Comparative Analysis between KNN and Decision Tree Models for Classifying Adult and Wilt Datasets.

Le-Li Mao (260800098), Marco Caniglia (260929489), Victor Livernoche (260926276)
September 30th 2021

Abstract

There exists various machine learning models used to classify data. This paper aims to investigate and compare the performance of two traditional classification models, the k-nearest neighbors and decision tree models. Two datasets will be used to perform the classifications. The first being the Adult dataset. Our model will use the dataset to predict whether an adult's income exceeds $50k/year based on various factors related to employment. The second being the Wilt dataset, where our model uses a high-resolution remote sensing dataset to predict whether a section of trees is diseased based on the satellite image attributes. For the adult dataset, the decision tree has better performance than the knn model while for the wilt dataset the knn tends to perform better than wilt. In terms of training speed when performing cross validation, we found that the knn trains slower than the decision tree due to its bigger time complexity.

1 Introduction

In light of the no free lunch theorem, which states that there are no universal best-performing algorithms, comparing two different classification techniques on a dataset gives strong insights on the performance of these models on real data [1]. Decision trees are well known for their fast running-time, as they require no data preprocessing, but can easily overfit. On the other hand, k-nearest neighbours is less efficient, but requires no supervision while only having a single hyperparameter. Both classification techniques were used on the Adult and Wilt datasets to determine which model is more accurate and efficient on a target dataset [2]. The algorithms have to predict on the Adult dataset whether or not an person makes more than $50,000 per year based on some employment information such as their age, education, workclass, sex, capital and more. For the Wilt dataset, the model will predict if a patch of trees is deceased based on the features given by a satellite image such as mean green, red, texture pan value, and more. The project shows that the decision tree outperforms KNN to predict incomes with the Adult dataset (see Table 1), while k-nearest-neighbor performs better to predict the correct label with the Wilt dataset. It also shows that the decision tree running time is much faster than KNN for both datasets.

	Adult dataset	Wilt dataset
KNN	82.6 %	80.6 %
Decision Tree	85.1 %	76.0 %

Table 1. Comparison of model accuracy on different datasets

2 Datasets

2.1 Adult Dataset

The adult dataset contains 14 features, 8 of which are discrete features, while the 6 remaining are continuous. In the preprocessing phase, the continuous features were encoded, using One-Hot encoding, to allow the classification of the data. All data entries missing any information were removed for the prediction as they did not represent a significant part of the dataset. The min-max normalization was applied to restrict the values between 0 and 1 in order for all values to have the same weight in the classification. Without this normalization, we found that KNN was less accurate since some features were overshadowed by others.

The features represent information about adults such as their age, workclass, education, marital status, relationship, race, sex, capital gain and loss, hours of work per week, and native country. Each data entry contains a label that defines whether or not the individual's income is greater than $50,000 per year. For example, the histogram

below represents the distribution of age of the dataset.

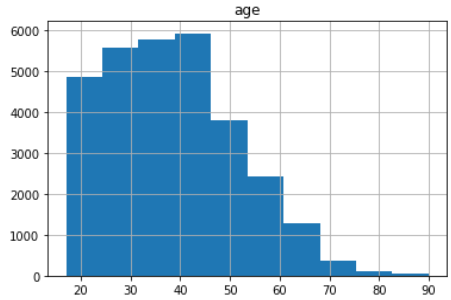

Figure 2. Age distribution of the Adult dataset

Feature creation was investigated for this dataset, but dismissed as many features are discrete in nature and could not provide additional insights by logically combining them.

2.2 Wilt Dataset

The wilt dataset label contains 5 labels, all of which are continuous. The prediction label is binary, the outcome is either 'w' which represents diseased trees or 'n' represents non disease trees. The 5 input parameters of the data are the following: GLCM mean texture, mean green value, mean red value, mean NIR value, and standard deviation (Pan band). These attributes are extracted from the quickbird imagery [2] and all attributes are floating point numbers representing the intensity.

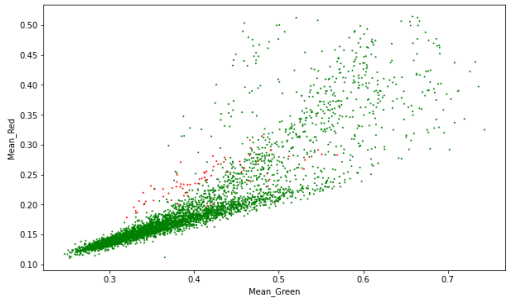

Figure 3: Mean intensity of mean green vs red values of image within the dataset with red dot being the diseased tree and green being healthy.

In terms of preprocessing, we performed a re-mapping of the labels from a string format into a discrete number and converted the input features value into numpy floating point numbers. Although all the features are floating point numbers, we did not combine any features into new features due to the inability to find any correlation when examining the data.

3 Results

3.1 Adult Dataset

In order to evaluate the accuracy for knn on the Adults dataset, the model was tested with 25 different values for the hyperparameter. After having performed a 5-fold cross validation on the set of k values, we observe that the highest accuracy is observed with a k of 9, which provides an accuracy of 0.826.

