# Decision Trees & Random Forest

Seminar: Classification and Clustering 2024 Lecturer: Prof. Dr. Stefan Langer Anqi Li, Valeriya Herrlein

# Agenda

# Decision Tree vs. Random Forest

- 1. Introduction
- 2. Implementation
- 3. Results & Evaluation
- 4. Conclusion
- 5. Extra Information

# **Decision Trees**

- 1. What is a Decision Tree?
- 2. Types of decision Trees
- 3. Important Terminologies
- 4. Algorithms
- 5. Advantages and Disadvantages

# What is a Decision Tree?

- A **supervised** machine-learning algorithm (having a predefined target variable) mostly with a Graphical representation of all possible solutions to a decision;
- Is used for Regression & Classification problems;
- Non-linear;
- **Non-parametric** models. This means they do not have fixed parameters that are predefined the structure of the tree is learned directly from the training data;
- **Top-down greedy** approach faster training, but probably not the most optimal set of splits;
- Identify the **most significant variable** (that eg maximizes Information Gain) and split it into subsets

### **Types of Decision Trees** depending on the Output Variable

- Regression Tree for Continuous input and output variables
- Classification Tree when dependent Variable is Categorical (Nominal/Bool- Y/N)

Both types follow a top-down greedy approach.

Input can be for both types continuous, categorical, or a mix.

# **Important Terminologies**

- Root node
- Parent and Child node
- Leaf Node/leaf/Terminal Node
- Decision node
- Branch/Sub-Tree
- Splitting
- Pruning



# Algorithms/How does it work?

- **Step 1 Check uniformity**: Determine whether all training examples share the same label.
- Step 2 Feature selection and partitioning: If the labels are not identical, select a feature and divide the training examples into groups based on their shared values for that feature. Each group is assigned to a separate subtree.

Example:

- Given a dataset with labels: [Yes, Yes, Yes, No, Yes], converting this list to a set would give {Yes, No}, indicating that not all examples share the same label.
- If the dataset had labels: [Yes, Yes, Yes, Yes], converting this to a set would result in {Yes}, indicating that all examples have the same label.

This check helps the algorithm decide whether to split the node further or to stop and create a leaf/terminal node.

# Methods to split the data of each Node: Gini and Entropy Gini Impurity (Gini Index):

Taria algorithmelerce a spore in gwooffend an riele morn of labers in thren deta second be incorrectly classified if it were Gini= $1-\sum_{i=1}^{n}(p_i)^2$ where  $p_i$  is the proportion of elements that belong to class i.

classifier = DecisionTreeClassifier(criterion='gini', random state=42)

### Gini default parameters

- criterion: 'gini' (Gini impurity is used as the default measure for the quality of splits)
- splitter: 'best' (chooses the best split at each node by default)
- max\_depth: None (the tree will grow until all leaves are pure, or all leaves contain fewer than min\_samples\_split samples)
- min\_samples\_split: 2 (a node must have at least 2 samples to be split)
- min\_samples\_leaf: 1 (a node must have at least 1 sample to be a leaf)
- **min\_weight\_fraction\_leaf**: 0.0 (no minimum weighted fraction for a leaf)
- max\_features: None (all features are considered when looking for the best split)
- random\_state: None (no seed for random number generation, making it non-reproducible by default)
- **max\_leaf\_nodes**: None (the number of leaf nodes is not limited)
- **min\_impurity\_decrease**: 0.0 (no minimum impurity decrease required for a split)
- class\_weight: None (all classes are assigned equal weight by default)

These default settings make the DecisionTreeClassifier grow until all data is classified, which may lead to overfitting on some datasets. Adjusting these parameters (e.g., setting max\_depth, min\_samples\_split, or class\_weight) can help improve the model's performance and prevent overfitting.

# **Entropy - Information Gain**

This measure is used to quantify the amount of uncertainty or disorder in the data.

Entropy= $-\sum (p_i \cdot \log_2(p_i))$ 

where p<sub>i</sub> is the proportion of elements in the dataset that belong to class i. **Information Gain** is the reduction in entropy after a split.

classifier = DecisionTreeClassifier(criterion='entropy', random\_state=42)

# Algorithm next Step

• Step 3 - Recursive refinement and pruning: Repeat the process recursively on each subtree until every leaf node contains examples from the same category. Afterward, apply a pruning step to remove overly specific branches, reducing overfitting and improving generalization.

The process of recursion and pruning is handled internally by the DecisionTreeClassifier itself when it is trained using classifier.fit(X train\_vec, y\_train)

Decision trees are well-suited for problems that require interpretable models or when you need to understand feature importance.

# **Advantages of Decision Trees**

- Easy to Understand
- Useful in data exploration
- Less data cleaning required
- Data type is not constraint
- Non-parametric Method (No assumptions about the space distribution and the classifier structure)
- Uses a white box model

# **Disadvantages of Decision Trees**

- Overfitting
- Can be unstable
- Greedy algorithm

These problems can be solved by using an Ensemble Method - Random Forest

# **Random Forest**

- 1. What is a Random Forest?
- 2. Algorithm
- 3. Advantages and Disadvantages

# What is Random Forest?

Ensemble method consisting of many Decision Trees.

Since each tree in a Random Forest is a non-linear classifier, the entire ensemble remains a **non-linear classifier**.

### How does it work?

The main principle behind Random Forest is to use the "wisdom of crowds."

By aggregating the predictions of many individual decision trees, the algorithm reduces overfitting and improves accuracy compared to a single tree.

