Mini Project 1: Machine Learning 101

Anthony Porporino (260863300), Tyler Watson (260867260), Kynan Nedellec (260866794)

Abstract - In this project our team compared and contrasted two machine learning algorithms, k-nearest neighbors (KNN) and decision trees, by measuring their performance in a classification problem on two datasets. The first dataset (adult dataset) contained census information about individuals such as age and education level as well as whether or not their annual income was over 50 thousand dollars. The second dataset (news dataset) contained information about news articles including title length and number of links contained as well as whether or not the number of shares was above 1400. For both datasets we had access to the target labels. By tweaking certain hyperparameters in both algorithms as well as size of the training data set, we were able to determine the optimal model for each dataset classification problem. For the first dataset, we found that KNN with 12 neighbors using robust scaling resulted in the best accuracy of 86% on the testing data. Contrastively, a decision tree algorithm gave the best results with an accuracy of 63% for the second classification problem.

I - Introduction

The project's main task was to explore how preprocessing of data, hyperparameter choices, and size of the training data set each affect the accuracy of two machine learning algorithms tasked with two different binary classification problems. The two machine learning algorithms that were used were k-nearest neighbors (KNN) and decision trees. Both algorithms have hyperparameters that can be modified in order to improve performance. Model performance comparisons were done on two separate datasets. The first contained adult census information to determine if an individual's annual income was above 50 thousand dollars and the second contained information about online new articles to determine whether or not they had been shared over 1400 times.

Before running the experiments on each algorithm, we performed some initial research by skimming through some research papers and experiments on this exact comparison. The one that stuck out the most, was out of King Saud University where they compared Naive Bayesian, decision trees, and KNN on a sentiment classification problem. The stand out points from this experiment were that decision trees were much faster than KNN, and that an increasing training data size has different effects on each algorithm [1]. We were therefore extra vigilant for anything that would confirm these two observed trends.

The experiments we ran generated 5-fold cross validation mean accuracy values for all possible combinations of hyper parameters and preprocessing decisions for both algorithms on both datasets. We also ran this same process using different subset sizes of the data. The results from these experiments indicate that as the size of the training dataset increases, both KNN and decision trees models increase in accuracy, with the exception that KNN tends to drop off when using the entirety of each available dataset, as stipulated by the aforementioned study. We also concluded that KNN was more accurate for the adult dataset, whereas a decision tree performed better on the news dataset.

II - Datasets

The first dataset (adult dataset) is extracted from the 1994 US census covering basic personal information such as age, work-class, education, race, native-country, and more, for a total of 17 features. For the principal hyperparameter analysis, we removed all records containing missing entries (2400 on 32543).

Both the relatively few missing records (on the total dataset), and the fact that out of the 2400 records with missing instances 1836 were missing data on another feature, motivated this decision. We also decided to remove the 'education' feature because of its direct correlation with the 'education-num' feature, as shown below. This choice removed redundant information, and an unwarranted bias towards the ‘education’ feature (most notably for KNN models).

