ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP354

Agricultural Data Analysis using Machine Learning: A Study on Dry Bean Classification

Archith Shankar[1]; Arushi R Kadam[2]; Nishita Senthilkumar[3]; Shradha A Venkatachalam[4]; Shivandappa[5]; Narendra Kumar[6]
Biotechnology, R V College of Engineering Mysore Road, Bengaluru 560059

Abstract:- The classification of Dry Beans using various techniques such as Support Vector Machine (SVM) classification, K-means clustering, Decision Trees and Random Forest (RF) classification using an ipython notebook. To refine the model, performance matrix graphs of Cross entropy vs Epoch number, True value vs Predictive value and Accuracy vs Epoch. This analysis is often used in agricultural practices for improved crop management, increasing yield, resource optimization, enhancing sustainability etc.

Keywords:- *Dry Beans, Phaseolus Vulgaris L, Machine Learning, Classification Methods, KNN Cluster, Support Vector Machine.*

I. INTRODUCTION

Dry bean is a major part of the diets of many cultures and regions in the world, as it is a major source of protein and fibre. There are various types and varieties of these dry beans that are of different origins and genetic variations which are characterised by their size, dimensions and sizes.[2] In order to distinguish between the various types of dry beans that ultimately affect the value of the product and the quality of the dry beans that are sold in the marketplace, these types and varieties are crucial to the agricultural sector's categorization needs.

The importance of this process comes with the ailment of a tiring and wearing out manual process that is not only difficult but also time consuming and inefficient. Due to this issue a model for a computer based machine learning algorithm using the python language along with various other exploratory data analysis and machine learning aiding libraries. This model will aid in the process of dry bean classification allowing a more efficient classification process and an easier method. Along with these advantages, building a machine learning model also allows repeatability and consistency in the prediction as it can be implemented any number of times on various datasets with the required attributes.[3]

The dataset used for this has been characterised by a computer vision system that takes images of dry beans that are then characterised to provide attributes including area, perimeter, major axis length, minor axis length, eccentricity, solidity, roundness, etc.. These attributes are fed into machine learning models like neural networks and Linear regression to create a model that can be fed with data attributes that can effectively be able to characterise the inputs into the types of dry bean varieties. Here, there are 6 different varieties listed with all their attributes with over 1500 of each type of dry bean.which makes the data of higher standard allowing for better accuracies.

II. DATASET

This study uses machine learning to predict the 7 multivariate dry beans for which the dataset takes into account the shape, size, colour and other features. This dataset contains 13611 images which were captured in high resolution that are used to distinguish between the varieties of dry bean. It contains 16 features, four forms of shape and used 12 dimension pictures, some of the examples being area, perimeter, axis ratio, convex area etc. The seven varieties of dry beans that are used to classify in this study are Barbunya, Bombay, Cali, Sira,Horoz, Dermason and Seker.[1]

III. BACKGROUND

- *The Following are the Classification Methods, Quantitative and Qualitative Measures used in Building and Enhancing the Predictive Model:*

- **K-Means Clustering:** This is an unsupervised machine learning algorithm used to classify data into clusters having ‘k’ points as the centres. Each centre is assigned data points which are closest to them in distance and the process of assigning the centres is repeated until we get fixed points.
- **Silhouette Score-** This is a measure of the optimal number of clusters for a data sample that is already divided into ‘k’ clusters.
- **Decision Tree Classification**: Decision trees are a tool that is used to classify data by splitting a dataset according to measures such as Gini impurity, information gain or entropy to find the best attribute. This process is repeated for each subset until an outcome is obtained.
- **Random Forest (RF)**: RF is a classification method that aggregate the outputs of parallelly arranged decision trees, known as forests, to identify the most popular output. It often uses a technique known as bootstrap aggregation (bagging) i.e, a method that uses the average of predictions

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP354

generated by independently trained data samples to provide a more accurate result [4].

- **Artificial Neural Network (ANN)**: ANN is a computing model which mimics the neural system of the brain. Artificial neurons are arranged in different layers, and each of these layers are connected, often through a numerical relation by an 'activation function'. Input layers pass by the different neuronal layers and their connections to give the desired output signal.
- **Cross Entropy**: This is a measure of the difference between the predicted probability and the actual probability of a sample belonging to a particular class. The value must be close to 0 to generate a model with minimum error.
- **Linear Regression:** It is a machine learning algorithm that uses the relationship between a dependent and an independent variable for prediction values of a data sample.
- **Epoch:** This is one complete pass of a training dataset through an algorithm. Through each epoch, the internal parameters are updated. The epoch number is often more than 1 to reduce errors and increase the accuracy.
- **SVM Classification**: Support Vector Machine (SVM) Classification uses hyper-planes to separate data points in an N-dimensional space. It uses the hyper-plane that can maximise the distance between the closest data points of different classes to classify the data [5].
- **Confusion matrix:** It is a matrix that measures the number of times the model correctly and incorrectly identifies the instances of a class.
- **Accuracy:** It is the measurement of the closeness between the predicted results and the actual results of a training dataset.
- **Precision:** It is the measurement of the accuracy of positive (correct) predictions.
- **Recall:** It is the measurement of the ability of a model to identify relevant cases in a dataset.
- **F1 Score**: It is the measurement of harmonic mean of precision and recall.

