**CLASSIFICATION OF PUMPKIN SEEDS DATASET USING MACHINE LEARNING TECHNIQUES**

By: Meghana Gamidi

**ABSTRACT:** In this paper knowledge discovery in databases (KDD) techniques are applied for decision support in assessing pumpkin seeds. We utilize genetic optimization to try and balance interpretability and accuracy in assessing. Using data mining software, we compare multiple models to determine which is the most accurate and appropriate, so carefully statistically validating our proposed approach. To guarantee the integrity of the data, this process starts with a comprehensive data cleansing step and further aims to validate and compare statistical models to identify the most effective predictor of pumpkin seed quality.

# INTRODUCTION

Data mining is a technique for solving problems by analyzing existing data stored in databases. Knowledge Discovery in Databases (KDD) is a systematic approach for automatically analyzing and modeling large datasets. It involves developing algorithms that examine the data, construct models, and identify previously undetected patterns. These models are then employed for data analysis, prediction and understanding various phenomena. To extract patterns and discover knowledge that could be useful for future decision-making, a variety of algorithms have been developed and implemented. Common data mining techniques include data preparation, pattern recognition, clustering, and classification.

Pumpkin seeds are a popular worldwide due to their balanced nutritional profile, containing carbohydrates, fats, proteins, and various essential minerals. Their composition includes 37% carbohydrates, 35-40% fats and proteins, and a rich mineral profile with calcium, potassium, phosphorus, magnesium, iron, and zinc. In our study we used 2500 records of Pumpkin seeds dataset published and accessible through Kaggle's repository. We mainly focus on the two most prominent and high-quality pumpkin seed varieties ‘Urgup\_Sivrisi' and 'Cercevelik,' to develop predictive models for seed quality assessment. Our journey begins with meticulous data cleansing to ensure the integrity and reliability of our analysis.

In this study we use Classification algorithms a supervised learning where the objective is to specify a target class or attribute which aim’s to estimate the value of a single class based on other attributes. Clustering algorithms is an unsupervised learning, attempt to separate the data into meaningful subsets closely related to one another. Association algorithms is an unsupervised learning which do not target a single attribute and instead tries to find relations among multiple attributes.

Classification is one of the essential jobs in data mining. In this study we applied various classification algorithms, including Navie Bayesian, Logistic Regression, Multilayer Perceptron, Stochastic Gradient Descent, KNN classifier, Decision Table, ZeroR, Decision Stump, J48, Random Forest, Random tree and REP to the given dataset and analyze accuracy, precision, mean absolute error, root mean squared error and relative absolute error to provide best classification algorithm. In clustering we used simple k Means and EM algorithms. The K-Means algorithm is a clustering algorithm that aims to partition a dataset into K clusters, where each data point belongs to the cluster with the nearest mean. The EM algorithm is a general framework for finding maximum likelihood estimates of parameters in models with latent variables. It is commonly used for clustering with Gaussian Mixture Model.

1. **DATASET**

**Data Set Understanding**: Identifying the challenges, we will face and having a context to suggest feasible and practical solutions are essential. Understanding the characteristics, limitations and restrictions of the data and obtain the desired outcomes.

**Data Selection**: From the available data, a relevant subset is chosen that aligns with the analysis objectives. This involves integrating data from various sources into a unified dataset that supports the research goals. The selected data should beconsistent and relevant to the problem at hand.

**Cleaning and Pre-Processing**: At this stage, the data's quality is assessed by performing operations that guarantee its usability. Data cleaning involves handling

missing values, removing outliers, and checking for inconsistencies. It may also involve eliminating variables or attributes with missing data, as well as non-relevant material such as text or images.

**Data Transformation**: In this stage, data quality is enhanced through transformations such as dimensionality reduction, which involves reducing the number of variables in the dataset.

**Choosing Data Mining Algorithms:** The appropriate algorithm or combination of algorithms is selected to extract patterns and uncover knowledge. Different selection strategies exist for each technique. Each algorithm has its own unique characteristics, methodology and approach to producing results. It is therefore essential to understand the properties of the algorithms and choose the one best suited to the data.

**Applying Data Mining Algorithms**: Once the chosen techniques have been selected, they are applied to the preprocessed data. This may involve executing the algorithms multiple times to fine-tune parameters for optimal results. The specific parameters vary depending on the selected method.

