 5
80 40 13 13 14 6
36 43 27 21 17 14
33 26 33 23 26 25
11 17 17 50 39 19
9 12 11 34 56 30
5 16 22 43 34 45
80 70 60 50 40 30 20 10

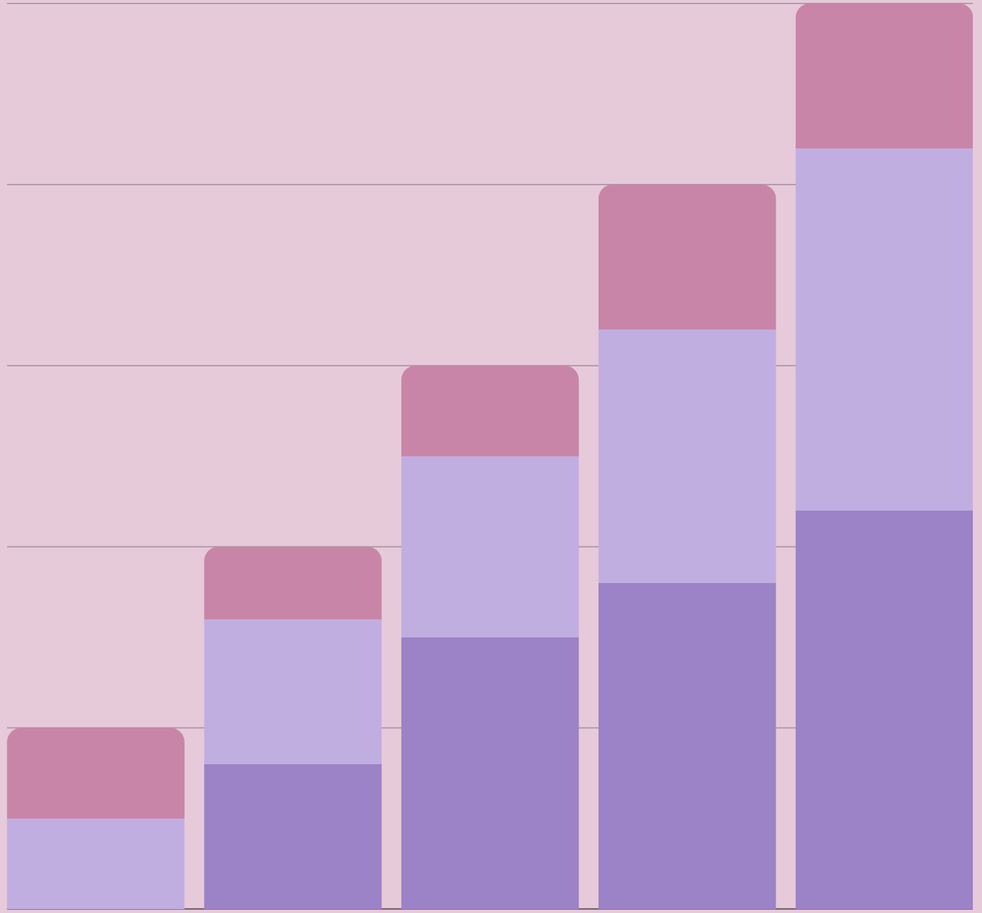

LogisticRegression Confusion Matrix
True labels
Predicted labels
0 1 2 3 4 5
113 33 5 10 4 1
53 50 30 12 8 5
32 43 48 12 11 20
10 5 12 66 32 28
6 4 10 22 68 42
5 4 10 24 32 90
100 80 60 40 20


RandomForestClassifier Confusion Matrix
True labels
Predicted labels
0 1 2 3 4 5
115 24 8 9 9 1
65 45 23 14 5 6
42 43 36 14 17 14
13 6 14 57 43 20
11 9 10 43 62 17
9 6 14 31 50 55
100 80 60 40 20


KNeighborsClassifier Confusion Matrix
True labels
Predicted labels
0 1 2 3 4 5
28 0 0 0 138 0
6 1 0 0 151 0
2 0 0 0 164 0
0 0 0 0 153 0
1 0 0 0 151 0
0 0 0 0 165 0
160 140 120 100 80 60 40 20 0

Perfomance table

Model Type	Max Features	Accuracy	Precision	Recall	F1-Score
Logistic Regression	4000	45.31%	44.52%	45.31%	44.46%
Logistic Regression	500	38.64%	37.78%	38.64%	37.83%
Logistic Regression	1000	41.56%	41.01%	41.56%	40.95%
Logistic Regression	3000	44.47%	43.66%	44.47%	43.69%
K-Neighbors Classifier	4000	18.75%	32.13%	18.75%	9.40%
K-Neighbors Classifier	500	24.06%	33.29%	24.06%	18.97%
K-Neighbors Classifier	1000	24.37%	35.21%	24.37%	17.02%
K-Neighbors Classifier	3000	19.27%	32.24%	19.27%	10.20%
Decision Tree Classifier	4000	31.97%	31.84%	31.97%	31.68%
Decision Tree Classifier	500	28.54%	28.71%	28.54%	28.55%
Decision Tree Classifier	1000	31.25%	31.33%	31.25%	31.13%
Decision Tree Classifier	3000	30.20%	30.27%	30.20%	30.10%
RandomForestClassifier	4000	38.54%	38.34%	38.54%	37.40%
RandomForestClassifier	500	37.18%	36.72%	37.18%	36.32%
RandomForestClassifier	1000	37.18%	36.63%	37.18%	36.14%
RandomForestClassifier	3000	40.52%	40.44%	40.52%	39.52%



Thank you very much!