**US WHEAT CLASSIFICATION**

A Project Report Submitted in fulfillment on the Requirement for the Data Science Program.

Post-Graduation (Data Science)

By

**Aniket Chavan**

Under the estimate guidance of

Mr. RITIK RAJ VAISHYA

Trainer



IMARTICUS LEARNING

Thane, MAHARASHTRA

2023-24

1. **Introduction**

Wheat is known to be one of the most important agricultural crops throughout the world. Due to its mass storage in warehouses, tonnes of wheat grains rotten every year that eventually affects its market price. This paper presents an end-to-end automatic system that utilizes computer vision techniques for the quality grading of wheat grains. The main purpose of this work is to determine most discriminatory features and a suitable classifier that may classify the given wheat sample into two classes ‘fresh’ and ‘rotten’. At first, shadow removal, segmentation, and separation of each grain is performed as parts of the pre-processing step. The pre-processing is followed by the features extraction step where 7 color and 16 texture handcrafted features are determined for each grain. The four binary classification models, namely, [Support Vector Machine](https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/support-vector-machine) (SVM), K-Nearest Neighbour (KNN), Multi-Layer Perceptron (MLP), and Naïve Bayes (NB) are then built using 10-fold cross validation approach. The classifiers are compared on the basis of performance metrics- accuracy, error rate, recall, specificity, precision, and F1-score. The comparative analysis depicts that based on color features, the SVM classifier outperforms other classifiers by achieving the accuracy of 93%. In contrast, based on texture features, the NB classifier achieved accuracy at 65%; highest among all classifiers. Experimental results encourage the utility of SVM classifier modelled on color features in automatic quality grading of wheat grains.

**Objective:**

1. Predictive Modeling for Surgical Complications:

* Objective: Develop a logistic regression model of us wheat classification.
* Rationale: To identify and quantify the impact of various wheat classification.
* Objective: Utilize logistic regression to stratify patients into risk categories for surgical interventions.

3.Identification of Predictive Factors:

* Objective: Identifying predictive factors in wheat classification involves determining which features or characteristics of wheat contribute significantly to distinguishing between different classes.
* Rationale: To gain insights into the most influential variables affecting wheat , informing future interventions and highlighting areas for improvement.
* Scope of project :The scope of a wheat classification project in machine learning is broad and can encompass various aspects related to the identification, categorization, and understanding of different wheat varieties or classes.

5. Decision-making :

* Objective: Develop a logistic regression model to estimate the probability of achieving specific outcomes and accuracy about the wheat classification .

6. Quality Improvement Initiatives:

* Quality improvement initiatives in wheat classification involve implementing strategies and practices to enhance the accuracy, efficiency, and reliability of the classification process.
* Here are some key initiatives that can contribute to the improvement of wheat classification:

These objectives align with the application of logistic regression in the wheat classification domain, where the aim is often to model and predict binary outcomes or probabilities classifications of wheat and their associated outcomes

1. **Work Description**

Machine learning project on wheat classification involves various tasks and responsibilities aimed at developing a model capable of accurately classifying different types of wheat. Here's a work description outlining key steps in such a project:

1.**Project Objectives:** Clearly articulate the specific goals of the wheat classification in us project. For example, the objectives may include, the classification of different types of wheat developing a risk assessment model for a particular surgical procedure, or identifying significant predictors of surgical success.

**2.Data Collection:** Specify the types of data to be collected. This could include different wheat types , best selling wheats , quality of wheats, imaging data, and details od wheat . Emphasize the importance of obtaining a comprehensive dataset to ensure the robustness of the logistic regression model.

**3. Data Preprocessing:** Describe the steps involved in preparing the collected data for analysis. This may encompass handling missing values, standardizing variables, and encoding categorical variables. Emphasize the significance of data quality in the accuracy of the logistic regression model.

**4. Variable Selection:** Explain the rationale behind selecting specific variables for inclusion in the logistic regression model. Considers different types of wheat.

**9.Limitations and Future Directions:** Acknowledge the limitations of the logistic regression model and propose avenues for future research. This may include exploring additional variables, validating the model on external datasets, or integrating real-time data for dynamic risk assessment of the different types of wheat present in us.

