GROUP J

DATA MINING FRAMEWORK

INSTRUCTOR: DR. NGUYEN THI THANH SANG

TEAM MEMBER

NGUYỄN VĂN NGỌC THI
ITCSIU22251

PHẠM MINH KHÁNH
ITCSIU21194

TRỊNH VŨ THẾ ANH
ITCSIU22009

LÊ TRỌNG ĐẠI
ITITIU20176

TABLE OF CONTENTS

2

DATA PREPROCESSING

2.1 DATA PREPROCESSING

THE DATASET CONTAINS 13,200 INSTANCES (ROWS) AND 11 ATTRIBUTES (COLUMNS):

11 Atributes	**Temperature**	**Humidity**	**Wind Speed**	**Precipitation**	**Cloud Cover**
Atmospheric Pressure	**UV Index**	**Season**	**Visibility**	**Location**	**Weather Type**

KEY CHARACTERISTICS

Mix of Numerical and Categorical Data	Class Imbalance
Data Range and Scaling	Duplicate-Free

2.2 DATA CLEANING PROCESS

1 **NORMALIZATION**

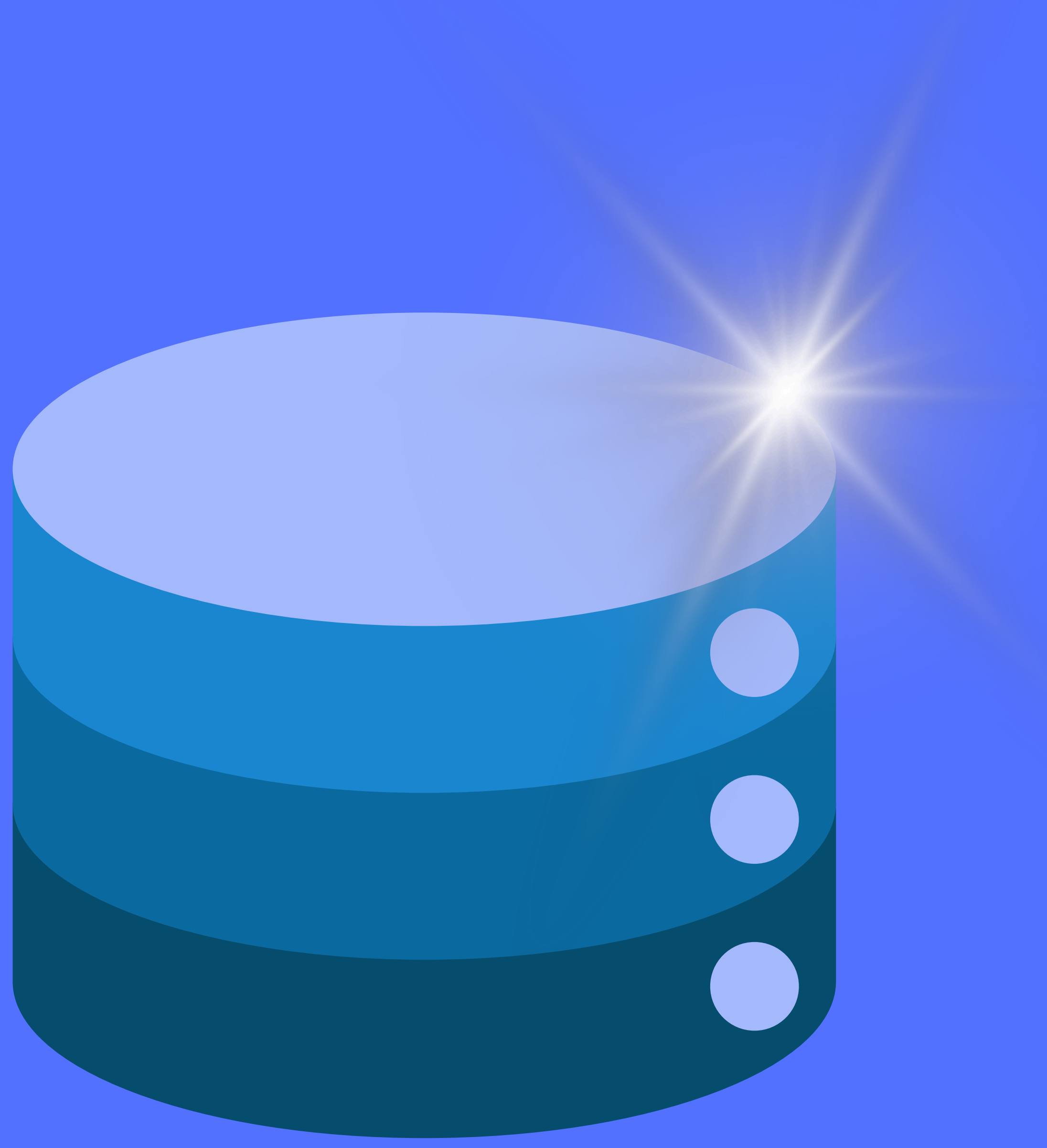
Scales numerical data into a uniform range