**Analyzing**: After applying the algorithms to the dataset, the generated patterns and performance are evaluated to ensure they align with the objectives set in the initial phases. A technique called cross-validation is used for this evaluation. It involves dividing the data into training and testing sets. The training set is used to create the model, while the testing set is used to assess the model's effectiveness.

**Interpretation**: If all steps have been followed correctly and the evaluation results are satisfactory, the final stage involves applying the extracted knowledge to the relevant context and addressing the identified problems. If the results are not satisfactory, it is necessary to revisit previous stages to make adjustments, starting from data selection to the evaluation phase.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.n o** | **Name** | **Type** | **Categorical Value** |
| **1** | **Area** | **Discrete** | **Numeric** |
| **2** | **Perimeter** | **Discrete** | **Numeric** |
| **3** | **Major\_Axis\_ Length** | **Discrete** | **Numeric** |
| **4** | **Minor\_Axis\_ Length** | **Discrete** | **Numeric** |
| **5** | **Convex\_Area** | **Discrete** | **Numeric** |
| **6** | **Equiv\_Diame ter** | **Discrete** | **Numeric** |
| **7** | **Eccentricity** | **Discrete** | **Numeric** |
| **8** | **Solidity** | **Discrete** | **Numeric** |
| **9** | **Extent** | **Discrete** | **Numeric** |
| **10** | **Roundness** | **Discrete** | **Numeric** |
| **11** | **Aspect\_Ratio n** | **Discrete** | **Numeric** |
| **12** | **Compactness** | **Discrete** | **Numeric** |
| **13** | **Class** | **Continous** | **Nominal**  **(UrgupSivrisi , Cercevelik)** |

In this study we use pumpkin seed dataset from Kaggle citation request from KOKLU, M., SARIGIL, S., & OZBEK, O. (2021). The use of machine learning methods in classification of pumpkin seeds (Cucurbita pepo L.). Genetic Resources and Crop Evolution, 68(7), 2713-2726. This Pumpkin\_Seeds\_Dataset.arff dataset is classified as training and testing data sets. Pumpkin\_Seeds\_Dataset.arff that contains various examples corresponding to 2500 objects and 13 columns.

* It is analyzed using cross validation of 10 &
* percentage split of 70% of the dataset

**3. WEKA TOOL IMPLEMENTATION**

WEKA stands for **Waikato Environment for Knowledge Analysis** it’s a popular tool used for data mining and machine learning algorithms developed using JAVA by the university of Waikato, New Zealand.

It contains tools for

* Data pre-processing,
* Classification,
* Clustering,
* Association rules
* Select Attribute &
* Visualization.

**4. BEST ALGORITHM**

The best algorithm depends on the specific characteristics of the data and the requirements problem. We tried multiple algorithms and compare their performance using cross-validation and percentage split evaluation metrics to determine which one works best for your particular task.

Additionally, considering the more accuracy and low mean root errors for the model.

1. **CONCLUSION**

This research investigated the effectiveness of ensemble regression models in predicting high-quality pumpkin seeds using a real-world dataset dedicated to pumpkins. The study involved comprehensive data preprocessing steps, including cleaning, outlier removal, attribute selection, and discretization, aiming to enhance the information gain and accuracy. Multiple classification methods were applied to both discretized and non-discretized datasets.

The results of the study shows that **Random Forest (RF), logistic regression and multilayer perceptron** exhibited the most favorable outcomes in terms of analysis, demonstrating higher accuracy and fewer errors. Upon scrutinizing the accuracies of the ten machine learning models, it was concluded that **Random Forest (RF)** is the best model for the "Pumpkin Seed" dataset.