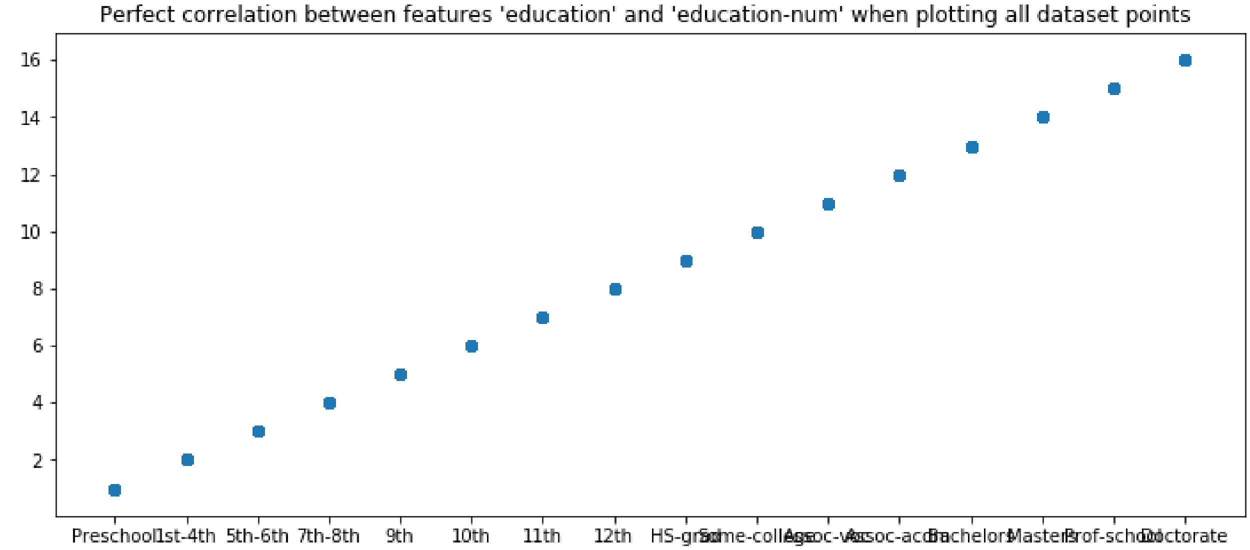


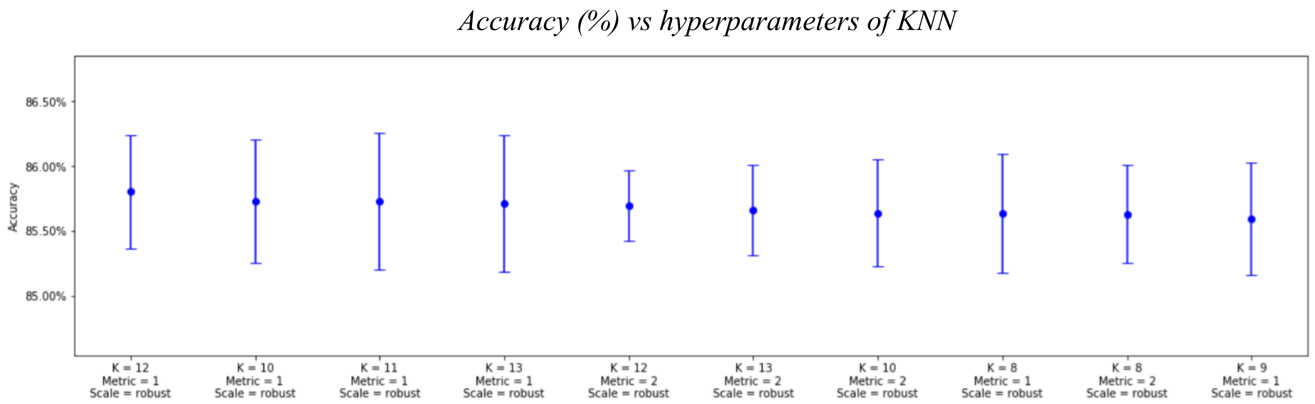
Figure 1: Scatter plot illustrating feature redundancy

The second dataset we decided to analyze contained information about online news articles including the number of words in the title, the average length of words in the content, the number of links contained, etc. The target classification we chose to do was to determine whether the number of times an article was shared was above or below 1400 since this essentially divided the dataset in half. This dataset originally had 60 features which were reduced to 23 to improve and remove seemingly irrelevant features for our prediction.