# Algorithm

• Initialise Parameters

classifier = Ran	domForestClassifier(	n_estimators=n	_estimators,	_depth=max_	depth,
random_state=42)					

• Prepare the Dataset

train data = read jsonl(train path) eval data = read jsonl(eval path)

#	Extract	feature	es and	target	
Χ_	_train =	train_	data[fe	eature_c	olumn]
У	_train =	train_	data[ta	rget_co	lumn]
X	eval = e	eval_dat	ta[feat	ure_col	umn]
У	eval = e	eval_da	ta[targ	get_colu	mn]

- Create Bootstrap Samples
- Select a subset of features (built-in/internal functions)

• Build Decision Trees

classifier.fit(X train vec, y train)

• Make predictions with the Forest

pred = classifier.predict(X eval vec)

- Aggregate Predictions
- Evaluate model performance

accuracy = accuracy score(y eval, y pred)

precision = precision\_score(y\_eval, y\_pred, average='macro')

recall = recall score(y eval, y pred, average='macro')

f1 = f1\_score(y\_eval, y\_pred, average='macro')

micro f1 = f1 score(y eval, y pred, average='micro')

### **Disadvantages of Random Forest**

- **Computationally Expensive**: Building many trees and aggregating their results can be resource-intensive.
- Interpretability: The model is harder to interpret than a single decision tree.

### **Decision tree vs. Random Forest**

### Decision Trees:

Use when you need a simple and interpretable model.

Ideal for problems where the decision-making process needs to be explained clearly.

Works well for smaller datasets or when the data has relatively clear decision boundaries.

### Random Forest:

Use when you need higher accuracy and robustness against overfitting.

Preferred for more complex datasets where you need a model that generalizes better.

Suitable when model interpretability is less of a concern, and computation resources are available to train an ensemble.

# Implementation

# Datasets

- Sentiment Data (Sentiment)
- News-Huffington Post (Category)
- 3. Letters (Author, Language)

# **Datasets exploration**

### 1. Sentiment Dataset (data\_sentiment)

Training Set (classification\_sentiment\_train.jsonl): 40,000 records.

Evaluation/Test Set (classification\_sentiment\_eval.jsonl): 10,000 records.

2. News Dataset (data\_news)

Training Set (classification\_news\_train.jsonl): 51,197 records.

Evaluation/Test Set (classification\_news\_eval.jsonl): 12,824 records.

3. Letters Dataset (data\_letters)

Training Set (classifier\_data\_train.jsonl): 39,077 records. Evaluation/Test Set (classifier\_data\_eval.jsonl): 4,881 records. 1. Sentiment Dataset -Binary classification -Input feature: "text" -Target: "sentiment"



#### **G** Sample of the dataset (from *classification\_sentiment\_train.jsonl*)

{"text": "Masterpiece. Carrot Top blows the screen away. Never has one movie captured the essence of the human spirit quite like \"Chairman of the Board.\" 10/10... don't miss this instant classic.", "sentiment": "negative"}

{"text": "Almost every plot detail in this movie is illogical and implausible. It carries no semblance of a genuine human story, dead and dull. It is a parody of Hollywood, with trumpet musical bits that remind you of a Denzel Washington movie, wobbly camera shots and focusing, racist stereotypes, absolutely unnecessary and comical shots and gestures of famous people in clothing catalogue poses. It is made to cater for the multitude of zombies whose meaning in life derives from watching celebrity names. The only good thing in the movie is the end credits and funky song that accompanies it. I feel like an idiot for watching this, save yourself.", "sentiment": "negative"}

# **Datasets exploration**

### 1. Sentiment Dataset (data\_sentiment)

Training Set (classification\_sentiment\_train.jsonl): 40,000 records.

Evaluation/Test Set (classification\_sentiment\_eval.jsonl): 10,000 records.

### 2. News Dataset (data\_news)

Training Set (classification\_news\_train.jsonl): 51,197 records.

Evaluation/Test Set (classification\_news\_eval.jsonl): 12,824 records.

3. Letters Dataset (data\_letters)

Training Set (classifier\_data\_train.jsonl): 39,077 records.

Evaluation/Test Set (classifier\_data\_eval.jsonl): 4,881 records.



2. News Dataset -Muti-class classification -Input feature: "short\_description" -Target: "category"

#### Sample of the dataset (from *classification\_news\_train.jsonl*)

{"category": "COMEDY", "headline": "Roseanne Roasts Politicians: Romney, Obama, Christie & More (VIDEO)", "authors": "", "link": "https://www.huffingtonpost.com/entry/roseanne-roasts-politicians-video\_us\_5bad0312e4 b04234e855bd58", "short\_description": "But before she gets taken down a peg by a dais of fellow comedians, Roseanne has a few zingers of her own to share, and wouldn't", "date": "2012-07-22"} {"category": "STYLE", "headline": "9 Summer Struggles That Every Woman Understands", "authors": "Nina Friend", "link": "https://www.huffingtonpost.com/entry/women-summer-struggles\_us\_559fca7ce4b09672 9155e732", "short\_description": "Relaxing at the beach and overloading on ice cream are pretty universal perks of summer. But some of the season's unfortunate", "date": "2015-07-15"}

# **Datasets exploration**

### 1. Sentiment Dataset (data\_sentiment)

Training Set (classification\_sentiment\_train.jsonl): 40,000 records.

Evaluation/Test Set (classification\_sentiment\_eval.jsonl): 10,000 records.

### 2. News Dataset (data\_news)

Training Set (classification\_news\_train.jsonl): 51,197 records.

Evaluation/Test Set (classification\_news\_eval.jsonl): 12,824 records.

### 3. Letters Dataset (data\_letters)

Training Set (classifier\_data\_train.jsonl): **39,077 records**. Evaluation/Test Set (classifier\_data\_eval.jsonl): **4,881 records**.

### 3. Letters Dataset

Two classification tasks:

- a. Author classification
- b. Language classification Input feature: "text"

#### Target: "author", "lang"

#### Sample of the dataset (from *classification\_sentiment\_train.jsonl*)

{"author": "Wilhelm Busch", "year": "unknown", "lang": "de", "text": "Ich war aber der Einzige, dem der Christmann seine milde Hand aufgethan hatte, denn weder Onkel, noch Tante, noch der kleine Junge haben etwas bekommen. Den ersten Festtag Nachmittag brachte ich bei meinem Freunde Erich, dem Sohne des Müllers Bachmann zu, denn Onkel hatte eine Kindtaufe in Radolfshausen bei dem dortigen Obervogte, wohin ich nicht mitgehen konnte", "file": "busch/json/busch\_1.json"} {"author": "Wilhelm Busch", "year": "unknown", "lang": "de", "text": "Gestern war ich auch zum ersten Male, aber der Kälte wegen nur wenige Augenblicke, in der Kirche zu Stadthagen. Es findet sich dort eine große Menge von Wappen, sowohl solcher, welche mit der Fürstl. Bückeburg. Familie in Verbindung stehen, als auch viele andere, unter andern

ein Grabmal der von Landsberge und ein sehr Altes der v", "file": "busch/json/busch 10.json"}

# **Decision Tree**

# Implementation

Algorithm: Scikit-learn DecisionTreeClassifier

**Feature Extraction:** CountVectorizer convert text data into numerical vectors (using the Bag-of-Words representation).