IV. METHODOLOGY

A. Data Collection and Preprocessing :

The data used for a study is characterised based on its volume, veracity, velocity and variety, which talk about the, sheer size of the data used, the accuracy and the the reliability of the data, the speed at which it is gathered, and the diversity of the data's sources and types, in that order. To study data for dry bean in the agricultural field, a dataset on dry bean from the UC Irvine Machine learning repository.

The raw data acquired from the repositories are not always readily usable, they contain quality issues including null values, outliers, inconsistencies, etc. These quality issues in this dataset have been eliminated by using data cleaning methods, feature scaling, etc.

B. Feature analysis and Model selection

The features in the dataset must be evaluated to see if any of them might have a bias due to reasons that are not required. Detecting such types of features will, on their neglect, allow the model to focus more on the features that are seen to show a recurring relation with the target variable.

➢ *Model Selection:*

The kind of data being evaluated and the desired output are taken into consideration while choosing the model. Here, the models used are:

- Random Forest Classifiers: The algorithm is for classification tasks which uses multiple decision trees to reduce overfitting which increases the precision and quality of the predictions.
- Decision Tree Classifiers: The algorithm is used to take into consideration each attribute of the data and identify the most suitable pathway for decision making.
- Artificial Neural Network (ANN): Artificial Neural Networks are machine learning models that are highly complex. They are flexible and model non-linear relationships by calculating different weights and biases specific to the model that is executed.
- Support Vector Machine (SVM): SVMs are algorithms that work by mapping, separating, transforming and then classifying the data. They are employed in the process of ultimately identifying the best hyperplane to divide the classes that have been found.
- Models of linear regression: In order to identify the features that work better in prediction models, linear regression is a mathematical model that is used to discover relationships between the various qualities in a dataset or between the target variable and the attributes.

C. Implementation:

Implementation of the models is done using the python language in an ipython notebook using Google Colab, which is a cloud based service that can run on any browser providing cloud-based ram.

➢ *The Dataset is Uploaded by First Making it Accessible through Google Drive. the Different Tools Used Here for Data Attribution and Pipeline Include:*

- Pandas: Data analysis and manipulation
- Scikit-learn: Allows the use of all the machine learning models that are apt for the dataset
- Matplotlib: Data Visualisation
- Tensor Flow: Building and training for neural networks

The different libraries are imported into the notebook along with the specific sub libraries that are essential in implementing the processes to split the data, develop the model and train the model. Visualisation of the model includes that of cross-entropy loss vs. epoch graphs, true vs. predicted values

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP354

graphs, accuracy vs. epoch graphs and confusion matrices, these help in the detection of the accuracy of predictions and the efficacy of the same.

The accuracy of the predictions are not only seen by the above visualisations but also by using the random forest classifiers that give a quantitative measure of the accuracy quotient

V. RESULT

A. Classification Report

➢ *Classification Report Using Random Forest Algorithm*

The accuracy score of 92.62% suggests that many percentage of test data points were correctly plotted.The machine learning model was able to classify the type of dry beans: barbunya, sira, horoz, dermason, cali, bombay, and seker.The precision, recall, f1 score were calculated in this classification report.

Table 1: Classification Report Using Random Forest Algorithm

Accuracy: 92.62%
Classification Report:

	precision	recall	f1-score	support
BARBUNYA	0.93	0.91	0.92	261
BOMBAY	1.00	1.00	1.00	117
CALI	0.92	0.94	0.93	317
DERMASON	0.90	0.92	0.91	671
HOROZ	0.97	0.95	0.96	408
SEKER	0.97	0.94	0.95	413
SIRA	0.87	0.88	0.88	536
accuracy			0.93	2723
macro avg	0.94	0.94	0.94	2723
weighted avg	0.93	0.93	0.93	2723

➢ *Classification Report Using Decision Tree*

The accuracy score of 89.28% suggests that many percentage of test data points were correctly plotted.The machine learning model was able to classify the type of dry beans: barbunya, sira, horoz, dermason, cali, bombay, and seker.The precision, recall, f1 score were calculated in this classification report.