2 **ENCODING**

Converts categorical values into numeric representations

3 **EXECUTION**

Filters the dataset to retain only specific features



2.2 DATA CLEANING PROCESS

4 **MISSING DATA HANDLING**

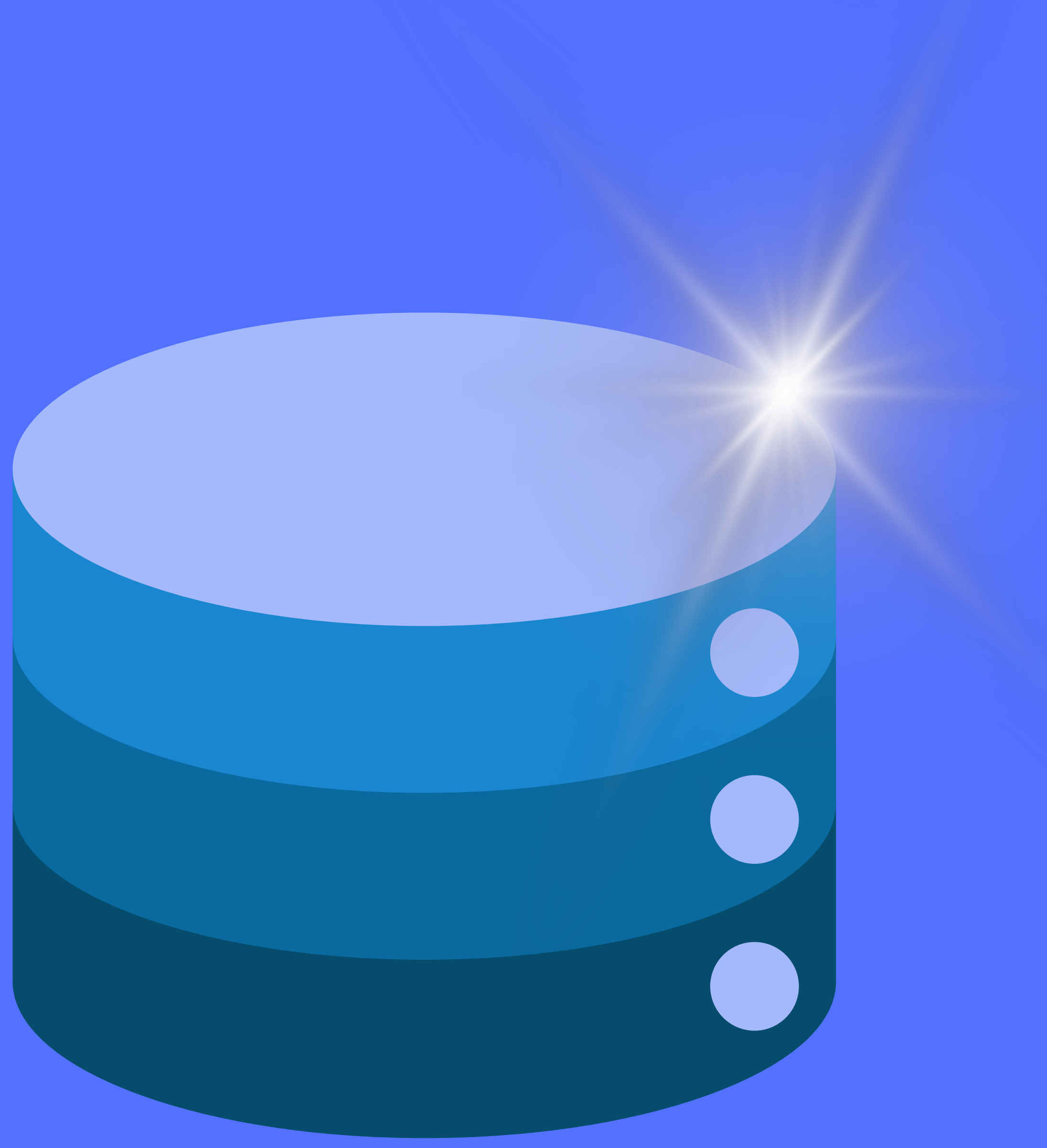
Replaces missing/invalid values with column modes

