Big Homework

Farhana Noor

December 5, 2023

1 Introduction

The report explores the application of ordered sets in data analysis using various datasets and machine learning models. It focuses on the analysis of three commonly used datasets: Diabetes, Iris, and Breast Cancer. The utilization of machine learning models—K Nearest Neighbour (KNN), Decision Tree, Random Forest, Logistic Regression, and Lazy FCA Model—is investigated to determine their efficacy in predictive analysis.

2 Datasets

2.1 Diabetes Dataset

- **Reference**: [https://archive.ics.uci.edu/dataset/34/diabetes]
- **Features**:
  - Pregnancies
  - Glucose
  - BloodPressure
  - SkinThickness
  - Insulin
  - BMI
  - DiabetesPedigreeFunction
  - Age
- **Objects**: 768 instances of diabetic and non-diabetic patients.

2.2 Iris Dataset

- **Reference**: [https://archive.ics.uci.edu/dataset/53/iris]
- **Features**:

  - Sepal Length
  - Sepal Width
  - Petal Length
  - Petal Width
- **Objects**: 150 instances representing iris flowers categorized into three species.

2.3 Breast Cancer Dataset

- **Reference**: [https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic]
- **Features**:
  - Mean Radius
  - Mean Texture
  - Mean Perimeter
  - Mean Area
  - Mean Smoothness
  - Mean Compactness
  - Mean Concavity
  - Mean Concave Points
  - Mean Symmetry
  - Mean Fractal Dimension
- **Objects**: 569 instances of Malignant and Benign samples.

3 Machine Learning Models Used

3.1 K Nearest Neighbours (KNN)

- Configured the KNN algorithm with varying 'k' values to determine the optimal value. - Trained the model on the training set and evaluated its performance on the test set.

3.2 Decision Tree

- Constructed decision trees with different criteria (e.g., Gini impurity, entropy) to analyze their impact on model performance. - Fine-tuned hyperparameters like maximum depth or minimum samples per leaf to optimize the tree structure.

3.3 Random Forest

- Implemented Random Forest ensembles with different numbers of trees and analyzed their impact on model accuracy and F1 score. - Conducted feature importance analysis within the ensemble.

3.4 Logistic Regression

- Utilized Logistic Regression to model the relationship between the independent variables and diabetes occurrence. - Fine-tuned regularization parameters (e.g., L1 or L2 penalties) to prevent overfitting.

3.5 Lazy FCA Model

- Implemented the Lazy FCA model to analyze the ordered sets representation of the diabetes dataset. - Evaluated performance metrics such as accuracy and F1 score using this model.

4 Methodology

4.1 Data Preprocessing:

4.1.1 Data Binarization

4.2 Model Selection and Training:

4.2.1 K-fold Cross Validation

4.2.2 Splitting Data:

Divide datasets into training and testing subsets (e.g., 70-30 ratio).

4.2.3 Model Implementation:

Apply KNN, decision tree, Random Forest, logistic regression, and lazy FCA on the datasets.

4.2.4 Model Training:

Train each model on the training data using appropriate parameters and techniques.

4.2.5 Model Evaluation:

Evaluate models' performance using accuracy and F1-score.

4.3 Results Analysis:

4.3.1 Performance Metrics:

Calculate and compare accuracy and F1-score for each model.