Table 2: Classification Report Using Decision Tree

Accuracy: 89.28%
Classification Report:

	precision	recall	f1-score	support
BARBUNYA	0.86	0.88	0.87	261
BOMBAY	1.00	1.00	1.00	117
CALI	0.89	0.91	0.90	317
DERMASON	0.88	0.88	0.88	671
HOROZ	0.95	0.93	0.94	408
SEKER	0.93	0.91	0.92	413
SIRA	0.83	0.83	0.83	536
accuracy			0.89	2723
macro avg	0.91	0.91	0.91	2723
weighted avg	0.89	0.89	0.89	2723

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP354

B. KNN Clustering

➢ *Elbow Method for Optimal K*

Inertia is a measure of how well the K-Means algorithm clusters the data. It calculates the sum of squared distances between each data point and its assigned cluster center.Lower inertia suggests tighter clusters.As k increases, inertia decreases due to the fact that adding more clusters allows data points to be closer to their cluster centers.The sudden decrease of inertia from k=2 to k=3 (in Fig 1) suggests that adding a third cluster significantly improves the clustering.The rate of inertia decreases slowly from k=3 to k=4 thus suggesting that adding another cluster does not significantly impact the clustering . Thus k=3 or k=4 is considered to be the ideal number of clusters.

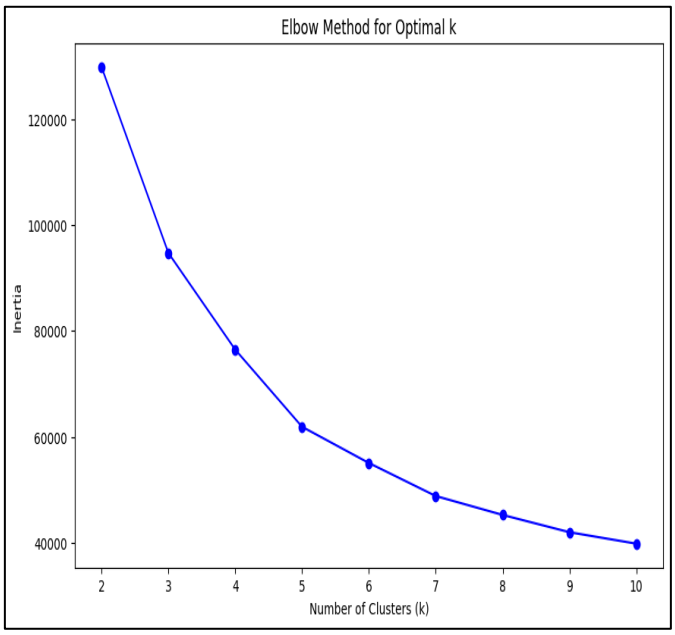

Fig 1: Elbow Method for Optimal K

- The X-axis represents the number of clusters k being tested and the Y-axis shows the inertia

➢ *Silhouette Score For Optimal K*

Silhouette score plot interprets the relationship between the number of clusters and the silhouette score.At k=3, the silhouette score is found to be the highest around 0.40 (in Fig 2). This suggests that 3 clusters provide the best separation of data.The decline in silhoutte score suggests that numbers beyond that point result in less distinct grouping.