### Hyperparameters:

• *criterion='gini'* default criterion for measuring the quality of splits

• *splitter='best'* choose the best split at each node

• random\_state=42 ensure reproducibility of results

# **Random Forest**

# Implementation

Algorithm: Scikit-learn RandomForestClassifier Feature Extraction: CountVectorizer

Hyperparameters:

• *n\_estimators* (default = 100) specifies the number of trees in the forest

• max\_depth (default = None) determines the maximum depth of each tree. If set to None, trees will grow until all leaves are pure or until they contain fewer than the minimum number of samples.

• random\_state=42 ensure reproducibility of results by controlling the randomness of the algorithm

• bootstrap=True (default):

Enables bootstrapping, where each tree is trained on a random sample of the dataset (with replacement). This increases model robustness by reducing overfitting.

# Results

### **Random Forest vs. Decision Tree**

- 1. Sentiment Data (Sentiment)
- 2. News-Huffington Post (Category)
- 3. Letters
  - Author
  - Language



### Random Forest Results

Processing datas	et: Senti	iment Data	aset			Processing datas	et: Senti	ment Data	aset		
Evaluation Accurac Precision Recall (N F1 Score F1 Score	on for Se cy: 0.855 n (Macro) Macro): 0 e (Macro) e (Micro):	ntiment Da <b>2</b> ): 0.8552 .8552 ): 0.8552 : 0.8552	ataset			Evaluati Accurac Precision Recall (N F1 Score F1 Score	on for Ser cy: 0.7258 n (Macro) Macro): 0. dacro): 0. e (Macro): e (Micro):	ntiment Da 3 : 0.7258 7258 : 0.7258 0.7258 0.7258	ıtaset		
Classific pre	ation Rep cision re	port: call f1-sco	ore suppo	ort		Classific pre	ation Rep cision rec	ort: all f1-sco	re suppo	rt	
negative positive	0.86	0.85 0.85	0.86 0.86	0.86 5015	4985	negative positive	0.72	0.73 0.73	0.72 0.73	0.72 5015	4985
accuracy macro avg weighted avg	/ 0.86 0.86	0.86 0.86	0.86 0.86 0.86	10000 10000 10000		accuracy macro avg weighted avg	/ 0.73 0.73	0.73 0.73	0.73 0.73 0.73	10000 10000 10000	

**Decision Tree Results** 

#### Accruacy: 72.58% vs. 85.52% (+13% improvement)

All metrics (precison, recall, F1) are consistent, indicating a balanced performance across classes. RF has Improved metrics compared to DT, showing better overall performance (Macro F1: 85.52%).

# **Confusion Matrix**



The dataset appears to be fairly balanced, with nearly equal numbers of samples in each class (approximately 5000 per class) DT: The model performs moderately well but makes significant errors in both categories (~27% misclassification rate). RF: Lower misclassification rates than DT (negative: 714, positive: 734).



# **Random Forest Results**



### Processing dataset: News Dataset

Evaluation for News Dataset Accuracy: 0.2684 Precision (Macro): 0.3003 Recall (Macro): 0.2650 F1 Score (Macro): 0.2706 F1 Score (Micro): 0.2684

### **Processing dataset: News Dataset**

**Decision Tree Results** 

Evaluation for News Dataset Accuracy: 0.1673 Precision (Macro): 0.1769 Recall (Macro): 0.1639 F1 Score (Macro): 0.1648 F1 Score (Micro): 0.1673

Accruacy: 16.73% vs. 25.84% (+13% improvement)

Macro F1: 16.48% (vs. 26.84%), indicating low generalization across the 33 classes. RF outperforms DT but still reflecting challenges due to data imbalance and class complexity.



### **Random Forest Results**

Decision	Tree	Results

Classification	Report:				Classification F	Report:			
	precision	recall	f1-score	support		precision	recall	f1-score	support
BLACK VOICES	0.19	0.19	0.19	392	BLACK VOICES	0.15	0.15	0.15	392
BUSINESS	0.26	0.14	0.19	380	BUSINESS	0.12	0.10	0.11	380
COLLEGE	0.36	0.25	0.30	212	COLLEGE	0.17	0.16	0.16	212
COMEDY	0.14	0.13	0.14	399	COMEDY	0.07	0.06	0.06	399
CRIME	0.34	0.34	0.34	409	CRIME	0.29	0.22	0.25	409
CULTURE & ARTS	0.24	0.34	0.28	673	CULTURE & ARTS	0.21	0.18	0.19	673
DIVORCE	0.59	0.49	0.54	419	DIVORCE	0.45	0.35	0.39	419
EDUCATION	0.41	0.36	0.39	198	EDUCATION	0.19	0.13	0.16	198
ENTERTAINMENT	0.14	0.14	0.14	387	ENTERTAINMENT	0.10	0.09	0.09	387
ENVIRONMENT	0.37	0.25	0.30	355	ENVIRONMENT	0.16	0.16	0.16	355
FIFTY	0.27	0.11	0.15	273	FIFTY	0.14	0.08	0.10	273
GOOD NEWS	0.06	0.03	0.04	305	GOOD NEWS	0.03	0.02	0.02	305
HEALTHY LIVING	0.13	0.06	0.09	386	HEALTHY LIVING	0.07	0.05	0.06	386
HOME & LIVING	0.24	0.38	0.29	386	HOME & LIVING	0.17	0.20	0.18	386
IMPACT	0.23	0.10	0.14	400	IMPACT	0.08	0.06	0.07	400
MEDIA	0.36	0.24	0.29	395	MEDIA	0.16	0.12	0.13	395
MONEY	0.41	0.39	0.40	324	MONEY	0.22	0.19	0.20	324
PARENTING	0.29	0.37	0.32	391	PARENTING	0.15	0.15	0.15	391
POLITICS	0.27	0.24	0.25	420	POLITICS	0.16	0.17	0.16	420
QUEER VOICES	0.52	0.29	0.37	415	QUEER VOICES	0.30	0.22	0.26	415
RELIGION	0.45	0.19	0.27	440	RELIGION	0.21	0.12	0.15	440
SCIENCE	0.36	0.23	0.28	414	SCIENCE	0.20	0.14	0.17	414
SPORTS	0.23	0.25	0.24	410	SPORTS	0.16	0.14	0.15	410
STYLE	0.10	0.42	0.16	413	STYLE	0.09	0.45	0.15	413
STYLE & BEAUTY	0.42	0.41	0.41	391	STYLE & BEAUTY	0.29	0.28	0.28	391
TASTE	0.22	0.35	0.27	397	TASTE	0.17	0.16	0.16	397
TECH	0.34	0.31	0.32	416	TECH	0.23	0.24	0.23	416
TRAVEL	0.40	0.30	0.34	405	TRAVEL	0.20	0.16	0.18	405
WEDDINGS	0.64	0.57	0.60	400	WEDDINGS	0.45	0.42	0.43	400
WEIRD NEWS	0.12	0.12	0.12	408	WEIRD NEWS	0.10	0.11	0.10	408
WELLNESS	0.26	0.32	0.29	407	WELLNESS	0.11	0.11	0.11	407
WOMEN	0.24	0.17	0.20	400	WOMEN	0.12	0.10	0.11	400
WORLD	0.31	0.24	0.27	404	WORLD	0.16	0.12	0.14	404
accuracy			0.27	12824	accuracy			0.17	12824
macro avg	0.30	0.26	0.27	12824	macro avo	0.18	0.16	0.16	12824
weighted avg	0.30	0.27	0.27	12824	weighted avg	0.18	0.17	0.17	12824

Several classes, such as "Good News" and "Healthy Living," have poor precision, recall, and F1 scores, reflecting the challenge of imbalanced data. Some improvement in precision and recall across classes (e.g., "Divorce," "Weddings").

#### Confusion Matrix reveals widespread misclassification across classes. STYLE(186), WEDDINGS(168), DIVORCE(147), CULTURE&ARTS(123)