III - Results

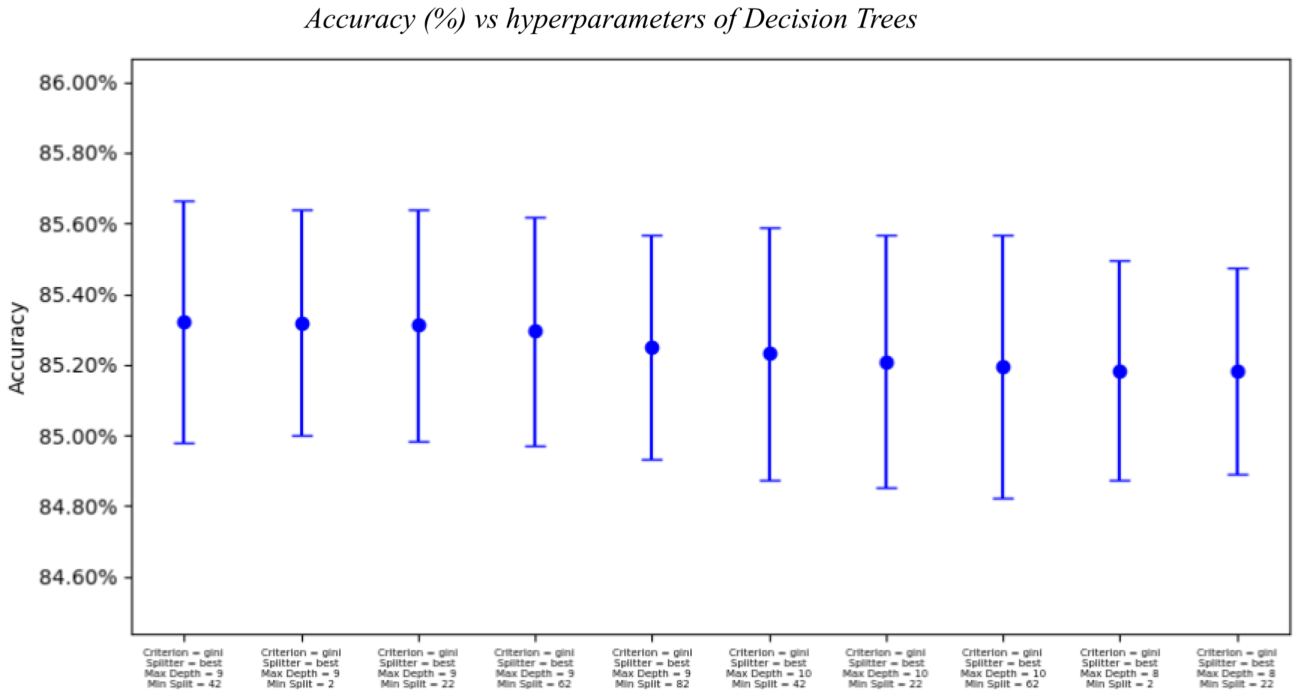
To optimize hyperparameters and preprocessing decisions, we ran all possible combinations for both KNN and decision trees. The hyperparameters for KNN were the value of k (1 to 14) and the distance metric (euclidean or manhattan distance). For data preprocessing we tried four options including no preprocessing, normalization, robust, and min max scaling algorithms. The hyperparameters explored for the decision tree included the depth of the tree, the minimum number of samples to split, the splitter (best or random) and finally the criterion to measure the quality of the splits at each step (gini or entropy).

The results of our experiments show that both KNN and decision trees perform very well for the salary classification problem. Figure 2 below shows the hyperparameters of the top 10 KNN models with the highest mean value for 5-fold cross validation as well as the error bars showing the standard deviation. Figure 3 shows the same metric for decision trees. In the appendix, Figure 6 to 13 show the top 10 KNN models and decision tree models with the highest mean on the training set.



Accuracy (%) vs hyperparameters of KNN

Figure 2: Top 10 KNN models with the highest mean cross validation score on adult validation set



Accuracy (%) vs hyperparameters of Decision Trees

Figure 3: Top 10 decision tree models with highest mean cross validation score

The results indicate that KNN performs slightly better on the adult dataset. We decided that the best KNN algorithm was one that was scaled using the robust scaling algorithm, used a euclidean distance (metric 2) for distance calculations and a k set to 12. This model had a similar mean score to the other best performing model but had a significantly lower standard deviation.

The confusion matrix showing the results when running this specific k nearest neighbor algorithm on the test set is shown in Figure 4:

	Predicted <=50K	Predicted >50K
Actual <=50K	10601	759
Actual >50K	1374	2326

Accuracy: 0.8583665338645419

Figure 4: Confusion matrix for KNN model on adult test dataset

Although we saw better results with the KNN model, the computation time for predicting values using the KNN model was orders of magnitude larger than that for decision trees, and that for best models with similar mean accuracy (both around 85-86%). This is definitely a benefit of using decision trees over KNN for this problem.