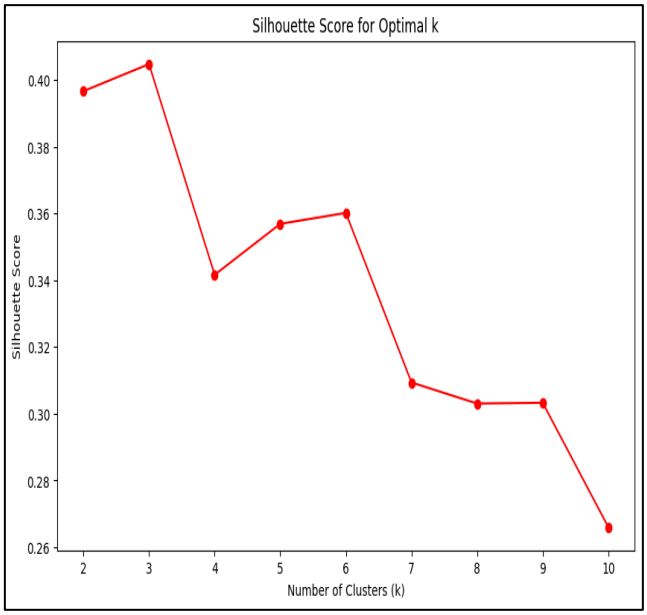

Fig 2: Silhouette Score For Optimal K

C. Confusion Matrix for SVM Classifier

The confusion matrix interprets the performance of an Support Vector Machine (SVM) classifier on a multiclass classification.The rows correspond to the actual class whereas the columns correspond to the actual predicted class.The diagonal values indicate the correctly calculated values whereas the other elements indicate miscalculations.Since the dataset was evenly split among the 7 varieties, Dermason bean has the highest correctly classified elements indicating the models proficiency. The highest number of miscalculations are that of the dermason as sira indicating some similarity in the features of these beans. Thus this matrix helps us identify areas of improvement.

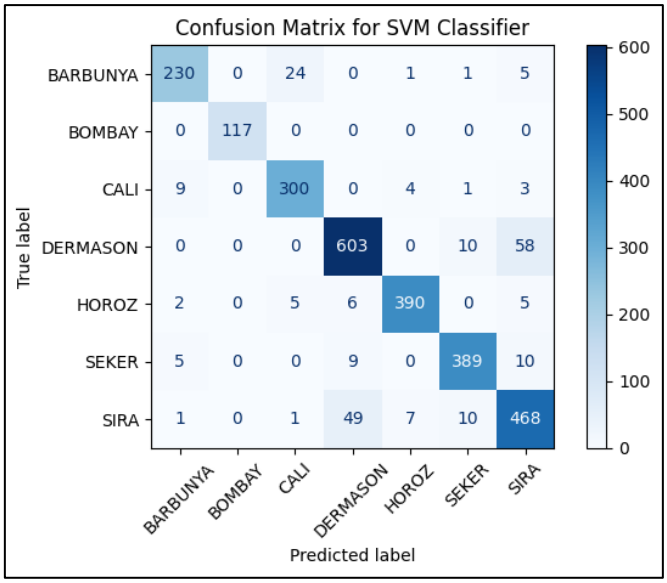

Fig 3: Confusion Matrix For SVM Classifier

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP354

D. Performance Metrics

➢ *Cross-Entropy Vs Epochs*

The training loss is represented by the blue line whereas the yellow line represents the validation loss .The training loss is found to be highest at the first epoch (in Fig 3). After about 10 epoch cycles, the training loss stabilizes.The validation loss also starts out high but stabilises over time. After about 30 epochs, the training loss loosely aligns with the validation loss.The validation loss indicates that the model struggles with some aspects of generalisation but performs better after more epoch cycles.

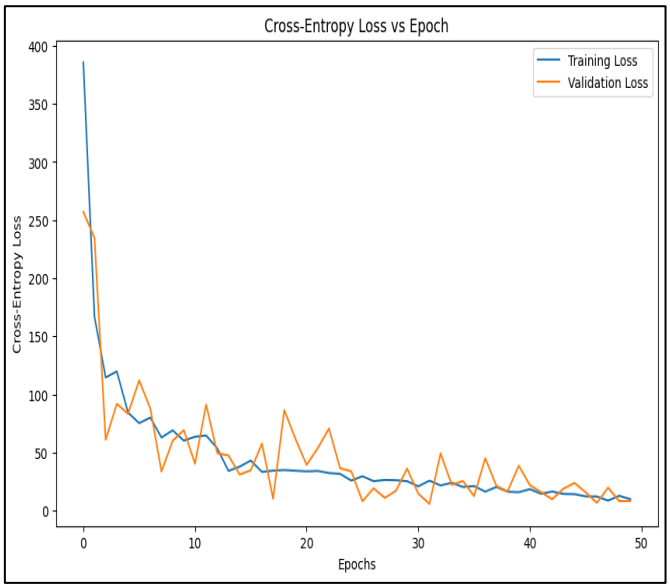

Fig 4: Cross-Entropy Vs Epoch Graph

➢ *Accuracy Vs Epoch Graph*

The blue dotted represents the training accuracy whereas the validation accuracy is represented by the orange line(in Fig 4).The training accuracy starts at a low point of 0.2 but steadily increases with the number of epochs.Similarly the validation accuracy starts low and increases with the number of epoch cycles peaking at epoch cycle 31 with the validation accuracy of 0.7613.After this epoch both the validation accuracy and the training accuracy decline due to overfitting or learning saturation.