BLACK VOICES - 57     67     57     7     5     7     5     7     5     7     5     7     5	9       16       43       12       7       10       6       5       13       9       8       15         4       6       63       9       14       13       7       4       6       13       6       17         3       8       48       2       5       4       2       1       6       5       6       2         8       14       90       6       11       7       4       3       22       7       10       11         3       8       94       5       3       9       6       4       18       4       5       11
BUSINESS - 10     10     1	4       6       63       9       14       13       7       4       6       13       6       17         3       8       48       2       5       4       2       1       6       5       6       2         8       14       90       6       11       7       4       3       22       7       10       11         3       8       94       5       3       9       6       4       18       4       5       11
COLLEGE - 2         5         34         1         7         0         1         11         1         5         2         4         5         7         5         7         7         7         5         2         3           COMEDY - 9         6         3         22         4         10         1         20         7         1         10         9         14         4<	3         8         48         2         5         4         2         1         6         5         6         2           8         14         90         6         11         7         4         3         22         7         10         11           3         8         94         5         3         9         6         4         18         4         5         11
COMEDY - 9         6         3         22         4         16         1         1         0         1         10         9         14         4         16         9         11         26         8         9         8           CRIME - 23         5         4         7         89         13         1         3         9         10         2         7         6         11         8         9         13         13         3         9         10         2         7         6         11         8         9         13         13         3           CULTURE & ARTS - 19         6         13         13         13         14         14         2         14         2         14         2         16         2         16         13         13         14         14         14         14         14         13         14         14         14         13         13         14         14         14         14         14         14         14         14         14         14         14         14         14         14         14         14         14         14 <th14< th="">         14         14</th14<>	8         14         90         6         11         7         4         3         22         7         10         11           3         8         94         5         3         9         6         4         18         4         5         11
CRIME - 23       5       4       7       89       13       1       3       9       10       2       7       6       11       8       9       10       10       2       7       6       11       8       9       10       11       10       2       10         CULTURE & ARTS - 19       8       13       13       14       13       4       20       10       2       12       10       13       15       14       15       16       10       10       12       14       13       14       13       14	3 8 94 5 3 9 6 4 18 4 5 11
CULTURE & ARTS       19       8       8       13       8       13       8       13       8       13       8       14       9       16       2       12       6       16       2       12       6       16       2       12       6       16       13       15       14       15       16       13       16       12       16       2       12       16       2       12       16       17       10       13       15       14       15       16       16       16       16       16       17       16       17       16       17       16       17       16       17       16       17       16       17	
DIVORCE - 10       5       5       9       11       9       147       3       3       7       7       0       7       13       5       6       6       22       6       6       2       6       6       2       6       6       2       6       6       2       6       6       2       6       6       2       6       6       2       6       6       2       6       6       2       6       6       13       6       12       4       12       4       6       2       6       7       7       6       12       4       13       6       12       4       13       6       13       14       13       6       14 <th14< th=""> <th14< th="">       14</th14<></th14<>	21 17 135 17 18 20 10 6 20 19 13 14
EDUCATION - 6       6       22       2       1       4       0       26       2       3       2       3       6       12       4       12       4       6       2       2         ENTERTAINMENT - 15       11       2       17       6       2       1       2       36       5       3       15       8       13       4       13       6       12       4       10       6       2       3         ENVIRONMENT - 5       7       5       6       9       3       6       2       3       6       7       12       5       4       16       2       9       12         GOODNEWS - 7       7       1       10       9       9       5       10       16       10	2 3 7 13 6 9 11 32 5 18 15 6
ENTERTAINMENT       15       11       2       17       6       21       1       2       36       5       3       15       8       13       4       13       6       3       10       6       3       4         ENVIRONMENT       5       7       5       6       9       20       5       0       6       5       2       3       4       9       13       7       5       4       16       2       9       12         FIFTY       1       5       1       5       3       6       10       5       2       5       23       3       6       7       12       5       4       10       4       7       4       6         GOOD NEWS       7       7       1       10       9       9       5       10       15       8       10       6       13       4       4       7       4       3       6       14       16       20       6       11       5       14       4       13       14       14       14       14       16       11       16       11       13       12       14       14       14	2 3 19 3 5 7 7 4 1 8 1 5
ENVIRONMENT       5       7       5       6       9       20       5       0       6       58       2       3       4       9       13       7       5       4       16       2       9       12         FIFTY       1       5       1       5       3       6       10       5       2       5       23       3       6       7       12       5       4       10       2       9       13       7       12       5       4       10       2       10 <th10< th="">       10       10<td>4 8 89 11 9 7 4 5 20 5 18 7</td></th10<>	4 8 89 11 9 7 4 5 20 5 18 7