1. **EXPLORATORY DATA ANALYSIS ( EDA )**

Exploratory Data Analysis (EDA) is a crucial step in the machine learning pipeline, including tasks like wheat classification. EDA involves examining and understanding the characteristics and patterns within the dataset before building a model. Here's a guide for conducting EDA in the context of wheat classification:

**1 LOAD AND ISPECT THE DATA :**

* Load the dataset containing information about different wheat samples.
* Check the first few rows of the dataset to get a sense of its structure and format.

**2 SUMMARY STATISTICS :**

* Compute summary statistics to get an overview of the central tendency, dispersion, and shape of the dataset.

**3 DATA TYPES AND MISSING VALUES**Check the data types of each column and identify any missing values.

**4 CLASS DISTRIBUTION :**Understand the distribution of wheat classes to ensure a balanced dataset.

**5 FEATURE DISTRIBUTION :**Visualize the distributions of individual features to identify patterns and potential outliers.

**6 BOX PLOT FOR OUTLIERS:**

* Use box plots to identify potential outliers in the dataset

**4 DATA SPLITTING.**

**`1 TRAINING DATASET :**

In machine learning, a training dataset is a subset of a dataset used to train a model. It consists of input-output pairs, where the input represents the features or attributes of the data, and the output is the target variable or the label that the model aims to predict

**2 TESTING DATASET :**

A testing dataset, also known as a validation or evaluation dataset, is a separate subset of the data that is not used during the training phase. The primary purpose of a testing dataset is to evaluate the performance of the trained model on new, unseen data. The model's generalization ability and accuracy are assessed by making predictions on this independent dataset.

**5 LOGISTIC REGRESSION MODEL.**

**1 MODEL SELECTION :**

Model selection is a critical step in the machine learning pipeline, involving the choice of an appropriate algorithm or model architecture for a specific task. The goal is to select a model that performs well on the given dataset and is suitable for the problem at hand. Here are key considerations and steps for model selection in machine learning:

**2 MODEL TRAINING :**

Choose an appropriate machine learning algorithm based on the nature of your problem (e.g., linear regression, decision trees, support vector machines, neural networks). Split the dataset into training and testing sets. The training set is used to train the model, and the testing set is reserved for evaluating its performance.

**3 HPPER PARAMETER TUNNING :**

Hyperparameters are configuration settings for a model. Examples include learning rates, regularization strengths, or tree depths. Specify a grid of hyperparameter values to search through during optimization. This can be done using techniques like grid search or randomized search. Evaluate the model's performance with different hyperparameter combinations using cross-validation. This helps to avoid overfitting to a specific set of hyperparameters.

**4 CROSS VALIDATION :**

Cross-validation involves splitting the dataset into multiple folds, training the model on different subsets, and evaluating its performance on the remaining data. This helps assess how well the model generalizes to unseen data. In K-Fold cross-validation, the dataset is divided into K subsets. The model is trained K times, each time using K-1 folds for training and the remaining fold for validation. Ensures that each fold maintains the proportion of classes present in the original dataset, useful for imbalanced datasets.

**6 MODEL EVALUATION**

**1 CONFUSION MATRIX :**

* A confusion matrix is a table that describes the performance of a classification model. It presents a summary of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values.
* Useful for understanding the model's behavior, especially in imbalanced datasets.

**2 ACCURACY :**

* Accuracy is a basic metric that measures the overall correctness of the model.
* Formula: Accuracy= ( TP+TN / TP + TN + FP + FN )

**3 PRECISION , RECALL , F1-SCORE :**

* Precision is the ratio of correctly predicted positive observations to the total predicted positives. It is a measure of the model's accuracy when it claims a positive prediction.
* Formula: Precision= ( TP / TP+FP )

**RECALL :**

Recall is the ratio of correctly predicted positive observations to the all observations in actual class. It is a measure of the model's ability to identify all relevant instances.