5 **DEDUPLICATION**

Ensures unique rows in the dataset

6 **DATA SPLITTING**

Divides data into training and test sets



2.3 DATA TRANSFORMATION

1. NORMALIZATION

Purpose: To scale numerical data into a consistent range (usually between 0 and 1) for easier comparison and to prevent features with large magnitudes from dominating the learning algorithm.

Implementation:

- *Finding Min/Max Values: The program computes the minimum and maximum values for each numeric column across the dataset.*
- *Scaling: Each numeric value is transformed using the formula*

$$\text{normalized\_value} = \frac{\text{value} - \text{min}}{\text{max} - \text{min}}$$

- *Handling Zero Range: If max == min for a feature (e.g., all values are the same), the normalized value is set to 0.*
- *Impact: Ensures numerical features are on a comparable scale, enhancing the performance of distance-based algorithms (e.g., k-NN, clustering) and gradient-based optimizations.*

2.3 DATA TRANSFORMATION

2. ENCODING CATEGORICAL VARIABLES

Purpose: To convert categorical data (non-numeric) into numeric form, as most machine learning algorithms require numerical inputs.

Implementation:

- *A unique integer is assigned to each unique categorical value in a column.*
- *The mapping for each categorical column is stored in a map (encoders), where the key is the categorical value, and the value is its numeric encoding.*
- *Rows in the dataset are transformed based on these mappings, with unknown or missing values defaulting to -1.*

Impact: Transforms non-numeric features into numeric form while preserving the uniqueness of each category. This step is a prerequisite for applying most machine learning algorithms.

2.3 DATA TRANSFORMATION

3. FEATURE SELECTION

Purpose: To reduce dimensionality by selecting only the most relevant features, thereby simplifying the model and potentially improving its performance.

Implementation:

- *The program accepts a list of selected features (selectedFeatures) as input.*
- *It determines the indices of the selected features in the header and extracts only these columns from the dataset.*

Impact:

- *Reduces computational overhead and risk of overfitting.*
- *Focuses the model on features likely to have the most predictive power, improving interpretability and performance.*

3. SEQUENTIAL PATTERN MINING

METHODOLOGY

- DATA PREPROCESSING: REMOVE NUMERIC ATTRIBUTES
- MINING FREQUENT PATTERNS: USE APRIORI ALGORITHM
- RULE EVALUATION

3. SEQUENTIAL PATTERN MINING

RESULTS AND FINDINGS

- BEST RULES FOUND

1. CLOUD COVER=CLEAR ⇒ WEATHER TYPE=SUNNY (CONFIDENCE: 100%, LIFT: 4)
2. CLOUD COVER=OVERCAST, WEATHER TYPE=SNOWY ⇒ SEASON=WINTER (CONFIDENCE: 97%)
3. LOCATION=INLAND, WEATHER TYPE=SNOWY ⇒ SEASON=WINTER (CONFIDENCE: 96%)
4. LOCATION=MOUNTAIN, WEATHER TYPE=SNOWY ⇒ SEASON=WINTER (CONFIDENCE: 96%)
5. WEATHER TYPE=SNOWY ⇒ SEASON=WINTER (CONFIDENCE: 94%)