FIFT - 1       5       1       5       3       6       10       5       2       5       23       3       6       7       12       5       4       20       4       7       4       6         GOOD NEWS - 7       7       1       10       9       9       5       0       16       5       2       5       8       10       8       5       2       5       6       5       3       3         HEALTHY LIVING - 10       11       5       7       4       10       9       3       10       2       10       6       10       6       10       10       8       3       3       4       3       3       3         HEALTHY LIVING - 10       11       5       7       4       10       9       3       10       2       11       6       10	12 8 60 9 4 10 6 4 8 15 3 21
GOOD NEWS - 7       7       1       10       9       9       5       0       16       5       2       5       8       10       8       5       2       5       6       5       3       3         HEALTHY LIVING - 10       11       5       7       4       13       9       3       15       8       11       6       20       6       11       5       9       17       4       4       7       9         HOME & LIVING - 17       12       3       14       2       20       10       2       11       8       4       7       10       77       6       5       10       12       6       2       5       6       5       10       12       6       2       5       6       5       10       12       6       2       5       6       5       16       13       14       2       10       12       14       14       14       14       14       14       14       14       14       15       14       14       14       14       14       14       14       14       14       14       14       14       14       14	6 2 74 2 4 4 10 5 0 18 6 4
HEALTHY LIVING - 10       11       5       7       4       13       9       3       15       8       11       6       20       6       11       5       9       17       4       4       7       9         HOME & LIVING - 17       12       3       14       2       20       10       2       11       8       4       7       10       77       6       5       10       12       6       2       5       6         IMPACT - 5       15       11       6       9       18       9       11       7       4       10       12       6       2       5       10         MEDIA - 15       11       5       13       3       12       1       0       6       8       2       7       9       7       6       46       8       4       4       6       2       14         MEDIA - 15       11       5       13       3       12       1       0       6       8       1       16       9       6       4       4       6       2       1         MONEY - 3       22       4       11       4       16	3 9 87 3 9 14 5 8 25 3 7 4
HOME & LIVING - 17       12       3       14       2       20       10       2       11       8       4       7       10       77       6       5       10       12       6       2       5       6         IMPACT - 5       15       11       6       9       18       9       11       7       4       10       14       25       10       12       6       2       5       14         MEDIA - 15       11       5       13       3       12       1       0       6       8       2       7       4       10       14       25       10       15       18       11       13       12       14         MEDIA - 15       11       5       13       3       12       1       0       6       8       2       7       9       46       8       4       4       4       4       6       2         MONEY - 3       22       4       11       4       18       3       3       4       10       6       4       13       11       4       6       4       10       10       10       10       10       10       10<	9 8 86 6 7 9 14 8 10 32 8 4
IMPACT - 5       15       11       6       9       11       9       11       7       4       10       14       25       10       15       18       11       13       12       14         MEDIA - 15       11       5       13       3       12       1       0       6       8       2       7       9       6       46       8       4       4       6       2         MONEY - 3       22       4       11       4       18       3       3       3       4       10       6       4       13       14       4       6       2         PARENTING - 9       6       5       12       11       15       7       9       8       10       6       4       15       16       9       61       10       9       5       8       10 </td <td>6 13 16 14 22 11 10 9 19 9 7 7</td>	6 13 16 14 22 11 10 9 19 9 7 7
MEDIA - 15       11       5       13       3       12       1       0       6       8       2       7       9       7       6       46       8       6       44       4       6       2         MONEY - 3       22       4       11       4       18       3       3       4       10       6       4       13       11       16       9       61       10       9       5       8       9         PARENTING - 9       6       5       12       11       15       7       9       8       13       9       3       10       12       22       7       7       60       2       10       9       10	14 8 39 6 13 8 10 1 9 18 14 15
MONEY - 3       22       4       11       4       18       3       3       4       10       6       4       13       11       16       9       61       10       9       5       8       9         PARENTING - 9       6       5       12       11       15       7       9       8       13       9       3       10       12       22       7       7       60       2       10       9       10	2 12 82 8 4 15 5 2 18 4 9 15
PARENTING - 9 6 5 12 11 15 7 9 8 13 9 3 10 12 22 7 7 60 2 10 9 10	9 4 9 4 6 12 7 9 6 14 5 5
	10 8 22 9 11 12 8 12 7 20 20 6
POLITICS - 10 15 7 15 10 11 5 6 11 18 5 8 9 11 8 17 7 3 70 6 11 10	10 16 62 6 6 12 4 2 18 2 9 10
QUEER VOICES - 11 4 2 10 4 19 2 6 13 4 4 5 11 10 10 10 5 10 18 93 10 4	4 12 71 12 6 5 8 8 10 5 8 5
RELIGION - 7 9 3 8 9 18 5 5 12 7 4 3 15 13 10 5 6 14 16 9 53 9	9 8 118 4 5 9 12 4 9 6 12 13
SCIENCE - 10 10 12 9 6 15 7 0 15 11 1 4 11 10 4 8 3 6 5 5 8 58	58 10 89 7 7 18 7 4 18 19 9 8
SPORTS -         9         3         6         19         12         18         2         2         12         8         2         9         10         15         6         9         2         7         17         5         7         3	3 58 85 8 13 11 9 3 16 4 8 12
STYLE - 10 5 2 10 2 11 2 0 21 2 2 8 7 12 2 2 12 6 8 0 4	4 8 186 19 12 11 2 5 23 4 10 3
STYLE & BEAUTY - 16 9 8 11 5 18 5 1 8 4 2 8 11 19 8 6 6 10 7 5 6 0	0 9 27 109 7 7 15 13 10 6 9 6
TASTE - 10 8 6 14 6 23 10 0 14 11 3 8 4 24 11 5 9 10 9 2 6 8	8 13 27 11 62 19 9 5 24 16 8 2
TECH - 9 20 1 14 4 20 2 3 11 7 6 3 15 9 6 11 10 7 12 7 8 8	8 15 23 10 16 101 6 8 17 6 9 12
TRAVEL - 9 11 3 6 4 24 9 5 3 15 8 6 10 12 15 2 10 16 10 6 7 6	6 8 32 15 19 12 64 9 10 23 11 5
WEDDINGS - 10 6 2 5 1 8 35 1 9 5 7 3 6 15 2 2 5 13 6 7 8 8	8 3 10 8 6 8 12 <mark>168</mark> 6 7 5 3
WEIRD NEWS - 10 9 3 16 12 11 2 1 22 15 0 5 9 19 5 5 2 7 14 3 3 8	8 20 111 2 10 14 5 3 44 1 11 6
WELLNESS - 3 15 9 6 0 22 9 8 5 18 16 3 26 17 16 8 12 15 11 9 5 16	16 9 6 8 21 11 21 12 6 43 12 9
WOMEN - 20 10 8 5 4 12 5 0 16 10 8 4 9 7 15 9 6 14 15 12 10 10	10 6 58 11 15 15 9 6 11 14 39 7
WORLD - 13 11 3 8 19 13 4 2 6 18 7 6 7 12 8 7 6 7 15 8 6 10	10 16 77 7 3 10 10 4 18 6 7 50
ACK VOICES - BUSINESS - COLLEGE - COLLEGE - COLLEGE - CONEDY - CONEDY - DIVORCE - DIVORCE - DIVORCE - EDUCATION - RETAINMENT - FIFTY - SGOD NEWS - MAPACT - MEDIA - MONEY - MONEY - PARENTING - PARENTING - PARENTING - RELIGION - SCLENCE - SCLENCE - SCLENCE -	



