Contents

1 Abstract

This homework report investigates the performance of different classification methods, namely Decision Trees, Random Forest, Logistic Regression, and k-Nearest Neighbors (kNN), over a custom created lazy classification classifier based on Formal Concept Analysis (FCA). The study uses three diverse datasets from Kaggle namely, 'Heart Attack Analysis & Prediction Dataset', 'Stroke Prediction Dataset' and 'Water Quality Dataset', to assess the performance of these methods across varied data characteristics. This report represents a comprehensive exploration of commonly used classification techniques as well as how well the custom FCA Binary and Pattern classifiers hold up to them, giving a practical understanding of their applicability in real-world scenarios.

2 Experiments and Results

2.1 Datasets

Any missing/null values were filled with respective means of their features across the dataset and dummy variables were created for categorical variables for all datasets wherever applicable.

2.1.1 Heart Attack Analysis & Prediction Dataset

The heart attack analysis and prediction dataset from Kaggle serves as a valuable resource for understanding and predicting factors associated with cardiac events. This dataset encompasses a range of features, each offering unique insights into the contributors to heart attacks. The key attributes are:

- **Age (age)**: The age of the individual, a crucial factor in heart health assessment.
- **Sex (sex)**: Denoting the gender of the patient, recognizing potential gender-based variations in heart attack occurrences.
- **Chest Pain Type (cp)**: Classifying chest pain into different types: 0-Typical Angina, 1-Atypical Angina, 2-Non-Anginal Pain, 3-Asymptomatic. All of which aides in the identification of symptomatic patterns.
- **Resting Blood Pressure (trestbps)**: Measuring the resting blood pressure, a fundamental metric in cardiovascular health assessment.
- **Serum Cholesterol (chol)**: The level of serum cholesterol, providing insights into lipid profiles and their impact on heart health.
- **Fasting Blood Sugar (fbs)**: Indicating the presence of fasting blood sugar, a factor associated with diabetes and its influence on heart conditions, 1-True, 0-False.
- **Resting Electrocardiographic Results (restecg)**: Describing the resting electrocardiographic results, 0-Normal, 1-ST-T Wave Normality, 2-Left Ventricular Hypertrophy. Aiding in the diagnosis of cardiac abnormalities.
- **Maximum Heart Rate Achieved (thalach)**: The maximum heart rate achieved during exercise, a dynamic indicator of cardiovascular fitness.
- **ST Depression Induced by Exercise Relative to Rest (oldpeak)**: Quantifying ST depression induced by exercise, offering insights into cardiac stress during physical activity.
- **Slope (slp)**: Slope, 0-Negative, 1-Flat, 2-Positive.
- **Coronary Arteries (caa)**: Number of major vessels. Ranges from 0 to 3

- **Thallium Stress Result (thall)**: Shows how well blood flows into the heart. 0-Thalassemia, 1-Fixed defect (no blood flow in some part of the heart), 2-Normal Blood Flow, 3-Reversible defect (a blood flow is observed but it is not normal)
- **Exercise Induced Angina (exang)**: Identifying the presence of exercise-induced angina, a symptom often associated with heart-related issues. 1-True, 0-False.
- **Target (output)**: Shows whether heart attack occurs or not. 1-True, 0-False.

2.1.2 Stroke Prediction Dataset

The stroke prediction dataset from Kaggle serves as a valuable resource for understanding and predicting factors associated with stroke events. This dataset encompasses a range of features, each offering unique insights into the contributors to strokes. The key attributes are:

- **Gender (gender)**: Denoting the gender of the patient, recognizing potential gender-based variations in stroke occurrences.
- **Age (age)**: The age of the individual, a crucial factor in stroke assessment.
- **Hypertension (hypertension)**: Binary indicator of hypertension status, a significant risk factor for strokes. 1-True, 0-False.
- **Heart Disease (heart_disease)**: Binary indicator of the presence of heart disease, contributing to overall cardiovascular risk. 1-True, 0-False
- **Marital Status (ever_married)**: Describing the marital status of the individual, which might have socio-economic implications on health. 1-True, 0-False
- **Work Type (work_type)**: Categorizing the type of work the person is engaged in, providing insights into lifestyle factors. The 5 categories are: Child, Never Worked, Self-Employed, Private, Government Job
- **Residence Type (Residence_type)**: Differentiating between urban and rural residences, capturing potential environmental influences on health. Rural-0, Urban-1.
- **Average Glucose Level (avg_glucose_level)**: Quantifying the average glucose level, offering insights into metabolic health.
- **Body Mass Index (bmi)**: Measuring the body mass index, a key indicator of weight-related health.
- **Smoking Status (smoking_status)**: Classifying individuals based on their smoking habits, a known contributor to stroke risk. The 4 categories are: Unknown, Never Smoked, Formerly Smoked, Smokes.
- **Target (stroke)**: The target variable indicating whether an individual has experienced a stroke or not. 1-True, 0-False.

The missing values in BMI are replaced by the column mean. The marriage and residence columns are binarized, while for the gender, work type and smoking status columns dummies are created.

2.1.3 Water Potability Dataset

The water potability dataset from Kaggle, presents a valuable resource for assessing the safety and quality of drinking water. The dataset encompasses various features, each offering insights into the factors influencing water potability. The key attributes are:

- **pH Value (ph)**: Measuring the acidity or alkalinity of the water, a critical parameter for assessing its quality.
- **Hardness (Hardness)**: Quantifying the concentration of minerals, particularly calcium and magnesium, providing information on water quality and potential scale formation.
- **Solids (Solids)**: Representing the total amount of dissolved substances in water, indicating its purity.
- **Chloramines (Chloramines)**: Measuring the presence of chloramines, commonly used as disinfectants, impacting taste and odor.
- **Sulfate (Sulfate)**: Indicating the concentration of sulfate ions, which can affect taste and have health implications.
- **Conductivity (Conductivity)**: Measuring the water's ability to conduct an electric current, offering insights into dissolved ion content.
- **Organic Carbon (Organic_Carbon)**: Assessing the presence of organic carbon compounds, which may impact water quality and treatment processes.
- **Trihalomethanes (Trihalomethanes)**: Quantifying the concentration of trihalomethane compounds, formed during water disinfection.
- **Turbidity (Turbidity)**: Reflecting the cloudiness or haziness of water, an important aesthetic and health-related parameter.
- **Target (Potability)**: The target variable indicating whether water is potable (safe for drinking) or not. 1-True, 0-False.

Missing values for pH, Sulfates and Trihalomethanes are replaced with the column means.

These datasets will be revisited later when performing FCA based lazy classification for binarization of data in order to make it suitable for entering into the classifier(s).

2.2 Evaluation of Standard Classification Models

A comprehensive analysis employing four classification methods was conducted, methods being: Decision Tree, Random Forest, Logistic Regression, and k-Nearest Neighbors (kNN). For each method, Functions were crafted to facilitate model training, testing, and evaluation across three distinct datasets. The objective was to produce the optimal parameters through a systematic grid search approach and evaluate the best models across the three datasets across 5 cross validation folds and test set predictions for later comparison with the Formal Concept Analysis (FCA) based classifiers. The metrics used were the weighted F1 score (to deal with heavy class imbalance) as well as accuracy.