3. SEQUENTIAL PATTERN MINING

RESULTS AND FINDINGS

- BENEFITS

- LIMITATIONS AND CHALLENGES

4. CLASSIFICATION ALGORITHM

IMPLEMENTATION PROCESS

Convert data to ARFF

```
public static void csvToArff(String filePath) throws Exception {  1 usage  ThiNVN
    //Load csv file
    CSVLoader loader = new CSVLoader();
    loader.setSource(new File(filePath));
    Instances dataset = loader.getDataSet();

    //Save as arff format
    ArffSaver saver = new ArffSaver();
    saver.setInstances(dataset);
    saver.setFile(new File( pathname: "src/data/weather_classification.arff"));
    saver.writeBatch();
}
```

4. CLASSIFICATION ALGORITHM

MODEL SELECTION

Comparison: OneR vs NaiveBayes

	OneR	**NaiveBayes**
Accuracy	Correctly Classified Intances: 66.89%	Correctly Classified Intances: 87.22%
Precision, Recall F-Measure	Precision: 55.23% Recall: 44.06% F-Measure: 49.02%	Precision: 83.22% Recall: 85.31% F-Measure: 84.25%
AUC	0.3838	0.8428
Error Measures	Mean Absolute Error: 0.1655 Root Mean Squared Error: 0.4069 Relative Absolute Error: 44.14% Root Relative Squared Error: 93.95%	Mean Absolute Error: 0.0816 Root Mean Squared Error: 0.2262 Relative Absolute Error: 21.76% Root Relative Squared Error: 52.22%

5. IMPROVEMENT PREVIOUS MODEL

METHODOLOGY

- DATA PREPROCESSING: ONE-HOT ENCODING
- CLUSTERING WITH SIMPLEKMEANS
- CLUSTER EVALUATION

5. IMPROVEMENT PREVIOUS MODEL

CLASSIFICATION WITHIN CLUSTERS

- CLUSTER SPECIFIC DATASET

- SPLIT TRAINING AND TESTING SET

- TRAIN J48 DECISION TREE

5. IMPROVEMENT PREVIOUS MODEL

CLASSIFICATION WITHIN CLUSTERS

- MODEL EVALUATION

	Cluster#0	Cluster#1	Cluster#2
Accuracy	92.3891 %	91.4857 %	96.6872 %

- OVERALL ACCURACY

$$\frac{\sum(Accuracy * Number\ of\ instances\ in\ testset)}{\sum(Number\ of\ instances\ in\ dataset)} = 93.535\%$$

5. IMPROVEMENT PREVIOUS MODEL

EVALUATION AND COMPARISION

	Initial model (NaiveBayes)	J48 Tree apply on full dataset	Improved model (clustering and classification)
Accuracy	87.2222 %	90.9091 %	93.535%

IMPROVE 6.32%

6. MODEL EVALUATION

PERFORMANCE METRICS

	Cluster#0 (5037)	Cluster#1 (3837)	Cluster#2 (4326)
Accuracy	90.8874 %	92.2075 %	96.0472 %
Precision	0.909	0.920	0.961
Recall	0.909	0.922	0.960
f-Measure	0.909	0.921	0.961
Runtime	0.11s	0.06s	0.08s

6. MODEL EVALUATION

INSIGHTS AND TRADE-OFFS

RECOMMENDATIONS FOR IMPROVEMENT

CONCLUSION

KEY FINDINGS

LESSON LEARNED

FUTURE DEVELOPMENT

THANKS FOR LISTENING