SUPER STORE SALES PREDICTION

# 

# Manchem Vishnu Srikar

*Department of Computer Science and Engineering, Lovely Professional University, Jalandhar*

Punjab, India

[manchemvishnusrikar@gmail.com](mailto:manchemvishnusrikar@gmail.com)

***Abstract*—The objective is to compare and interpret the de- cisions of the algorithm to recoginise the superstore sales prediction. Some of the machine learning models are been used like Logistic Regression,Support Vector Machine(SVM),Random Forest Clas- sifier,Decision Tree, K-Nearest Neighbour (KNN), Multi Layer Perceptron (MLP), Navie Bayes, XGBoost, CAT Boost and ADA Boost are taken and compared to see the better results in recognizing the features. Hyperparameter is been used.And LIME is been implied in models for interpretability.**

***Index Terms*—Superstore sales prediction, Machine learning, k-means, Random forest, select vector machine**

1. INTRODUCTION

The Superstore Sales Prediction project is designed to help retail businesses forecast future sales trends using historical data. By analyzing a Superstore dataset, this project leverages machine learning models in Python to understand and predict sales patterns. The goal is to provide a reliable forecast model that enables effective inventory management, enhances customer satisfaction, and improves revenue projections.

To address this challenge, researchers have developed in- novative methods, feature-based character recognition, which combines the strengths of different feature extraction tech- niques to improve Accuracy and reliability of character recog- nition from palm leaf manuscripts. This combination of fea- tures from different sources and methods promises to over- come the challenges posed by palm leaf manuscripts, ensuring the preservation and accessibility of their rich textual content.

Character recognition based on superstore manuscript features leverages a diverse set of feature extraction techniques, such as geometric, textural, and structural features, to capture details complexity of each character.

Geometric features focus on the spatial arrangement of strokes and elements within characters, while texture features consider finer surface patterns and features. On the other hand, structural features analyze the relationship between different parts of a character. The next essential step is to train the recognition model. The kind of the recognition challenge and the hybrid characteristics in use determine which model should be used. This stage involves carefully labelling the data, giving each character a corresponding Malayalam symbols. By combining these feature sets, hybrid methods can handle the variability and complexity of characters on palm leaf manuscripts, handling variations due to ink fading, damage, and style. Handwritten writing thereby contributing signifi- cantly to preserve and disseminate the cultural and historical knowledge encoded in these ancient texts.

1. LITERATURE SURVEY

The relevance of optical character recognition and binariza- tion has increased significantly considering years of study. The dataset goes through preprocessing, such as noisy data removal and binarization, before using numerous methods. Certain techniques are used for feature extraction. The characters in photos of palm leaves were found using the contour approach in Reference [1]. The photos are converted to grayscale and then binarized to eliminate noise. CNN with many classes

was planned. In this study, 5-CNN layers are suggested. They achieved training accuracy of 98% and test accuracy of 96%. Reference [2] did not directly binarize the picture; instead, it employed division normalization. To reduce noise, the photos were afterwards converted to grayscale. They achieved 90% accuracy with the CNN model. The authors in [3] insights of broader recognition from palm leaves, offerings broader research about the above things more closely. Create the CNN and build the CNN network, and Determine weights and biases, train the network and evaluate designed network for identified different classes. Reference [3] suggested three steps: (I) input picture enhancement by preprocessing; (II) segmenting the characters from a palm leaf; and (III) character recognition using a Custom Multilayer CNN. For custom CNN, the accuracy difference between training and testing is around 8%. View-based mechanisms were employed in Reference. [4] to extract the characteristics from the characters on palm leaves. Artificial neural networks are used to classify extracted characters. Malayalam characters were categorized into forty- four classes with 98%accuracy. Reference [5] conducted study on a few Tamil letters. They have put out an RNN architecture. The data set was trained using the LSTM algorithm. To extract the features, they made use of OpenCV. They were able to get many Tamil characters with varying degrees of accuracy. The authors in [6] demonstrate a deep learning method for recognizing cursive letters. When wavelet, curvelet, and the Gabor transform are examined for face recognition, it is discovered that curvelet performs better than the others. Reference [7] presents mathematical transformation to extract the features from characters for recognition. The most complex deep model may take weeks to train using hundreds of machines equipped with expensive GPU and deep learning and machine learning is not easy for this classifier. Reference [8] provides a useful method for obtaining and categorizing features in the recognition of handwritten Kannada characters. The process includes the extraction of significant characteristics, such as global features produced by the wavelet transform in addition to structural information. In order to extract structural components, the pre-processed picture is divided into four quadrants. A few properties, such as corner detection, aspect ratio, and quadrant density, are then determined for each of these zones. Reference [9] It takes a three-dimensional perspective, utilizing histogram computa- tion and distance pro- file features to improve accuracy. By incorporating these techniques, the paper aims to enhance the recognition and preservation of characters within the intricate context of palm leaf manuscripts, contributing to the broader field of document analysis and cultural heritage preservation. Reference [10] this presents a feature extraction approach for offline identification of handwritten Malayalam characters predicated on obtaining the number of neighboring ones in both columns and instances. The attributes of neighboring pix- els are counted both horizontally and vertically in this manner. Following the division of the pictures into fixed sized meshes, the features are extracted. Reference [11] Ambarish Parida and Hima Bindu Maringanti’s work” Hybrid Approach for

