ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

2022, x(x), pp. xx-xx, DOI: xx.xxxxx/xxxxx.xxxx

PUMPKIN SEED CLASSIFICATION

Bahar GÖRGÜN[1], İlayda Su İRDAY[2]

[1] Computer Engineering, Faculty of Engineering, Eskisehir Technical University, Eskisehir, Turkey.
[1] Computer Engineering, Faculty of Engineering, Eskisehir Technical University, Eskisehir, Turkey.

ABSTRACT

This study was carried out on the two most important and quality types of pumpkin seeds, “Ürgüp Sivrisi” and “Çerçevelik”, generally grown in Ürgüp and Karacaören regions in Turkey. Ürgüp Sivrisi and Çerçevelik are Class features. all the data were modeled with five different machine learning classification methods: Logistic Regression (LR), Naïve Bayes, Support Vector Machine (SVM) and Random Forest (RF), and k-Nearest Neighbor (k-NN), which further determined the most successful method for classifying pumpkin seed varieties with Accurracy Score, Precision and Recall Score.
Keywords: Classification,Random Forest,KNN,Logistic Reggression

1. INTRODUCTION

A pumpkin seed, also known in North America as a pepita (from the Mexican Spanish: pepita de calabaza, "little seed of squash"), is the edible seed of a pumpkin or certain other cultivars of squash. The seeds are typically flat and asymmetrically oval, have a white outer husk, and are light green in color after the husk is removed. Some cultivars are huskless, and are grown only for their edible seed. The seeds are nutrient- and calorie-rich, with an especially high content of fat (particularly linoleic acid and oleic acid), protein, dietary fiber, and numerous micronutrients. Pumpkin seed can refer either to the hulled kernel or unhulled whole seed, and most commonly refers to the roasted end product used as a snack.

In this study, we evaluated the performance of five different machine learning models on s dataset containing pumpkin seed features. The models included a Logistic Regression (LR), Naïve Bayes, Support Vector Machine (SVM) , Random Forest (RF), and k-Nearest Neighbor (k-NN). The dataset consistent of 2500 entries and 13 features, including area, perimeter, major axis length, minor axis length, convex area, equiv diameter, eccentricity, solidity, extent, roundness, aspect ration, compactness and class.

*Corresponding Author: xxxxxx@xxxxxxxx.xxx.xx
Recieving Date: ... Publishing Date: ...