The same process was done for the news dataset. Figure 5 in the appendix shows the results for KNN hyperparameter tweaking and Figure 6 shows it for the decision trees. This dataset proved to be much harder to train and accuracy levels were considerably lower than those achieved for the salary classification problem (the best models achieving less than 64% mean accuracy). One likely reason could be a lack of correlation between the selected features and the classification label. The removal was done with the goal of decreasing computation time.

The best performing model in this case was a decision tree model with hyperparameters set to 6 for max-depth, gini for greedy split calculation, and the splitter choosing the best feature to split on each time. However, the mean accuracy was not so different between the two criterion (gini and entropy) and it seems as though the min sample split also did not have much of an effect on these tests. This could be due to the max depth being so low that the min samples level was irrelevant. The confusion matrix for this can be seen in Figure 14 in the appendix.

Lastly we ran experiments to determine how varying the size of the dataset affects accuracy. We took 5 subsets each dataset in increasing size and ran all combinations of hyperparameters and preprocessing decisions for both models on both datasets. Figure 5 below shows the results for this process on the KNN algorithm for the adult dataset. As expected, the error rate decreases more significantly on the validation data than it does on the training data as the sets of training data are increased in size. This can be explained by reduced overfitting at each training iteration, as the model is less likely to capture spurious correlations between training instances as their number increases [3].

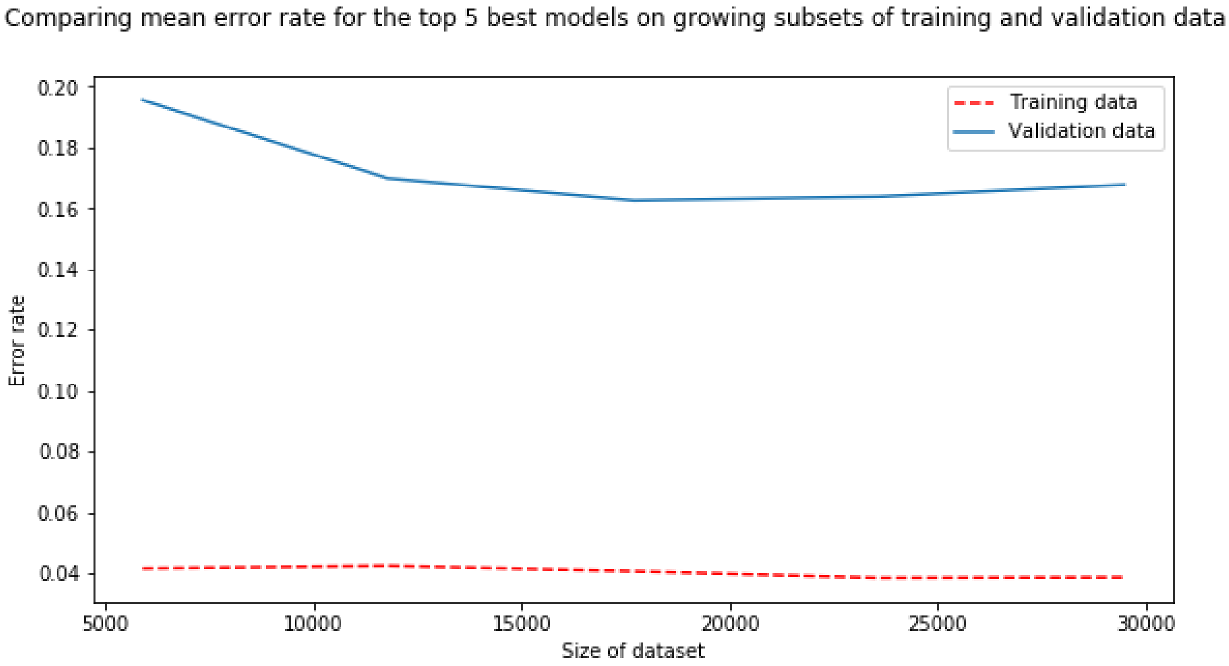


Figure 5: Accuracy for KNN algorithm on Adult dataset for different subsets with different sizes

These results, as well as the results from the other experiments shown in Figure 15 to 17 in the appendix, indicate that a larger data set generally increases performance decision trees. Although the same is generally true for KNN as well, in both KNN graphs we can see that the accuracy drops off when using the entire dataset. This seems to suggest that more data points may not always be advantageous for KNN.