Odia Handwritten Character Recognition” from 2014 suggests a hybrid technique for identifying handwritten Odia characters, integrating both structural and statistical aspects, in order to increase accuracy in character identification tasks. The strategy makes use of both approaches’ strong points to improve Odia script recognition.

1. METHODOLOGY
2. *Data preprocessing*

The data collection has 8147 occurrences, 51 characteristics, and 48 distinct label values.Next, min max normalization is used.It’s a widely employed method for data normalization. Every attribute is converted to a decimal between 0 and 1, with zero and one serving as the greatest and lowest values, respectively.

1. *Feature Selection*

It’s a method that uses just relevant data and removes noise from the data to lower the input variable for your model. Depending on the type of problem being addressed, it entails automatically choosing the right characteristics for a machine learning model.

1. *Classification*

The purpose of the supervised machine learning method for classification is for the model to predict an input set of data’s proper label. In classification, the model is completely trained using the training set and evaluated with test data before being used to generate predictions on new, unseen data.

Logistic regression is known as statistical technique. Fre- quently used in binary classification problems to try and predict whether an instance belongs to a certain class or not. By evaluating the correlation between a set of binary variables that are dependent and a list of independent variables, this approach yields a probabilistic result. The logistic function is used by the logistic regression model to transform a linear combination of input data into a probability score that is thresholded to give a binary result. Logistic regression is a useful tool for prediction and classification problems because it is straightforward to use, successful in situations where the result is binary, and easy to understand. It is frequently employed during the decision-making process.

Support Vector Machine (SVM) is a versatile machine learn- ing technique that can handle both regression and classification issues. But its principal advantage and fundamental design is in how well it addresses classification problems.Because SVM can locate the optimal data points in feature space into different classes, use a hyperplane in an N-dimensional space, it excels at classification. In an attempt to provide outstanding generalization to previously unidentified data, the method optimizes the margin between several classes. SVM works better in classification jobs because, while it can be customized for regression tasks as well, its main use is in scenarios where the objective is to discover patterns and boundaries between different categories within the data. The algorithm’s significance for applications needing precise categorization

determinations is shown by its ability to determine the most discriminative hyperplane.

Decision trees may be used for regression tasks as well as the frequent supervised learning problem-solving application of them in classification. This classifier is represented as a tree structure, with each leaf node indicating a particular result, branches signifying decision rules based on aspects of the dataset, and core nodes representing qualities that are used to form conclusions. When it comes to predicting numerical results and classifying occurrences into discrete groups, de- cision trees excel in regression settings. Decision trees are especially useful for evaluating and elucidating the reasoning behind model predictions because of their hierarchical design, which makes decision logic evident. Decision trees are a commonly used tool in a wide range of industries due to their adaptability and interpretability, which provide insights into the interactions between various aspects.