4.3.2 Tabulate Results:

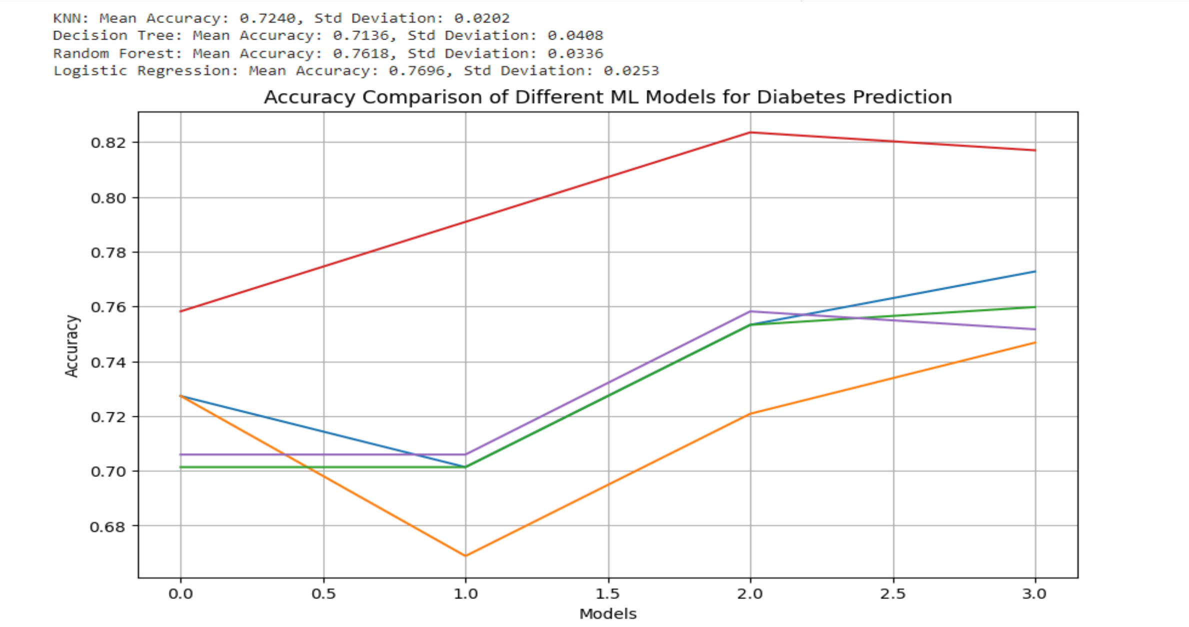
Create a table summarizing the performance metrics of all models across datasets.

Table 1: Accuracies and F1 Scores of Machine Learning Models on Datasets

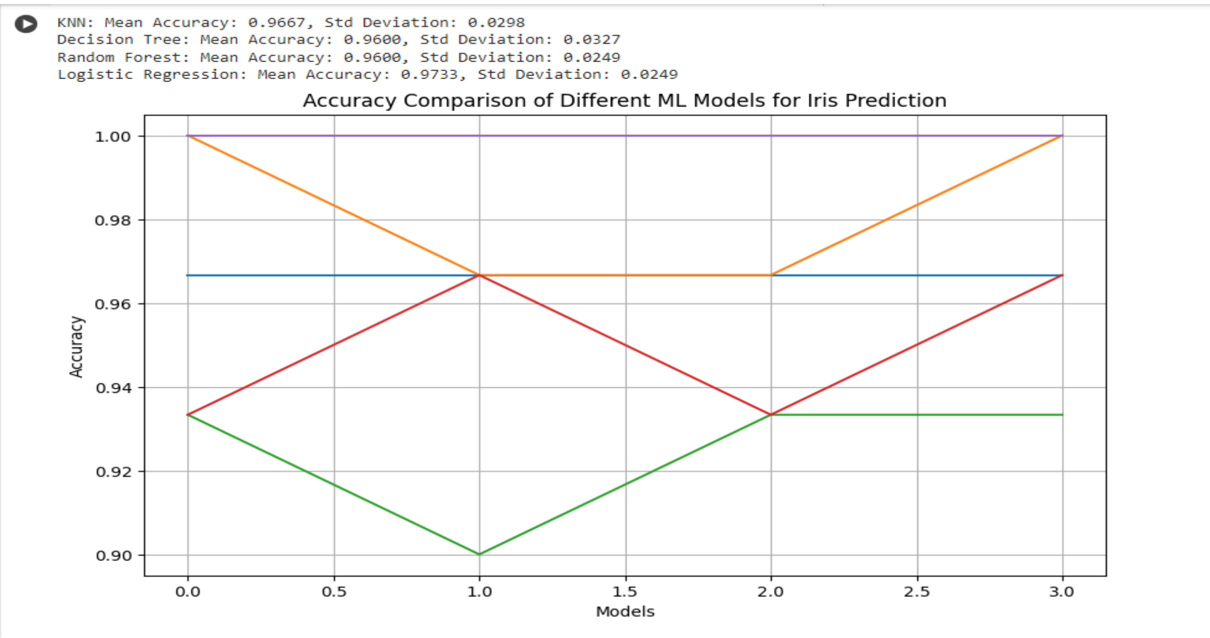
Dataset	Model	Performance	
		Accuracy (%)	F1 Score
Diabetes	KNN	72	71
	Decision Tree	67	65
	Random Forest	74	73
	Logistic Regression	75	74
	Lazy FCA Model		
Iris	KNN	96	96
	Decision Tree	93	93
	Random Forest	93	92
	Logistic Regression	94	94
	Lazy FCA Model		
Breast Cancer	KNN	95	95
	Decision Tree	90	92
	Random Forest	95	96
	Logistic Regression	97	97
	Lazy FCA Model		

4.3.3 Visual Representation:

Generate plots, illustrating the accuracies of different models on each dataset.