IV - Discussion and Conclusion

After analyzing the results from all the experiments, there are clearly some key takeaways from this project. The first, is that the size of the training dataset definitely matters to achieve better KNN and decision tree models. Indeed, for both of these we observed a general trend indicating that an increasing dataset size leads to a more accurate model. The second takeaway is that there is no “one model fits all” when it comes to different problems. This is clear by the fact that KNN performed better than a decision tree on the adult dataset, but not as well on the news dataset. It is therefore important to try multiple models to tackle whatever problem one is trying to solve, because it is definitely not “one size fits all”. The final takeaway is that modifying hyperparameters makes a major difference on model performance. This was fairly intuitive before the project, however, it is clear now that we have analyzed our results.

The difficulty, however, was in selecting the range of values to actually test the hyperparameters with. This is definitely an area that could be investigated further because the ranges we selected were, for the most part, arbitrary. Another optimization technique we had hoped to investigate further had time permitted, is the feature scaling of the KNN model. We believe that certain features such as education level and workclass are much better indicators on whether that person makes over 50 thousand dollars a year than other features such as marital status and age. Thus, we hypothesize that scaling those features and reducing others, to reduce noise, would be beneficial to our model’s performance.

V - Statement of Contributions

In terms of the code, Kynan worked on writing the cross validation function, the KNN experiments and data cleaning, Tyler worked on getting the new dataset, the decision tree experiments as well as the varying subset experiments, and Anthony worked on data cleaning, KNN scaling, subset experiments, and running chosen models on the test data set. The writeup was divided amongst team members and each member contributed an equal amount.

V - References

[1] M. Bilal, H. Israr, M. Shahid, and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naive Bayesian, decision tree and KNN classification techniques," *Journal of King Saud University Computer and Information Sciences*, 12-Dec-2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157815001330. [Accessed: 24-Sep-2021].

[2] D. McInerney and M. Nieuwenhuis, "A comparative analysis of kNN and decision tree methods for the Irish National forest inventory," *Taylor & Francis*, 22-Sep-2009. [Online]. Available: https://doi.org/10.1080/01431160903022936. [Accessed: 24-Sep-2021].

[3] Xue Ying, "An Overview of Overfitting and its Solutions", Journal of Physics: Conference Series, Volume 1168, Issue 2. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1168/2/022022. [Accessed: 25-Sep-2021].

VII - Appendix

Cross validation results - KNN

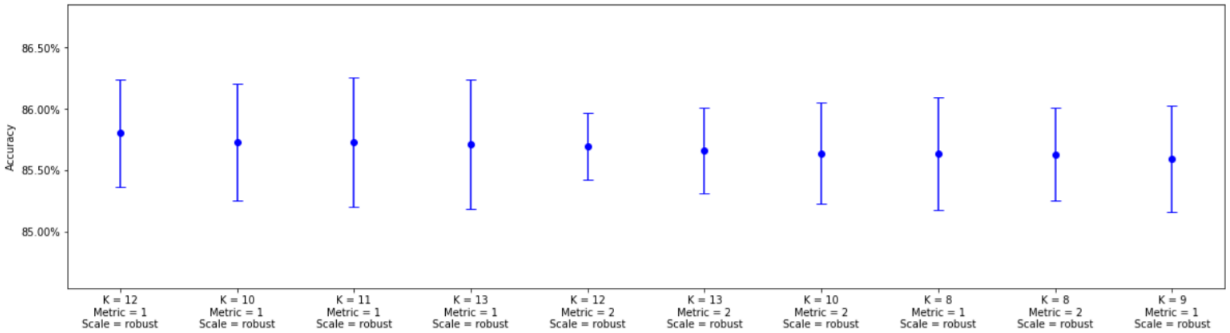


Figure 6: Top 10 KNN models with the highest mean cross validation score - adult dataset (validation)