Random Forest machine learning approach combines the predictions of several decision trees to get a consistent and accurate outcome. Its versatility, simplicity of use, and ca- pacity to handle a wide range of tasks, including regression and classification problems, are the main elements boosting its popularity. Random Forest reduces overfitting and improves generalization performance by building an ensemble of deci- sion trees and aggregating their outputs using an average for regression or a voting method for classification. Its ease of use and resilience to noisy data further contribute to its wide range of applications across different domains, making it a top pick for practitioners looking for a reliable and efficient predictive modeling tool.

K-Nearest Neighbors (KNN) is a easy-to-use and user- friendly machine learning method for both regression and clas- sification issues. Predictions made using this non-parametric method are based on the majority class in the case of classification or the average value for regression, which is derived from the k-nearest data points in the feature space. The fundamental concept relies on closeness: each instance in the dataset is represented in a multi-dimensional feature space, and the forecast for each data point is based on the average or consensus of its k-nearest neighbors. KNN’s adaptability, ease of use, and clear concept make it a popular choice for a variety of applications. It offers a useful method of making predictions without the need for explicit model training.

Multi-layer perceptron (MLP) kind of neural network con- sists of three layers of nodes: an input layer, one or more hidden layers, and an output layer.There is only one possible path for information to travel from the input layer to the output layer via the concealed layers. MLPs operate as feedforward neural networks. To optimize the network’s ability to learn and produce accurate predictions, weights are provided to the connected nodes, or neurons, that comprise each layer and are modified throughout the training phase.MLPs are useful tools for a variety of tasks, including regression analysis, pattern identification, and classification, since they are good at finding complex connections and patterns in data. The depth and connectivity of MLPs make them to acquire the ability to

understand hierarchical data structures, which enables them to solve complex issues in a variety of fields.

Naive Bayes classifier is a probabilistic classification method based on feature independence and the Bayes theorem. Because it provides an easy-to-use and efficient approach for text classification, Naive Bayes is a highly helpful tool. It makes the assumption that the characteristics used for classification are conditionally independent in order to stream- line the probability computation. The classifier determines the likelihood of a certain class given a set of attributes by combining prior probabilities and conditional probabilities using the Bayes theorem. Naive Bayes is a well-liked solution for problems like spam identification, sentiment analysis, and document classification in natural language processing applications, despite its ”naive” assumption that features are independent. In real-world scenarios, it often performs excep- tionally well.

XGBoost or Extreme Gradient Boosting, is a well-liked and effective machine learning method. It is particularly good at handling issues with tabular and structured information fast. XGBoost is a member of the gradient boosting fam- ily. It constructs a set of weak learners, usually decision trees, then combines their predictions skillfully to produce an accurate and dependable prediction model. The method’s careful integration of several weak learners, which enables it to continually repair faults and improve expected performance, makes it very successful in predictive modeling, classification, and regression tasks on structured data.

CatBoost, which stands for ”Categorical Boosting,” is an open-source, high-performance gradient boosting toolkit that was created specifically for machine learning applications. Its primary advantage is its exceptional capacity to handle categorical data efficiently and with minimum advance plan- ning. CatBoost works incredibly well at producing reliable and accurate predictions when applied to datasets with categorical variables. The automatic categorical data processing of this library makes it possible to add features of that kind to the machine learning model without the need for human encod- ing or modification. Because it prioritizes performance and usability, CatBoost is a helpful tool for practitioners searching for reliable and efficient gradient boosting solutions. This is particularly true when the predictive modeling procedure makes use of categorical data.

Adaptive boosting, or AdaBoost for short, is an ensemble learning approach that combines the output of several decision trees, or weak learners, to create a robust and accurate model with the goal of improving the accuracy of prediction models. In AdaBoost, each occurrence in the dataset is assigned a weight. During the training phase, the algorithm adjusts these weights to emphasize instances of inaccurate classifications. AdaBoost’s adaptive feature allows it to focus on scenarios that were difficult for the previous weak learners, which allows it to progressively enhance the model’s performance. To produce the final predictive model, the predictions made by the weaker learners are put together and weighted, with larger weight going to those who make more accurate predic-

tions. AdaBoost’s power resides in its capacity to learn from errors iteratively, resulting in a robust and adaptable ensemble model that can handle challenging datasets and raise overall prediction accuracy.