2.2.1 Decision Tree

The parameter grid for tuning the Decision Tree:

```
param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

2.2.2 Random Forest

The parameter grid for tuning the Random Forest:

```
param_grid = {
    'n_estimators': [25, 50, 100],
    'criterion': ['gini', 'entropy'],
    'max_depth': [None, 5, 10, 15],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

2.2.3 Logistic Regression

The parameter grid for tuning the Logistic Regression:

```
param_grid = {
    'penalty': ['l1', 'l2', 'elasticnet'],
    'C': [0.001, 0.01, 0.1, 1, 10, 100],
    'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'],
    'max_iter': [500, 750, 1000]
}
```

2.2.4 k-Nearest Neighbors (kNN)

The parameter grid for tuning kNN:

```
param_grid = {
    ’n_neighbors’: k_values,  # Try different k values
    ’weights’: [’uniform’, ’distance’],
    ’algorithm’: [’auto’, ’ball_tree’, ’kd_tree’, ’brute’],
    ’p’: [1, 2]  # 1 for Manhattan distance, 2 for Euclidean distance
}
```

2.2.5 Best Parameters

The outcomes of the grid search revealed the following best parameters for each method across the three datasets:

Decision Tree:

Heart Attack Dataset:

```
{
’criterion’: ’entropy’,
’max_depth’: 30,
’min_samples_leaf’: 1,
’min_samples_split’: 2
}
```

Stroke Dataset:

```
{
’criterion’: ’gini’,
’max_depth’: 10,
’min_samples_leaf’: 4,
’min_samples_split’: 10
}
```

Water Potability Dataset:

```
{
’criterion’: ’entropy’,
’max_depth’: 10,
’min_samples_leaf’: 4,
’min_samples_split’: 2
}
```

Random Forest:

Heart Attack Dataset:

```
{
’criterion’: ’entropy’,
’max_depth’: 15,
’min_samples_leaf’: 4,
’min_samples_split’: 2,
’n_estimators’: 50
}
```

Stroke Dataset:

```
{
’criterion’: ’entropy’,
’max_depth’: None,
’min_samples_leaf’: 2,
’min_samples_split’: 5,
’n_estimators’: 25
}
```

Water Potability Dataset:

```
{
’criterion’: ’gini’,
’max_depth’: 15,
’min_samples_leaf’: 1,
’min_samples_split’: 2,
’n_estimators’: 100
}
```

Logistic Regression:

Heart Attack Dataset:

```
{
’C’: 10,
’max_iter’: 500,
’penalty’: ’l1’,
’solver’: ’liblinear’
}
```

Stroke Dataset:

```
{
'C': 0.1,
'max_iter': 500,
'penalty': 'l2',
'solver': 'newton-cg'
}
```

Water Potability Dataset:

```
{
'C': 0.1,
'max_iter': 500,
'penalty': 'l2',
'solver': 'newton-cg'
}
```

k-Nearest Neighbors (kNN):

Heart Attack Dataset:

```
{
'algorithm': 'auto',
'n_neighbors': 8,
'p': 1,
'weights': 'distance'
}
```

Stroke Dataset:

```
{
'algorithm': 'auto',
'n_neighbors': 9,
'p': 2,
'weights': 'uniform'
}
```

Water Potability Dataset:

```
{'algorithm': 'auto',
'n_neighbors': 15,
'p': 1,
'weights': 'distance'
}
```

These optimal configurations highlight the significance of tailored parameter tuning for achieving superior performance in classification tasks across diverse datasets. Together both the validation and test results for all 3 of the datasets across all models can be shown:

Table 1: Standard Models Performance Metrics

Dataset	**Metric**	**Decision Tree**	**Random Forest**	**Logistic Regression**	**k-NN**
Heart Attack Dataset	CV Accuracy	0.756	0.805	0.822	0.839
	CV F1 Score	0.758	0.805	0.821	0.836
	Test Accuracy	0.852	0.852	0.885	0.885
	Test F1 Score	0.852	0.853	0.885	0.885
Stroke Dataset	CV Accuracy	0.942	0.954	0.955	0.955
	CV F1 Score	0.929	0.932	0.933	0.933
	Test Accuracy	0.931	0.939	0.939	0.94
	Test F1 Score	0.912	0.91	0.91	0.914
Water Potability Dataset	CV Accuracy	0.626	0.669	0.606	0.655
	CV F1 Score	0.595	0.63	0.459	0.621
	Test Accuracy	0.61	0.689	0.628	0.646
	Test F1 Score	0.594	0.661	0.485	0.615

2.3 Evaluation of FCA Binarized Binary Classification Model

Before using the classification model, the datasets must first be entirely binarized (wherever applicable) in order to be able to work with FCA. For that a copy is created and binarization is applied on it.