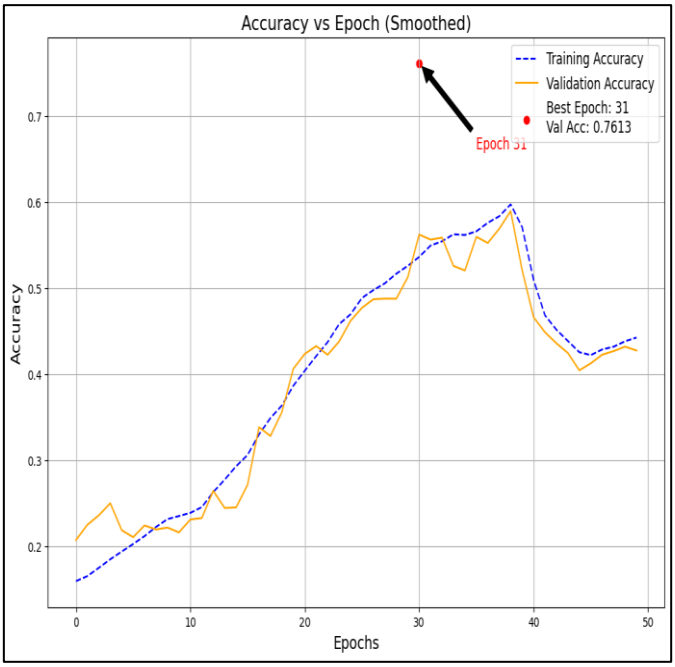

Fig 5: Accuracy Vs Epoch

➢ *True Values Vs Predicted Values*

The red dashed line indicates the perfect prediction lines whereas the blue dots represent the predicted values plotted against true values.(in Fig 5) The close alignment of the two indicates high accuracy and low error . There don't appear to be any significant outliers or large deviations which act in favour of model accuracy.

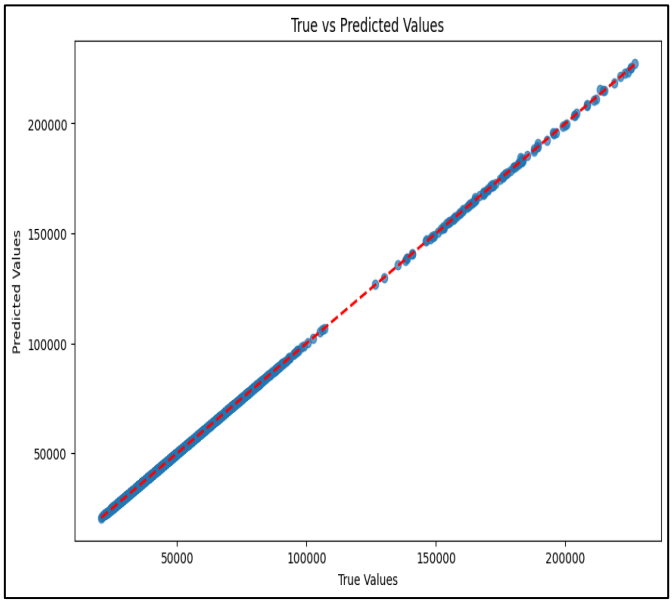

Fig 6: True Vs Predicted Values

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP354

VI. CONCLUSION

The accuracy score classification report of the random tree classifier is more than the accuracy score of the decision tree classifier as expected. Through unsupervised KNN clustering models, we identified that 3 clusters provide best separation of data. From the confusion matrix for SVM classifier we interpret that dermason has the highest correctly classified elements, Bombay has the least amount of misinterpretations,highest number of miscalculations are that of the dermason as sira indicating some similarity in the features of these beans.The cross-entropy vs epoch graph indicates that after 30 epochs the training and validation losses align.The accuracy vs epoch graphs helps us interpret that 31 cycles are optimal before overfitting of model starts to occur.Since the true values and predicted values align ideally and there are no outliers, we can conclude that the model is of high accuracy.

REFERENCES

[1]. Dry Bean [Dataset]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C50S4B.
[2]. Geng, Y., Li, Q., Yang, G., Qiu, W. (2024). Logistic Regression. In: Practical Machine Learning Illustrated with KNIME. Springer, Singapore. https://doi.org/10.1007/978-981-97-3954-7_4
[3]. Cantemir, E., Kandemir, O. Use of artificial neural networks in architecture: determining the architectural style of a building with a convolutional neural networks. Neural Comput & Applic 36, 6195–6207 (2024). https://doi.org/10.1007/s00521-023-09395-y
[4]. Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. Briefings in bioinformatics, 24(2), bbad002. https://doi.org/10.1093/bib/bbad002
[5]. Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine(SVM) Learning in Cancer Genomics. Cancer genomics & proteomics, 15(1), 41–51. https://doi.org/10.21873/cgp.20063