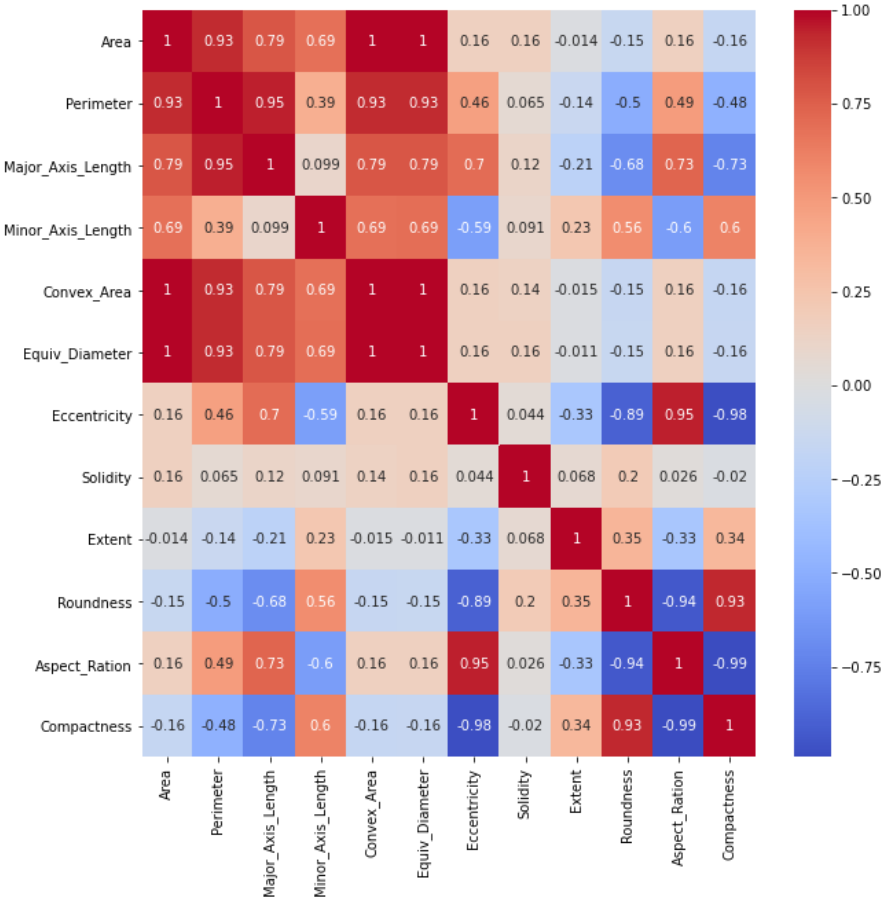


Figure 1. This figure is correlation matrices. A correlation matrix is a table that shows the correlation between multiple variables. It is used to summarize the relationship between different variables and to identify the strength and direction of the relationship. In a correlation matrix, each variable is represented by a column and a row, and the cells of the matrix contain the correlation coefficients between pairs of variables. The diagonal of the matrix is always 1, since each variable is perfectly correlated with itself. The off-diagonal cells contain values between -1 and 1, indicating the strength and direction of the relationship between the two variables. A value of -1 indicates a perfect negative correlation, a value of 1 indicates a perfect positive correlation, and a value of 0 indicates no correlation.

The models were trained and tested on the dataset using a 75/25 train/test split. The results showed that the Random Forest model had the highest accuracy, with a score of 87.20%.The KNN had the lowest accuracy, with a score of 81.12 %. These results suggest that machine learning models can be useful for predicting heart disease risk and that the Random Forest model may be the most suitable for this purpose.

To be give information about the algorithms that are used in the project.

1.1 Logistic Regression

Logistic regression is a type of supervised learning algorithm that is used for classification tasks. It is a linear model that is used to predict the probability that an example belongs to a particular class. In logistic regression, the model is trained on a set of input features and a target variable that consists of two or more classes. The model learns a set of weights and a bias term that are used to make predictions based on the input features. The predictions are made using the logistic function, which maps the input

data to a probability between 0 and 1. The predicted probability is then used to classify the example into one of the classes. Logistic regression is a popular choice for classification tasks because it is simple to implement and can handle a large number of input features. It is also robust to overfitting, which means that it can generalize well to new data. However, it is limited in its ability to model complex relationships between the input and output data. In conclusion, logistic regression is a powerful and widely-used machine learning algorithm that is used for classification tasks. Its simplicity and ability to handle a large number of input features make it a popular choice for many data scientists. While it is limited in its ability to model complex relationships, it is a useful tool for many applications.

1.2 Naive Bayes

In statistics, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features (see Bayes classifier). They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve high accuracy levels.

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

In the statistics literature, naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method.

1.3 (SVM) Support Vector Machines

Support vector machines (SVMs) are a type of supervised learning algorithm that is used for classification and regression tasks. They are a powerful and widely-used machine learning method that is known for their ability to handle high dimensional data and to perform well on a variety of tasks. In an SVM, the model is trained on a set of input features and a target variable. The model learns a set of weights and a bias term that are used to make predictions based on the input features. The predictions are made by finding the hyperplane in the feature space that maximally separates the classes. SVMs have a number of advantages, including their ability to handle high dimensional data and to perform well on a variety of tasks. They are also robust to overfitting, which means that they can generalize well to new data. However, they can be computationally expensive to train, especially on large datasets. In conclusion, SVMs are a powerful and widely-used machine learning algorithm that is used for classification and regression tasks. They are known for their ability to handle high dimensional data and to perform well on a variety of tasks, but can be computationally expensive to train on large datasets.

1.4 Random Forest

Random forests are a type of ensemble learning method that can be used for both classification and regression tasks. They are a popular choice for machine learning because of their high accuracy and robustness to overfitting. In a random forest, a large number of decision trees are trained on randomly selected subsets of the training data. During prediction, each decision tree in the forest makes a prediction, and the final prediction is made by taking the majority vote of the decision trees. This approach helps to reduce the overfitting that can occur when using a single decision tree.One of the key advantages of random forests is their ability to handle missing values in the input data.They can also handle high dimensional data, which means they can work with a large number of input features. Additionally, random forests are easy to use and have fast training and prediction times, making them a good choice for large datasets or real-time applications. In conclusion, random forests are a powerful machine learning algorithm that can be used for a variety of classification and regression tasks. Their high accuracy, robustness to overfitting, and ability to handle missing values and high dimensional data make them a popular choice for many data scientists.

1.5 KNN

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric supervised learning method first developed by Evelyn Fix and Joseph Hodges in 1951, and later expanded by Thomas Cover. It is used for classification and regression. In both cases, the input consists of the k closest training examples in a data set. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors. If k = 1, then the output is simply assigned to the value of that single nearest neighbor.

k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.

Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

2. MATERIALS AND METHODS (or EXPERIMENTAL)

Several metrics have been proposed for evaluating classification performances. 'Area Under (ROC) Curve' (AUC) has its origins in signal detection theory in the 1970s is considered as a best metric to state the performance, but there are other combined metrics that are useful for indicating the successful and unsuccessful aspects of a classification model.

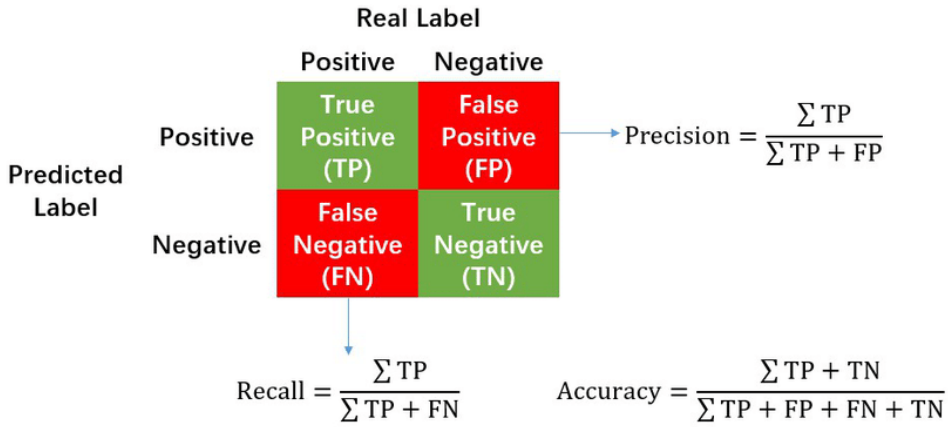


Figure 2. This figure shows performance metrics.

2.1 Accuracy Score

Accuracy is a metric commonly used to evaluate the performance of a classification model. It is calculated by dividing the number of correct predictions made by the model by the total number of predictions. For example, if a model makes 100 predictions and 75 of them are correct, the model's accuracy would be 75%. While accuracy is a useful metric for evaluating classification models, it can be misleading if the classes are imbalanced (e.g. if there are many more negative examples than positive examples). In these cases, other metrics such as precision and recall may be more informative.

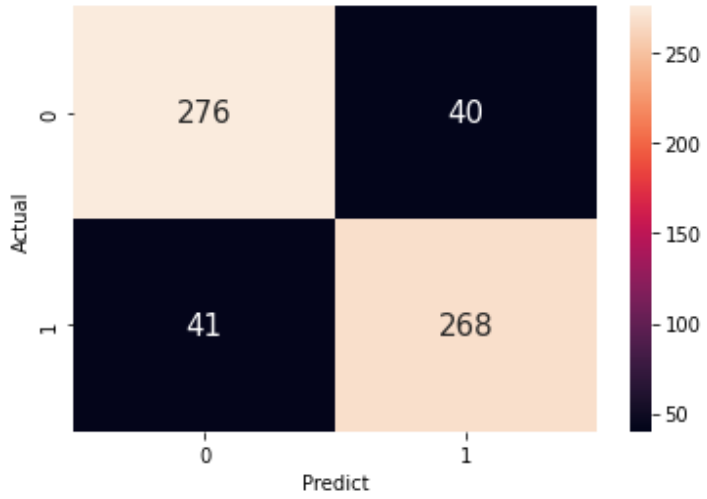


Figure 3. This figure shows accuracy score for Random Forest. Accuracy of Random Forest: % 87.039.

2.2 Precision Score

Precision Score is a metric that measures the proportion of true positive predictions made by a model to the total number of positive predictions made. It is calculated by dividing the number of true positive predictions by the total number of positive predictions made.

2.3 Recall Score

Recall Score is a metric that measures the proportion of true positive predictions made by a model to the total number of actual positive instances in the data. It is calculated by dividing the number of true positive predictions by the total number of actual positive instances. These three metrics are commonly used to evaluate the performance of classification models, particularly in the context of binary classification.

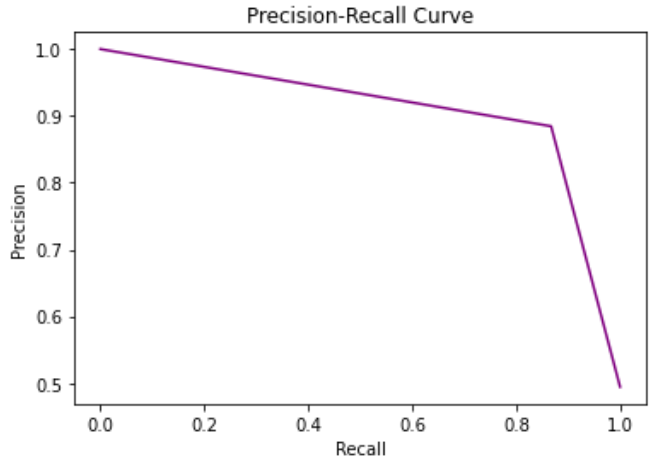
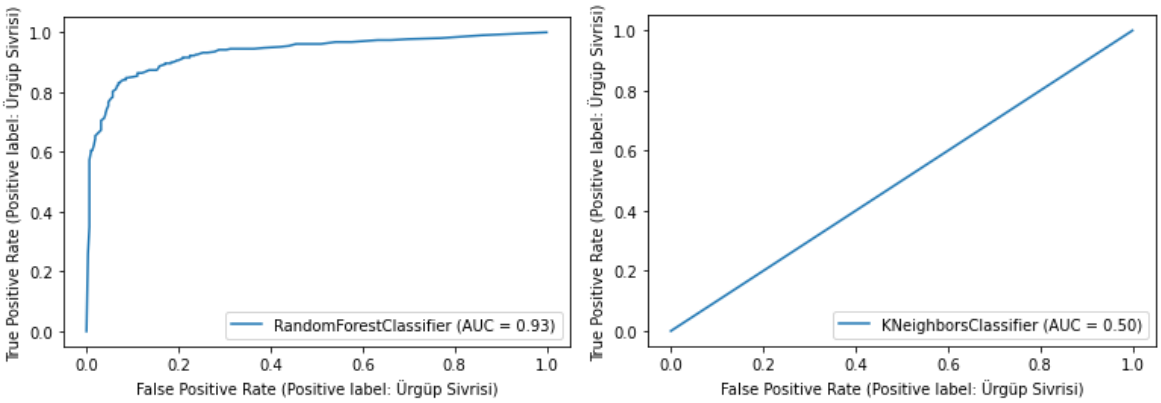


Figure 4. This figure shows precision recall curve. The precision-recall curve shows the tradeoff between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

2.4 Auc-Roc Curve

The AUC-ROC curve is a tool used to evaluate the performance of a classification model. It plots the true positive rate against the false positive rate at various classification thresholds. The AUC, or area

under the curve, represents the model's ability to distinguish between the positive and negative classes.A higher AUC indicates a better performing model. The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.True Positive Rate (TPR) is a measure of the accuracy of a classification model in predicting positive samples. It is defined as the proportion of positive samples that are correctly predicted as positive by the model. For example, if a model is designed to predict airplane radar signals and it correctly predicts 90 out of 100 signals as airplanes, the TPR would be 90%. False Positive Rate (FPR) is a measure of the accuracy of a classification model in predicting negative samples. It is defined as the proportion of negative samples that are incorrectly predicted as positive by the model. For example, if a model is designed to predict airplane radar signals and it incorrectly predicts 10 out of 100 signals as airplanes, the FPR would be 10%.



(a) (b)

Figure 5.(a)The most successful model is Random Forest because AUC value is highest. (b) The most unsuccessful model is KnearestNeighbors.

3. RESULTS

The Random Forest model had an accuracy of 5 87.20%, the Logistic Regression model had an accuracy of 85.60%,%, the SVM model had an accuracy of 85.60%, the Naive Bayes model had an accuracy of 83.84%, the KNN model had an accuracy of 81.12 %. To sum up, the Random Forest model had the highest accuracy, with a score of 87.20%.

4. DISCUSSION

The results of this study suggest that machine learning models can be useful for pumpkin seed classification. Of the five models tested, the Random Forest model had the highest accuracy, with a score of 87.20%. This suggests that the Random Forest model may be the most suitable for pumpkin seed classification. However, it is important to note that the results of this study should be interpreted with caution, as the dataset used was relatively small and may not be representative of the entire population. Further studies with larger datasets are needed to confirm these findings.

5. CONCLUSION

In conclusion, Five models were evaluated on a dataset containing 2500 rows and 13 features. The results showed that the Random Forest model had the highest accuracy, with a score of 87.20%, otherwise the KNN had the lowest accuracy, with a score of 81.12%. These findings suggest that machine learning models can be useful for pumpkin seed classification and that the Random Forest model may be the most suitable for this purpose. Further research is needed to confirm these results and to determine the most effective ways to apply these models in professional setting.

ACKNOWLEDGEMENTS

We would like to thank to Assit. Prof. Mehmet KOC. We are grateful for the education given to us.

CONFLICT OF INTEREST

The authors (İlayda Su İrday, Bahar Görgün) declare that they have no conflict of interest.

REFERENCES

https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/
Encyclopedia of Machine Learning, New York:Springer, 2011.
https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
https://deepchecks.com/how-to-check-the-accuracy-of-your-machine-learning-model/

WORK SHARING

İlayda Su İRDAY Data Pre-Processing, train and test Machine Learning Models and writing final report. Bahar GÖRGÜN Writing Project Proposal Report, Writing Progress Report, Data Analysis, ML Model's Scores, Writing Final Report