- 175

- 150

- 125

- 100

- 75

- 50

- 25

- 0

#### Confusion Matrix reveals widespread misclassification across classes. STYLE(186), WEDDINGS(168), DIVORCE(147), CULTURE&ARTS(123) ----> WEDDINGS(228), CULTURE&ARTS(228), DIVORCE(207), STYLE(175)(RF)



- 200

- 150

- 100

- 50

- 0

Random Forest: Confusion Matrix for News Dataset

	BLACK VOICES - 74	5	9	16	38	35	2	4	20	1	3	4	2	8	4	8	0	7	10	4	7	4	20	41	10	9	11	3	3	12	6	6	6	
	BUSINESS - 8	55	1	8	7	29	2	3	5	g	2	2	2	15	7	11	35	8	14	5	1	0	20	54	5	15	20	9	2	9	16	9	10	
	COLLEGE - 5	2	53	2	4	5	0	23	3	1	0	1	1	1	3	2	11	3	0	1	Å	2	7	47	2	5	3	2	0	2	8	5	5	
	COMEDY - 9	2	1	53	4	23	3	1	24	4	4	12	5	20	4	9	3	8	22	2	4	6	20	71	5	23	12	5	2	19	4	9	6	
	CRIME - 20	1	1	11	139	14	2	4	7	9	1	1	0	6	4	4	0	7	7	4	3	6	15	86	1	9	7	4	2	16	3	7	8	
	CUITURE & ARTS - 19	7	5	15	12	228	8	2	21	3	2	10	8	37	5	5	4	6	9	3	7	15	18	121	10	25	8	9	3	18	7	12	11	
	DIVORCE - 3	2	3	6	5	15	207	1 ī	8	1	1	2	3	12	1	3	8	33	1	6	2	2	4	0	6	16	4	3	32	3	13	11	2	
	FDUCATION - 2	3	19	2	1	11	0	72	0	3	î	0	3	6	3	2	4	19	3	0	0	2	3	17	1	3	0	4	0	2	8	1	3	
	ENTERTAINMENT - 23	5	3	24	13	41	1	1	54	2	2	6	1	5	1	12	1	6	9	2	1	2	19	90	7	14	4	3	4	23	3	3	2	
	ENVIRONMENT - 8	5	1	8	8	26	2	2	1	88	1	2	6	12	5	3	5	7	17	1	ĩ	13	6	51	6	13	10	7	3	8	11	1	17	
	FIFTY - 2	4	0	2	0	7	10	1	0	2	29	0	5	12	6	1	12	35	3	4	0	2	1	72	4	5	2	9	2	0	29	10	2	
	GOOD NEWS - 8	2	2	19	9	18	1	0	16	5	1	9	6	9	4	4	3	12	2	3	ĩ	4	10	87	1	13	9	2	5	29	5	4	2	
	HEALTHY LIVING - 7	10	ĩ	4	3	21	7	3	6	5	10	3	25	16	9	3	7	22	3	6	2	7	7	83	4	17	4	6	6	11	54	8	6	
	HOME & LIVING - 13	1	0	9	2	43	1	2	11	1	0	3	9	146	2	2	4	7	0	1	0	1	10	16	14	45	5	9	6	7	8	5	3	
S	IMPACT - 11	11	6	7	8	33	6	15	4	14	4	2	4	14	42	2	14	23	6	5	11	7	8	34	6	19	13	8	5	7	29	14	8	
bel	MEDIA - 10	7	2	15	6	18	1	1	11	2	0	4	3	1	0	96	2	7	43	1	3	1	14	76	8	13	13	3	0	15	2	6	11	
a	MONEY - 4	13	5	4	9	12	6	3	5	5	2	3	11	12	4	5	126	8	6	1	з	0	5	2	5	13	12	5	7	3	16	4	5	
e	PARENTING - 11	4	2	8	11	25	4	7	7	4	6	2	3	19	8	4	3	146	3	1	1	8	6	16	9	11	6	8	7	9	18	12	2	
T	POLITICS - 9	9	4	19	15	20	4	4	12	11	1	5	10	6	4	14	12	4	99	3	9	4	18	55	3	11	6	4	0	10	3	9	23	
	QUEER VOICES - 8	1	2	8	4	23	10	4	24	3	2	6	2	9	1	6	1	13	13	121	6	1	7	64	8	14	7	8	7	10	8	5	9	
	RELIGION - 13	2	6	3	7	16	2	3	6	5	3	8	3	9	11	3	2	13	15	11	85	6	14	114	2	6	9	9	2	10	15	9	18	
	SCIENCE - 11	7	2	9	7	22	1	3	10	8	1	6	9	15	3	7	1	6	7	5	4	95	10	76	5	13	8	2	3	21	19	11	7	
	SPORTS - 16	1	4	12	11	22	0	1	12	7	3	11	6	15	5	4	2	4	11	6	3	4	102	74	7	15	5	5	0	23	1	8	10	
	STYLE - 15	4	0	11	2	26	4	0	24	1	0	7	7	19	4	3	0	4	7	2	0	5	8	175	19	22	8	5	2	21	2	3	3	
	STYLE & BEAUTY - 6	3	1	8	3	30	3	1	17	1	3	4	2	29	1	6	6	8	3	4	2	4	4	10	159	12	7	9	8	12	8	13	4	
	TASTE - 10	6	1	17	1	31	0	0	9	1	3	5	6	40	3	2	4	10	3	4	2	6	14	23	7	139	13	11	2	12	8	3	1	
	TECH - 11	16	1	13	5	12	1	2	10	4	0	5	4	19	4	15	7	9	10	5	5	14	22	12	3	25	128	5	2	17	10	8	12	
	TRAVEL - 3	3	0	4	5	38	6	1	7	7	6	3	6	18	10	4	5	16	8	6	2	15	6	21	15	30	8	121	3	8	12	1	7	
	WEDDINGS - 2	2	0	4	3	19	33	0	9	2	2	1	2	15	2	1	5	7	1	1	2	2	5	2	18	7	4	5	228	0	10	4	2	
	WEIRD NEWS - 12	2	0	19	27	23	1	0	30	5	0	7	2	19	1	9	0	6	5	4	2	9	25	105	4	14	11	1	2	50	1	7	5	
	WELLNESS - 6	6	4	9	1	25	11	6	2	7	10	2	21	17	10	4	11	18	1	3	3	8	7	3	12	23	9	10	6	6	130	10	6	
	WOMEN - 16	3	4	12	6	27	9	4	13	4	6	5	6	18	8	7	4	20	5	5	6	3	11	52	13	8	6	0	3	19	22	70	5	
	WORLD - 9	5	4	8	27	21	3	0	6	16	0	3	3	7	5	6	4	8	21	4	7	9	16	68	3	11	2	7	0	14	3	6	98	
	N.	Ś	ш	7	ш	S	ш	Z	÷.	÷.	7	Ś	ق	Ū.	÷	A	7	5	Ś	S	z	ш	Ś	ц.	7	ш	Ŧ	Ĺ.	Ś	Ś	Ś	z	Ó	
	E	ES	B		M	RT	RC	0	EN	EN	E	Ň	IN	IN IN	AC	D	NE	Z	2	U	0	NC	RT	Υ	5	ST	ы	N	PN BN	N	ES	ЧE	RL	
	Q	N	E	M	Ю	X P	0	AT	Σ	Σ	Ē	z		È	MP,	ME	Q	L	5	0	9	山	PO	S	EA	TA	F	R	D	Z	L	Q	9	
	Ý	SUS	8	2		Щ	D	S	AIN	NO		OD	≿	S	=		~	RE	B	R	REI	S	S		B X			Г	ED	RD	ΈL	5	>	



### Random Forest Results

Processing dataset: Lette	ers Dataset				Processing dataset: Letters Dataset	
Evaluation for L Accuracy: 0.86 Precision (Maci Recall (Macro): F1 Score (Mac F1 Score (Micro	Letters Dataset 683 ro): 0.8346 : 0.7447 rro): 0.7700 o): 0.8683				Evaluation for Letters Dataset Accuracy: 0.8150 Precision (Macro): 0.6976 Recall (Macro): 0.6903 F1 Score (Macro): 0.6931 F1 Score (Micro): 0.8150	
Classification R	Report: precisi	on recall f1-	score supp	port	Classification Report: precision recall f1-score s	support
Franz Kafka	0.85	0.56	0.68	280	Franz Kafka 0.58 0.50 0.54	280
Friedrich Schille	er 0.70	0.68	0.69	266	Friedrich Schiller 0.46 0.50 0.48	266
Henrik Ibsen	1.00	0.97	0.99	897	Henrik Ibsen 0.99 0.98 0.98	897
James Joyce	0.97	0.63	0.77	682	James Joyce 0.75 0.72 0.74	682
Johann Wolfgang von Goe	the 0.72	0.39	0.50	228	Johann Wolfgang von Goethe 0.45 0.41 0.43	228
Virginia Woolf	0.88	1.00	0.94	1901	Virginia Woolf 0.90 0.91 0.91	1901
Wilhelm Busch	0.72	0.97	0.83	627	Wilhelm Busch 0.76 0.81 0.78	627
accuracy		0.87	4881		accuracy 0.81 4881	
macro avg	0.83	0.74	0.77	4881	macro avg 0.70 0.69 0.69	4881
weighted avg	0.88	0.87	0.86	4881	weighted avg 0.81 0.81 0.81	4881

**Decision Tree Results** 

#### Accruacy: 81.50% vs. 86.85% (+5% improvement)

Virginia Woolf and Henrik Ibsen dominate in classification accuracy; authors like Franz Kafka and Johann Wolfgang von Goethe have weaker recall scores( a higher number of texts by these authors are misclassified as being written by the others).