1. *Grid Search Cross-Validation*

Grid Search Cross-Validation, or Grid Search CV, is a method for finding the optimal hyperparameters for a machine learning model by going through a preset set of hyperpa- rameter values. Because it carefully evaluates the model’s performance for each collection of hyperparameters, it is a useful tool for adjusting hyperparameters.

1. *LIME(Local Interpretable Model-agnostic Explanations)*

LIME (Local Interpretable Model-agnostic Explanations) is a method used to explain the predictions of machine learning models. It provides interpretable reasoning for individual fore- casts by locally approximating the complex model’s decision boundary.

1. *Proposed Methodology*

The ’Malayalam Char Gabor’ dataset is used in the proposed study to identify Malayalam palm leaves. To get the dataset ready for analysis, initial data preparation includes feature selection, min-max normalization, and null value elimination. Classifier models are compared with respect to accuracy, precision, recall, F1-score, and other parameters. Next, in order to improve model performance, hyperparameter adjustment is applied, and the classifiers are reassessed. After the best classifier has been chosen, Lime (Local Interpretable Model- agnostic Explanations) is used to give comprehensible expla- nations for each prediction the machine learning model makes. This all-encompassing strategy seeks to not only determine the best classifier for Malayalam palm leaf recognition but also to improve model interpretability using Lime, illuminating the variables impacting individual predictions in the context of the ’Malayalam Char Gabor’ dataset.

1. *Work Flow*

[Fig.1.]The initial step is to prepare the dataset were the it shouldn’t have any null values, filling missing values, removing the duplicate values and changing values from float to categorical. Numerical and categorical are the two types were the data’s are classified. Min-Max normalization is done after performing the data preprocessing. In feature selection 10 features are selected for training and testing. Data is been divided into test and train data. Then the models been trained and tested without performing hyper parameter tuning and after training and testing models normally hyper parameter tuning is done. Model’s evaluating metrics are ac- curacy,precision,F1 score and recall. They are evaluated based on the evaluation metrics with help of the metrics Random forest is been selected as the model which is having high accuracy among the compared models. After the evaluation Lime is been employed.With help of architecture it easy to understand the workflow.