2.3.1 Heart Attack Dataset

- Binarizing age based on position around the mean
- Binarizing chest pain type (cp) whether or not they belong to types (0) (heart related) or (1,2,3) (non heart related)
- Binarizing resting blood pressure (trtbps) based on its position around normal blood pressure of humans being 120 mm/Hg
- Binarizing cholesterol (chol) based on its position around normal cholesterol level which is under 200 mg/dL
- Binarizing resting electrocardiographic results (restecg) based on whether it is normal (1) or not (0,2)

- Binarizing maximum heart rate achieved (thalachh) based on theoretical max heart rate calculated by (220-age)
- Binarizing oldpeak based on position around the mean
- Binarizing slope based on flat/positive (1,2) slope and negative slope (0)
- Binarizing number of major vessels based on whether there are any or not
- Binarizing thalassemia based on whether there is normal blood-flow (2) or not (0,1,3)

2.3.2 Stroke Dataset

- Binarizing age based on position around the mean
- Binarizing average glucose level based on position around the mean
- Binarizing bmi around the max normal limit 24.9
- Utilizing only first 2000 rows for classification due to large size of dataset and time constraints.

2.3.3 Water Potability Dataset

- Binarizing pH based on pH range of drinking water between 6.5 and 8.5
- Binarizing all other values around the mean of their columns
- Utilizing rows from 500 to 4000 for classification due to large size of dataset and time constraints.

2.3.4 Classification

The Heart Attack dataset perform the best on method='standard' while the Stroke and Water Potability datasets perform best on method='ratio-support'. It being the only hyper-parameter that can be modified. The metrics are as follows:

Table 2: FCA Binarized Binary Classifier Performance Metrics

Dataset	**Metric**	**Classifier**
Heart Attack Dataset	CV Accuracy	0.558
	CV F1 Score	0.527
	Test Accuracy	0.525
	Test F1 Score	0.407
Stroke Dataset	CV Accuracy	0.873
	CV F1 Score	0.815
	Test Accuracy	0.882
	Test F1 Score	0.827
Water Potability Dataset	CV Accuracy	0.583
	CV F1 Score	0.445
	Test Accuracy	0.601
	Test F1 Score	0.463

2.4 Evaluation of FCA Pattern Classification Model

For this classification model, the original datasets were used, however for Stroke Dataset: first 1000 rows are used while, Water Potability: 1000 rows from rows 500 to 1500 are used due to large size of dataset and time constraints. The datasets are passed along with an array of numbers corresponding to those columns in the dataset which are categorical in nature. The Heart Attack dataset works best with method='ratio-support' while Stroke dataset works best with method='standard-support' and Water Potability on method='standard'.

Table 3: FCA Pattern Classifier Performance Metrics

Dataset	**Metric**	**Classifier**
Heart Attack Dataset	CV Accuracy	0.773
	CV F1 Score	0.768
	Test Accuracy	0.852
	Test F1 Score	0.853
Stroke Dataset	CV Accuracy	0.747
	CV F1 Score	0.763
	Test Accuracy	0.685
	Test F1 Score	0.711
Water Potability Dataset	CV Accuracy	0.624
	CV F1 Score	0.537
	Test Accuracy	0.58
	Test F1 Score	0.461

3 Conclusion

Overall, the FCA based classifiers perform slightly worse than the regular standard models, though being close to Decision Tree. The Pattern Classifier however performs better than the Binarized Binary Classifier across all three datasets.