FORMULA = RECALL = ( TP / TP + FN )

**F1-SCORE :**

The F1 score is the harmonic mean of precision and recall. It balances precision and recall when one is more important than the other.

F1 SCORE FORMULA :

F1 SCORE = ( 2\* ( precision \* recall / precision+ recall ) )

**4 ROC CURVE :**

* A graphical representation of the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) at various thresholds.
* Useful for selecting the optimal threshold and understanding the model's performance across different decision boundaries

**5 MODEL STABILITY:**

* Model stability refers to how consistent a model's predictions are under different conditions or with different subsets of the data.
* Techniques such as cross-validation, bootstrapping, or stability selection can be used to assess model stability.

**7 FEATURE IMPORTANCE :**

1 COEFFICENT ANALYSIS :

Coefficient analysis in machine learning, particularly in the context of logistic regression, involves examining the coefficients assigned to each feature in the model. In logistic regression, the model predicts the probability of belonging to a certain class, and the coefficients represent the impact of each feature on the log-odds of the predicted probability.

**RESULTS :**

We got results in our wheat classification model like we considered various factors such as accuracy of the model,

Precision of the model , recall and f1 score of the model.

With all this factors we get to know about our models performance, how actually our wheat classification in us is performing , is it an overfit or underfit our best model.

In the case of overfitting model or underfitting model

Our model is overfitted as it gives us the

accuracy of 95 % , so in this case we used grid search tunning for our model.

which tunes our model by cross validation by simply cross validating the hyper parameter by selecting every possible combinations of the hyper parameters and then tunning the model which gave us the accuracy of 93 %.

**The accuracy we got before tunning was :**

|  |  |  |  |
| --- | --- | --- | --- |
| RESULTS | LOGISTIC  REGRESSION | SVM | KNN |
| ACCURACY | 93 | 95 | 95 |
| PRECISION | 86 | 87 | 87 |
| RECALL | 92 | 100 | 100 |
| F1-SCORE | 89 | 93 | 93 |

**The accuracy we got After tunning was :**

|  |  |  |  |
| --- | --- | --- | --- |
| RESULTS | LOGISTIC  REGRESSION | SVM | KNN |
| ACCURACY | 95 | 93 | 95 |
| PRECISION | 92 | 86 | 87 |
| RECALL | 92 | 92 | 100 |
| F1-SCORE | 92 | 89 | 93 |

**Model Development**

**1. Model Development:** Develop the logistic regression model using the selected technical factors as independent variables. Specify the logistic function, assess goodness-of-fit, and consider adjustments if necessary.

**2. Model Evaluation:** Evaluate the performance of the logistic regression model using appropriate metrics such as accuracy, sensitivity, specificity, and the AUC-ROC curve. Validate the model using techniques like cross-validation.

**3. Interpretation of Results:** Interpret the coefficients of the logistic regression model to understand the impact of each technical factor on the likelihood of the surgical outcome. Discuss the clinical relevance of the findings.

**4.** **Clinical Implications:** Discuss how the results of the technical survey can inform clinical decision-making. Consider the potential for modifying surgical techniques or introducing interventions to optimize outcomes based on the identified technical factors.

**5. Sensitivity Analysis:** Perform sensitivity analyses to assess the stability and robustness of the model under different scenarios or assumptions related to technical factors.

**6. Limitations and Future Directions:** Acknowledge limitations of the technical survey and propose avenues for future research. This may include exploring additional technical factors or validating the model in different surgical contexts.

**Conclusion :**

After performing different types of algorithm on the Wheat classification dataset

we could see that each algorithm's model perform differently in order to classify

the data and providing accuracy of doing the same.

from the table above we can clearly conclude that the linear regression model is

performing well on the dataset giving us accuracy of 92 %.

**Future directions :**

**Promising Results:**

Machine learning models have demonstrated high accuracy in classifying wheat types, accelerating the process and reducing errors.

**Improving Yield and Quality:**

Further research aims to utilize ML for predicting wheat yield and enhancing breeding processes for higher quality crops.

**Embracing Innovation :**

The future holds exciting potential in combining other digital technologies such as drones and IoT for precision agriculture.