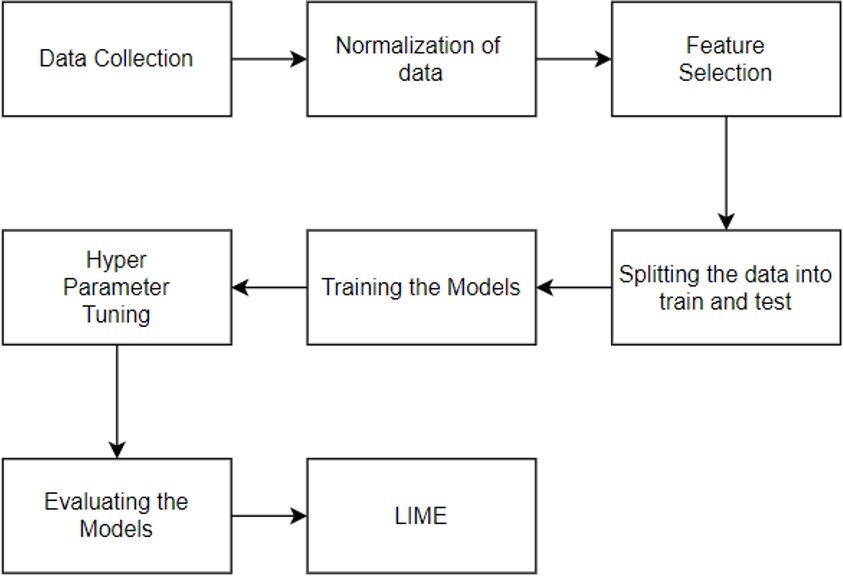


Fig. 1. Architecture of the proposed method

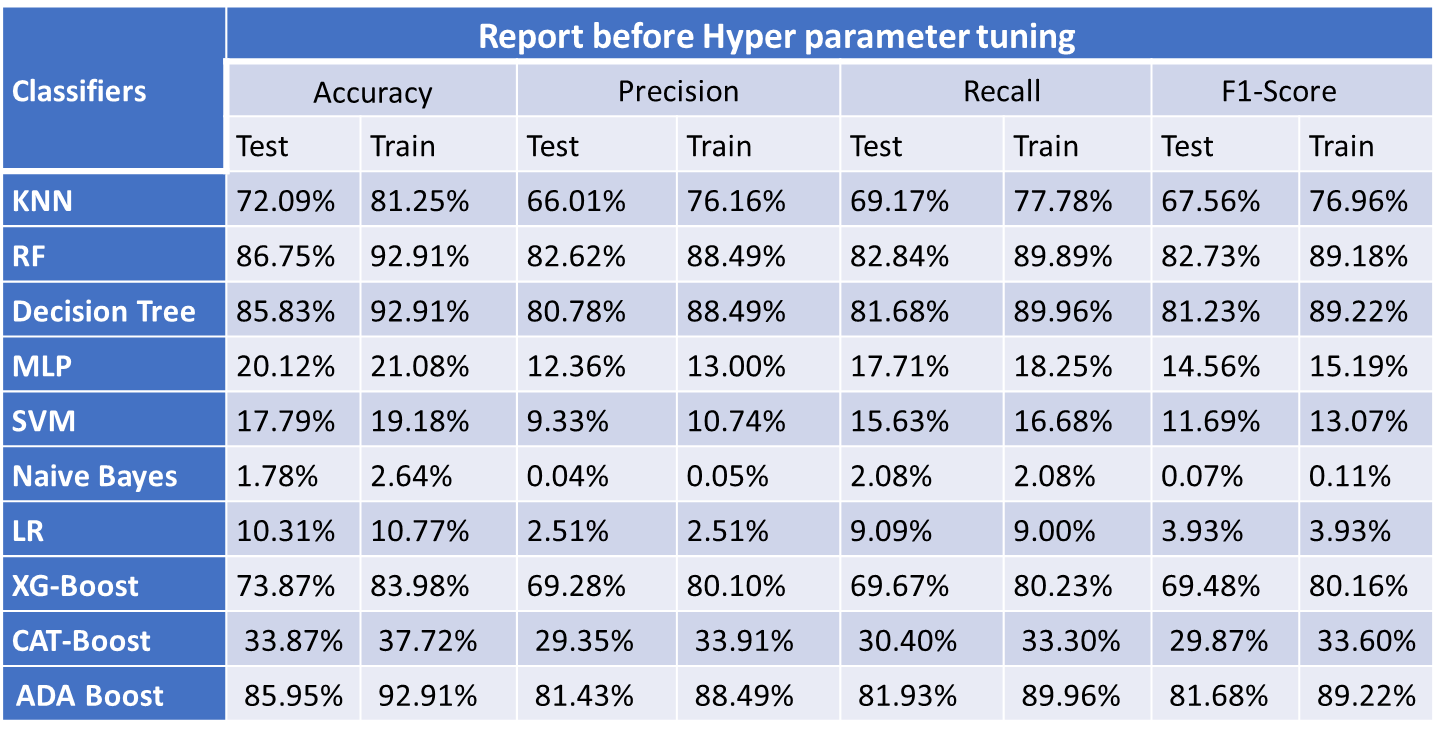


Fig. 2. Without Hyper Parameter Tuning

1. RESULTS AND ANALYSIS

Comparisons are made between classifier models such ADA Boost, XGBoost, CAT Boost, Random Forest Classifier, Deci- sion Tree, K-Nearest Neighbor (KNN), Multi Layer Perceptron (MLP), and Navie Bayes.

[Fig.2.],shows the metrices and the classifier models which has been compared without Hyper parameter Tuning.Random Forest was observed to have the highest test accuracy (86.75%) while ADA Boost followed the accuracy of (85.95%).Random Forest had the highest F1 score(82.73%) while ADA Boost has second higher F1 score of (81.68%).