Macro F1: 77.00% vs. 69.31%. RF showing improvement over DT, especially for underperforming classes like Freidrich Schiller and Franz Kafka.

# **Confusion Matrix**





The dataset is somewhat imbalanced, with authors like Virginia Woolf having many samples while Johann Wolfgang von Goethe has fewer.



Decision Tree: Confusion Matrix for Letters Dataset



### Random Forest Results

Processing datas	set: Letters I	_anguage E	Dataset		Processing dataset: Letters Language Dataset								
Evalua Accur Precis Recall F1 Sco F1 Sco	ation for Lette acy: 0.9982 ion (Macro): (Macro): 0.9 ore (Macro): 0 ore (Micro): 0	ers Languag 0.9810 763 0.9786 .9982	e Dataset		Evaluation for Letters Language Dataset Accuracy: 0.9953 Precision (Macro): 0.9827 Recall (Macro): 0.9449 F1 Score (Macro): 0.9627 F1 Score (Micro): 0.9953								
Classification Rep	ort: precision r	ecall f1-scc	ore support		Classification Report: precision recall f1-score support								
da de en fr it	1.00 1.00 1.00 0.97 0.94	1.00 1.00 1.00 0.95 0.94	1.00 1.00 1.00 0.96 0.94	844 1448 2519 38 32	da         0.99         1.00         1.00         844           de         1.00         1.00         1448           en         1.00         1.00         1.00         2519           fr         1.00         0.92         0.96         38           it         0.93         0.81         0.87         32								
accuracy macro avg weighted avg	0.98 1.00	1.00 0.98 1.00	4881 0.98 1.00	4881 4881	accuracy1.004881macro avg0.980.940.964881weighted avg1.001.001.004881								

**Decision Tree Results** 

Accruacy: 99.82% vs. 99.53% (+0.29% improvement)

Both models excelled, but RF outperformed DT slightly (accuracy: +0.29%, Macro F1: +1.59%).

The small number of classes and clear language distinctions likely contributed to the high performance.





Decision Tree: Confusion Matrix for Letters Language Dataset

Almost perfect performance across all classes, with minimal misclassification, mostly for classes like "fr" and "it." nostly for classes like "fr" and "it.".

# **Results comparison between Datasets**

### Sentiment Dataset:

- Both DT and RF performed well, but RF achieved a significant performance boost (accuracy: +13%).
- Binary classification is simpler for both models compared to other datasets.

### **News Dataset:**

- Both models struggled due to the high number of classes and imbalanced data.
- RF consistently outperformed DT across all metrics, showing it handles complex multi-class problems better.
- The need for feature engineering or resampling techniques is apparent to improve performance.

### **Letters Dataset:**

### Author:

- RF showed noticeable improvements over DT (accuracy: +5%, Macro F1: +8%,)
- Both models managed this multi-class problem reasonably well.

### Language:

- Both models excelled, but RF outperformed DT slightly (accuracy: +0.29%, Macro F1: +1.59%).
- The small number of classes and clear language distinctions likely contributed to the high performance.

# Conclusion: Decision tree vs. Random forest

- RF consistently outperforms DT across all datasets, demonstrating its robustness and ability to handle complex, imbalanced, and multi-class data.
- The largest performance gap is observed in the News Dataset, where RF's ensemble approach mitigated overfitting and improved predictions for minority classes.

Best Dataset Performance: Letters Language Dataset, with RF achieving near-perfect results.

Challenging Dataset: News Dataset due to high class imbalance and complexity.

RF is the superior choice across datasets, especially for imbalanced or multi-class problems, whereas DT is a simpler and faster baseline.

Further steps could include hyperparameter tuning for RF, oversampling techniques (e.g., SMOTE) for News Dataset, and advanced feature extraction for all datasets to boost performance.

# Literature

- Categorical Modeling/Automatic Interaction Detection by William A.V. Clark, Marinus C. Deurloo, in Encyclopedia of Social Measurement, 2005 (History and explanation of DecisionTrees);
- Boosting by Peter Wittek, in Quantum Machine Learning, 2014 (Decision Tree Classifier);
- Introduction to Random forest Simplified, Tavish Srivastava, 2020 ; https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/
- All About Decision Tree from Scratch with Python Implementation, 2024 ; https://www.analyticsvidhya.com/blog/2020/10/all-about-decision-tree-from-scratch-with-python-implementation/
- Sklearn, Decision Trees ; https://scikit-learn.org/1.5/modules/tree.html
- Tree Based Algorithms: A Complete Tutorial from Scratch (in R & Python), Analytics Vidhya, 2024 ; https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/?utm\_sourc e=blog&utm\_medium=decision-tree-vs-random-forest-algorithm
- Random Forest Classification with Scikit-Learn, Adam Shafi, 2024 ; https://www.datacamp.com/tutorial/random-forests-classifier-python
- Sklearn-Ensemble, RandomForestClassifier ; https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- Random Forest Classifier Tutorial ; https://www.kaggle.com/code/prashant111/random-forest-classifier-tutorial

# **History - First appearence**

- Ronald Fisher's paper on discriminant analysis (1936);
- AID project by Morgan and Sonquist and the 1966 publication by Hunt;
- Theta Automatic Interaction Detection (THAID) project by Messenger and Mandell (1972) the first classification tree;
- Berkeley Statistics professors Leo Breiman, Charles Joel Stone, Jerome H.
   Friedman and Richard Olshen from Stanford University, began developing the classification and regression tree (CART) algorithm, unveiled in 1977

# **History - Improvements**

- 1984 the first official publication with a CART software;
- Computer science researcher John Ross Quinlan invented a new concept: trees with multiple answers; continued to upgrade until It was ranked No. 1 in the Top 10 Algorithms in Data Mining at the IEEE ICDM Conference (2006);
- Random Forests The first such algorithm was created in 1995 by Tin Kam Ho

# Thank you!



python DT\_Countvec\_notNormalized.py 83.55s user 2.90s system 29% cpu
4:57.44 total

python RF\_classify\_extrametric\_noNorm.py 199.47s user 3.94s system 1
% cpu 2:52:27.59 total