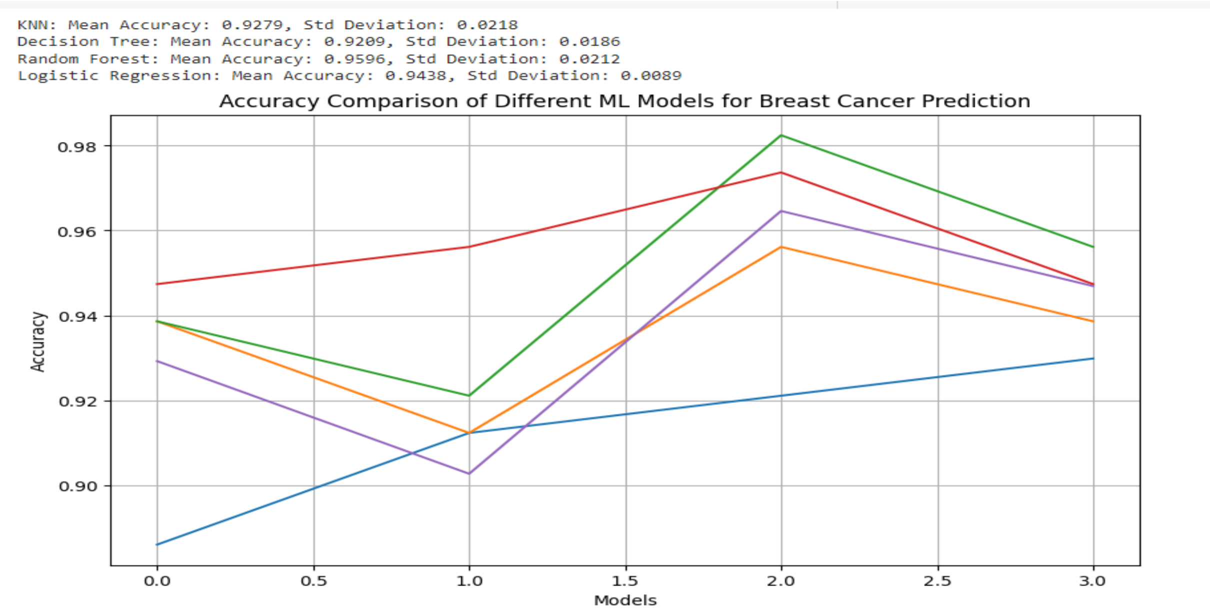


```
KNN: Mean Accuracy: 0.7240, Std Deviation: 0.0202
Decision Tree: Mean Accuracy: 0.7136, Std Deviation: 0.0408
Random Forest: Mean Accuracy: 0.7618, Std Deviation: 0.0336
Logistic Regression: Mean Accuracy: 0.7696, Std Deviation: 0.0253
```



```
KNN: Mean Accuracy: 0.9667, Std Deviation: 0.0298
Decision Tree: Mean Accuracy: 0.9600, Std Deviation: 0.0327
Random Forest: Mean Accuracy: 0.9600, Std Deviation: 0.0249
Logistic Regression: Mean Accuracy: 0.9733, Std Deviation: 0.0249
```

Accuracy Comparison of Different ML Models for Iris Prediction
Accuracy
1.00
0.98
0.96
0.94
0.92
0.90
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Models



```
KNN: Mean Accuracy: 0.9279, Std Deviation: 0.0218
Decision Tree: Mean Accuracy: 0.9209, Std Deviation: 0.0186
Random Forest: Mean Accuracy: 0.9596, Std Deviation: 0.0212
Logistic Regression: Mean Accuracy: 0.9438, Std Deviation: 0.0089
```

Accuracy Comparison of Different ML Models for Breast Cancer Prediction
Accuracy
0.98
0.96
0.94
0.92
0.90
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Models