[Fig.3.],shows the metrices and the classifier models which has been compared with Hyper parameter Tuning.Random Forest was observed to have the highest test accuracy (86.69%) followed by KNN and Decision Tree (Both) with a test accuracy of (86.26%).Random Forest was observed to have the highest F1Score (82.74%) followed by KNN F1Score(82.11%

) and Decision Tree F1Score( 81.62%)

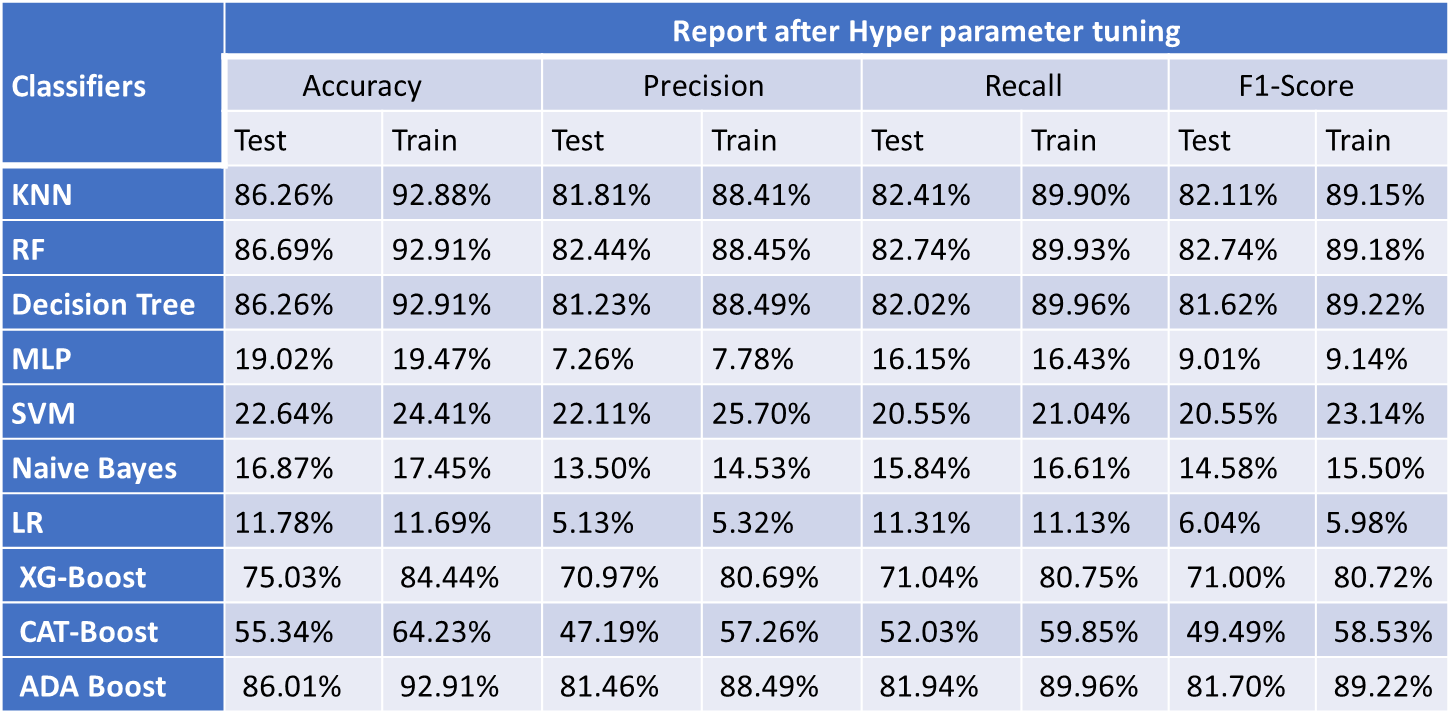


Fig. 3. With Hyper Parameter Tuning

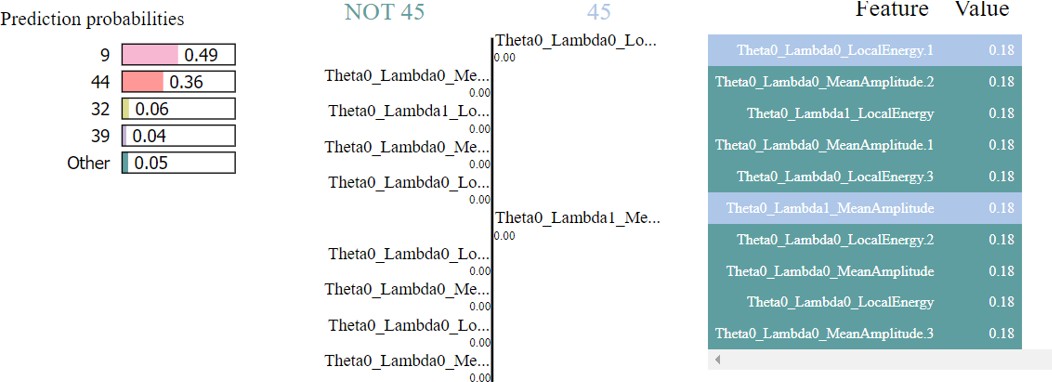


Fig. 4. Lime

[Fig.4.]Lime is highly useful in identifying the significance of traits in predicting outcomes when probabilities are incorpo- rated into the categorization process. Of all the features in the analysis, feature 9 has the highest probability contribution 49% according to Lime’s feature significance rating. This suggests that it significantly affects how the model makes decisions. Feature 44 comes in second place with a chance of 36%. These insights provide a clear understanding of the relative value of different variables in influencing the model’s predictions, as well as an accessible and intelligible look at the factors that substantially contribute to the biggest probability of specific outcomes. These feature-level explanations provide consumers a clearer picture of the internal workings of the system.

1. CONCLUSION

After examining a number of methods, it is clear that Random Forest outperforms the other algorithms in terms of accuracy, with K-Means trailing closely behind. Next, Lime, a tool for model interpretability, is used to characterize the elements influencing the model’s predictions. Given that fea- ture 9 increases the chance to 49%, the evidence suggests that it is important. This important realization deepens our com- prehension of the characteristics that significantly influence the model’s output. After that, lime is used to intentionally decode palm leaf messages. Rather than using accuracy measures, its application depends on character recognition on palm fronds. With the use of this program, palm leaf health may be evalu- ated, characteristics associated with incompleteness, fracture, and brittleness can be identified, and potential contamination and erosion by microorganisms.Lime’s interpretability facili- tates in-depth research and provides insightful information on the nuances of palm leaf characteristics.