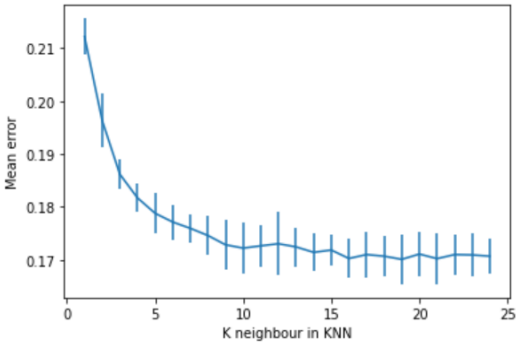

Figure 4: Cross validation of KNN across range of k values for Adult dataset

We then tested the same dataset with the decision tree classification model, which produced an accuracy of 0.851 with the max_depth hyperparameter set to 6.

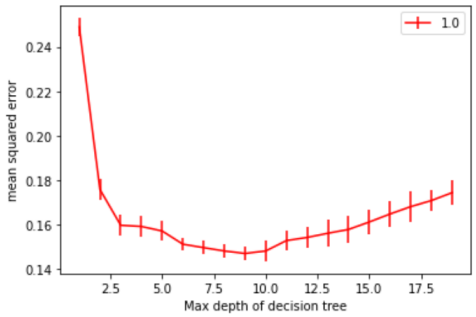

Figure 5: Cross validation of KNN across range of k values for Adult dataset

The obtained results varied with the size of the input dataset. After having run the model with sections of 0.1%, 1%, 10%, 50%, and 100%, we can observe that the variation of mean error is much higher when selecting smaller sections of the data.

3.2 Wilt Dataset

For the Wilt dataset, we achieved an accuracy of 0.796 for the k-nearest neighbor with the best number of neighbour parameters being set to 1 with rule of thumb when performing the cross validation.

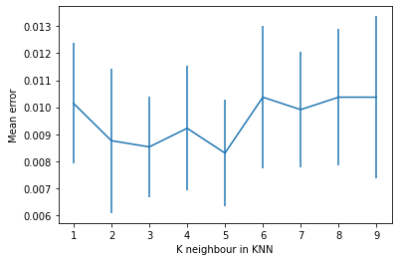

Figure 6: Cross validation of KNN across range of k values for Wilt dataset

For the decision tree, we achieved an accuracy of 0.760 with the best max depth parameters being set to 12 since the rule of thumb wasn't used due to high standard deviation that led to worse accuracy.

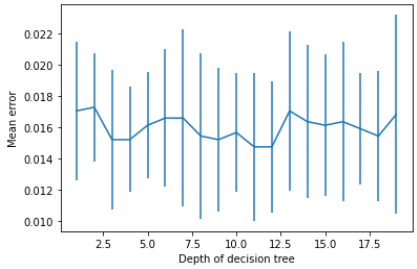

Figure 7: Cross validation of decision tree across range of k values for Wilt dataset

The accuracy during cross-validation was much better than the accuracy of the test data. This is due to the fact that most training data entries are marked as healthy while about half is in the test data. Mixing training and test datasets would fix the issue. We also performed the same test for different portions of the data size.

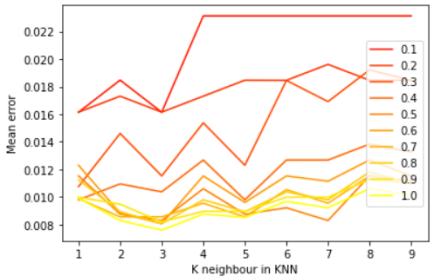

Figure 8: Mean error of KNN for different dataset size

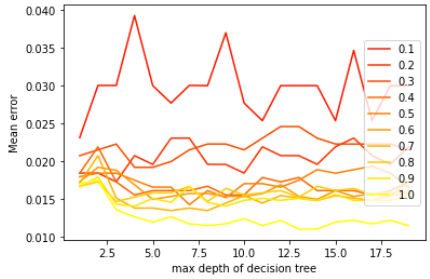

Figure 9: Mean error of decision tree for different dataset size

We discovered that reducing the data by a small amount may not affect the accuracy but larger reduction increases the error rate and deviation of the error substantially. Below shows the plot for different portion sizes ranging from 10% to 100% of the data size on both KNN and decision tree.

4 Discussion & Conclusion

The project showed that decision trees are more efficient in the classification than KNN and more accurate for Adult while KNN is more accurate for Wilt. This could be due to the greater size and number of features that Adult has compared to Wilt. It showed a real example of application of both classifiers and some of their advantages. Further work could be done to improve accuracy of the prediction done on the test datasets. For the decision tree, we could investigate other hyperparameters to optimize such as the minimum samples leaf and split and also modify the criterion function. We could try different normalization techniques for the KNN classifier. Different scaling and specific features of both datasets could be tested to see if it can improve accuracy. We noted that decreasing the size of the dataset often leads to worse accuracy. However, reducing the size of the dataset by a small amount can decrease the processing time while not necessarily losing accuracy. Another thing to note is that normalizing the data can lead to slower running time for KNN if we had integers since floating points are slower to compare. Lastly, we could see how KNN and decision trees perform against other machine learning algorithms, namely neural networks, linear regression, logistic regression, random forest and SVM to cite some of them.

5 Statement of Contributions

Le-Li Mao, Marco Caniglia and Victor Livernoche all discussed and contributed to all parts of the project in different proportions. For the most part, Le-Li Mao worked on data processing of Wilt and both models under the Wilt dataset while Marco Caniglia worked on KNN within the Adult dataset And Victor Livernoche worked on the preprocessing of Adult, modeling of decision tree, the writing of the introduction, adult dataset and discussion.

References

[1] Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation 1, 67

[2] Murphy, P. M., and Aha, D. W. 1996. UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn.