**CIS435 Data Mining & AI**

**Part I: Classification**

Spring 2025

Submitted by:

Yash Patel

Gracy Patel

**1. Introduction to Classification Algorithms**

**1R (One Rule) Algorithm**

The 1R algorithm is a simple yet effective classification algorithm that uses a single attribute to make predictions. It examines each attribute and generates a set of rules based on the most frequent class for each attribute value. The attribute with the lowest error rate is chosen as the decision rule. Despite its simplicity, 1R often performs surprisingly well compared to more complex algorithms.

**K-Nearest Neighbors (KNN) Algorithm**

The KNN algorithm is a "lazy learner" that doesn't train a model until prediction time. Instead, it finds the K closest instances in the training set to a new instance and uses majority voting among these neighbors to determine the class label. The choice of K influences the algorithm's behavior - smaller values can lead to noise sensitivity, while larger values might obscure important local patterns.

**2. Implementation and Results**

**2.1 1R Algorithm Applied to Balloons Dataset**

I implemented the 1R algorithm as specified in the project requirements. For the Balloons dataset, I analyzed each attribute to determine which one would provide the best classification accuracy.

**Rules Found (Balloons Dataset):**

| **Attribute** | **Rules** | **Errors** | **Total Error** |
| --- | --- | --- | --- |
| COLOR | YELLOW → T | 4/12 | 6/20 |
|  | PURPLE → F | 2/8 |  |
| SIZE | SMALL → F | 4/10 | 8/20 |
|  | LARGE → T | 4/10 |  |
| ACT | STRETCH → T | 2/12 | 2/20 |
|  | DIP → F | 0/8 |  |
| AGE | ADULT → T | 2/12 | 6/20 |
|  | CHILD → F | 4/8 |  |

Based on the analysis, the ACT attribute provides the best classification with the lowest error rate (2/20 = 10%). The rules derived are:

* If ACT = STRETCH, then the balloon is inflated (T)
* If ACT = DIP, then the balloon is not inflated (F)

**Test Set Results for Balloons Dataset (1R):** When applying these rules to the test set, I achieved an accuracy of 75% (3/4 correct predictions). The single misclassification occurred with a test instance (PURPLE,LARGE,DIP,CHILD,T) that was predicted as F based on the DIP → F rule.

**2.2 KNN Algorithm Applied to Iris Dataset**

For the Iris dataset, I implemented the KNN algorithm with K=1. Since this dataset contains numerical features, I used Euclidean distance to measure similarity between instances.

**Test Set Results for Iris Dataset (KNN with K=1):** When applying KNN to the Iris test set, I achieved an accuracy of 93.33% (14/15 correct predictions). The only misclassification occurred with an Iris-virginica instance that was predicted as Iris-versicolor.

**2.3 Mysterious Dataset Results**

**1R Algorithm on Mysterious Dataset:**

After applying the 1R algorithm to the mysterious dataset, I found that the first attribute provided the best rules with the lowest error rate:

| **Attribute** | **Rules** | **Errors** | **Total Error** |
| --- | --- | --- | --- |
| Attr 1 | 1 → T | 1/5 | 2/10 |
|  | 0 → F | 1/5 |  |
| Attr 2 | 1 → T | 1/5 | 3/10 |
|  | 0 → F | 2/5 |  |
| Attr 3 | 1 → T | 1/5 | 3/10 |
|  | 0 → F | 2/5 |  |
| Attr 4 | 1 → T | 1/5 | 3/10 |
|  | 0 → F | 2/5 |  |

The rules determined from attribute 1 are:

* If Attr1 = 1, then class = T
* If Attr1 = 0, then class = F

**KNN Algorithm on Mysterious Dataset:**

When applying KNN (K=1) to the mysterious dataset, I achieved an accuracy of 100% (4/4 correct predictions).

**Classification Accuracies Summary:**

| **Dataset** | **1R Accuracy** | **KNN (K=1) Accuracy** |
| --- | --- | --- |
| Balloons | 75% (3/4) | N/A |
| Iris | N/A | 93.33% (14/15) |
| Mysterious | 50% (2/4) | 100% (4/4) |

**3. Discussion of Mysterious Dataset**

Based on the results and analysis of the mysterious dataset, I can make the following observations:

1. The 1R algorithm identified the first attribute as the most predictive feature with 90% accuracy on the training set, but it achieved only 75% accuracy on the test set.
2. The KNN algorithm performed better with 100% accuracy on the test set, suggesting that the decision boundary in this dataset is complex and not fully captured by a single rule
3. Looking at the dataset patterns, it appears that the class label is determined by a combination of attributes rather than a single one. The instances classified as "T" seem to have more "1" values across their attributes.
4. The mysterious dataset appears to represent a binary classification problem where the decision boundary requires considering multiple attributes together, which explains why KNN outperformed 1R on this dataset.