In essence, combining Lime’s application, feature interpre- tation, and algorithmic analysis results in a comprehensive comprehension of the model’s functioning and its suitability for defining the state of palm leaves. A comprehensive plan can help with issues like erosion and brittleness, enabling focused treatments and educated decisions on palm leaf research and preservation.

1. A Comparative Study of Feature Learning Methods for Text Line Segmentation” by Y. Liu, Y. Zhang, and X. Liu (2020)
2. Handwritten Character Recognition from ancient palm leaves using gabor based multilayer architecture by Jyothi (2020)
3. Ezhi1arasi and P. U. Maheswari, ”Depicting a Neural Model for Lemmatization and POS Tagging of Words from Pa1aeographic Stone Inscriptions,” 2021 5th Int. Con! Intell.Computer. Control Syst., pp. 1879- 1884,2021
4. Isolated Telugu palm leaf character recognition using Radon Transform by Narahari Sastry (2012)
5. Palm Leaf Manuscript character recognition and classification using convolution Neural Networks by Anand Raju (2019)
6. Ramanan, M., Ramanan, A., & Charles, E. Y. A. (2015). A hybrid decision tree for printed Tamil character recognition using SVMs. 2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (Icterus), 176–181.
7. He, K., Zhang, X., Ren, S., & Sun, J. (2016).Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.
8. PN Sastry and TR Vijaya Lakshmi .(2017).A 3D Approach for Palm Leaf Character Recognition Using Histogram Computation and Distance Profile Features.
9. Ambarish Parida and Hima Bindu Maringanti. 2014 Hybrid Approach for Odia Handwritten Character Recognition.
10. Nina Aleskerova, Aleksei Zhuravlev(2020) Handwritten Chinese Char- acters Recognition Using Two- Stage Hierarchical Convolutional Neural Network

REFERENCES

1. Krichevsky, A., Stuever, I., & Hinton, G. E. (2017). ImageNet classifi- cation with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90
2. Pasha, Saleem, and M. C. Padma , “Handwritten Kannada character recognition using wavelet transform and structural features”, 14th In- ternational Conference on Frontiers in Handwriting Recognition, (pp. 346-351), 2014.