1. **COMPARSION**

|  |  |  |  |
| --- | --- | --- | --- |
| S.no | Model | Cross Validation (Folds-10) | percentage split(70%) |
| 1 | Navie Bayes | Correctly Classified:86.72 % Incorrectly Classified:13.28 % Mean Absolute Error: 0.1373 Root Mean Square Error: 0.3416 Relative Absolute Error: 27.5053 | Correctly Classified:86.67 % Incorrectly Classified:12.1333 % Mean Absolute Error: 0.1254 Root Mean Square Error: 0.3237 Relative Absolute Error: 25.134 |
| 2 | logistic Regression | Correctly Classified:88.68 % Incorrectly Classified:11.32 % Mean Absolute Error: 0.1726 Root Mean Square Error: 0.2931 Relative Absolute Error: 34.5703 | Correctly Classified:90.00 % Incorrectly Classified:10.00 % Mean Absolute Error: 0.1623 Root Mean Square Error: 0.2794 Relative Absolute Error: 32.5177 |
| 3 | Multi Layer Perceptron | Correctly Classified:88.4 % Incorrectly Classified:11.6 % Mean Absolute Error: 0.1614 Root Mean Square Error: 0.2964 Relative Absolute Error: 32.3312 | Correctly Classified:88.8 % Incorrectly Classified:11.2 % Mean Absolute Error: 0.1505 Root Mean Square Error: 0.2843 Relative Absolute Error: 30.1626 |
| 4 | Stochastic Gradient Descent | Correctly Classified:86.88 % Incorrectly Classified:13.12 % Mean Absolute Error: 0.1312 Root Mean Square Error: 0.3622 Relative Absolute Error: 26.282 | Correctly Classified:88.2667 % Incorrectly Classified:11.7333 % Mean Absolute Error: 0.1173 Root Mean Square Error: 0.3425 Relative Absolute Error: 23.5103 |
| 5 | k-Nearest Neighbor (k- NN) | Correctly Classified:82.28 % Incorrectly Classified:17.72 % Mean Absolute Error:0.1775 Root Mean Square Error: 0.4208 Relative Absolute Error: 35.5542 | Correctly Classified:83.0667 % Incorrectly Classified:16.9333 % Mean Absolute Error:0.1697 Root Mean Square Error: 0.4113 Relative Absolute Error: 34.0052 |
| 6 | Decision Table | Correctly Classified:86.76 % Incorrectly Classified:13.24 % Mean Absolute Error: 0.195 Root Mean Square Error: 0.3109 Relative Absolute Error: 39.0611 | Correctly Classified:87.0667 % Incorrectly Classified:12.9333 % Mean Absolute Error: 0.1829 Root Mean Square Error: 0.2952 Relative Absolute Error: 36.646 |
| 7 | Zero R | Correctly Classified:52.48 % Incorrectly Classified:48.00 % Mean Absolute Error: 0.4992 Root Mean Square Error: 0.4996 Relative Absolute Error: 100 | Correctly Classified:52.8 % Incorrectly Classified:47.20% Mean Absolute Error: 0.4991 Root Mean Square Error: 0.4993 Relative Absolute Error: 100 |
| 8 | Decision Stump | Correctly Classified:85.6 % Incorrectly Classified:14.4 % Mean Absolute Error: 0.2371 Root Mean Square Error: 0.3488 Relative Absolute Error: 47.5055 | Correctly Classified:87.4667% Incorrectly Classified:12.5333 % Mean Absolute Error: 0.2276 Root Mean Square Error: 0.3298 Relative Absolute Error: 45.5947 |

|  |  |  |  |
| --- | --- | --- | --- |
| S.no | Model | Cross Validation (Folds-10) | percentage split(70%) |
| 9 | J48 | Correctly Classified:87.56 % Incorrectly Classified:12.44 % Mean Absolute Error: 0.1777 Root Mean Square Error: 0.3159 Relative Absolute Error: 35.6013 | Correctly Classified:87.73333 % Incorrectly Classified:12.2667 % Mean Absolute Error: 0.1667 Root Mean Square Error: 0.3109 Relative Absolute Error: 33.3979 |
| 10 | Random Forest | Correctly Classified:88.72 % Incorrectly Classified:11.28 % Mean Absolute Error: 0.1664 Root Mean Square Error: 0.2977 Relative Absolute Error: 33.65 | Correctly Classified:89.8667 % Incorrectly Classified:10.1333 % Mean Absolute Error: 0.1558 Root Mean Square Error: 0.2861 Relative Absolute Error: 31.2179 |
| 11 | Random Tree | Correctly Classified:82.92 % Incorrectly Classified:17.08 % Mean Absolute Error: 0.1708 Root Mean Square Error: 0.4133 Relative Absolute Error: 34.2147 | Correctly Classified:83.0667 % Incorrectly Classified:16.9333% Mean Absolute Error: 0.1693 Root Mean Square Error: 0.4115 Relative Absolute Error: 33.9296 |
| 12 | REP Tree | Correctly Classified:87.24 % Incorrectly Classified:12.76 % Mean Absolute Error: 0.1765 Root Mean Square Error: 0.3171 Relative Absolute Error: 35.362 | Correctly Classified: 87.4667 % Incorrectly Classified:12.5333 % Mean Absolute Error: 0.1753 Root Mean Square Error: 0.3161 Relative Absolute Error: 34.7592 |