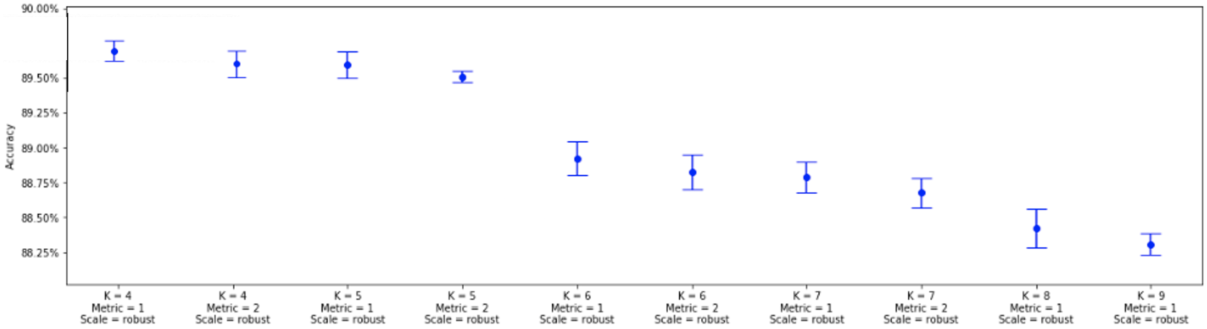


Figure 7: Top 10 KNN models with highest mean cross validation score - adult dataset (training)

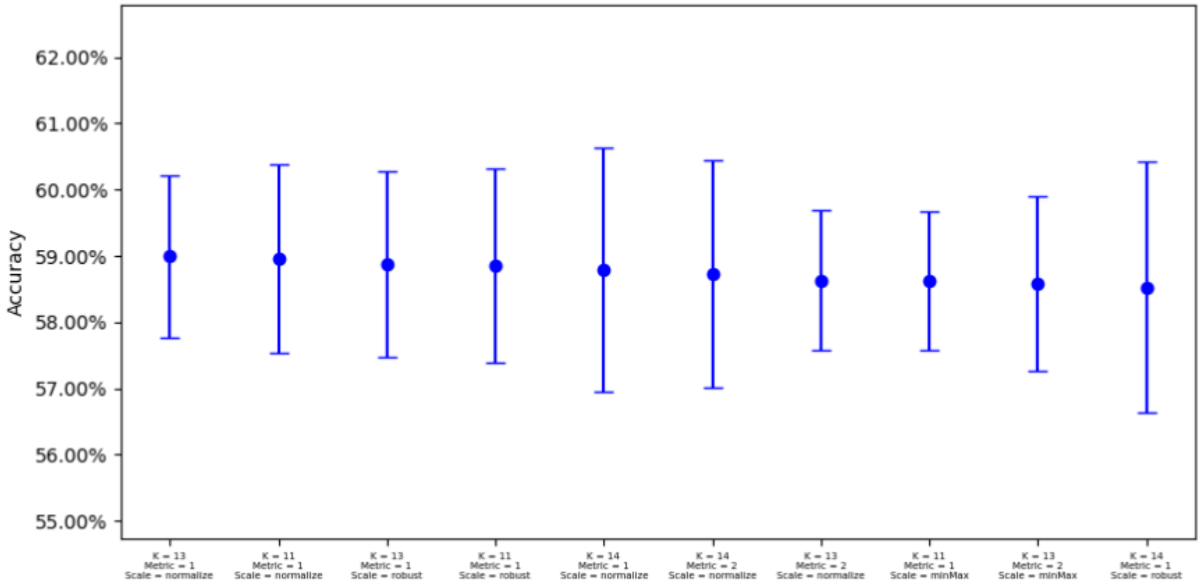


Figure 8: Top 10 KNN models with highest mean cross validation score - news dataset (validation)

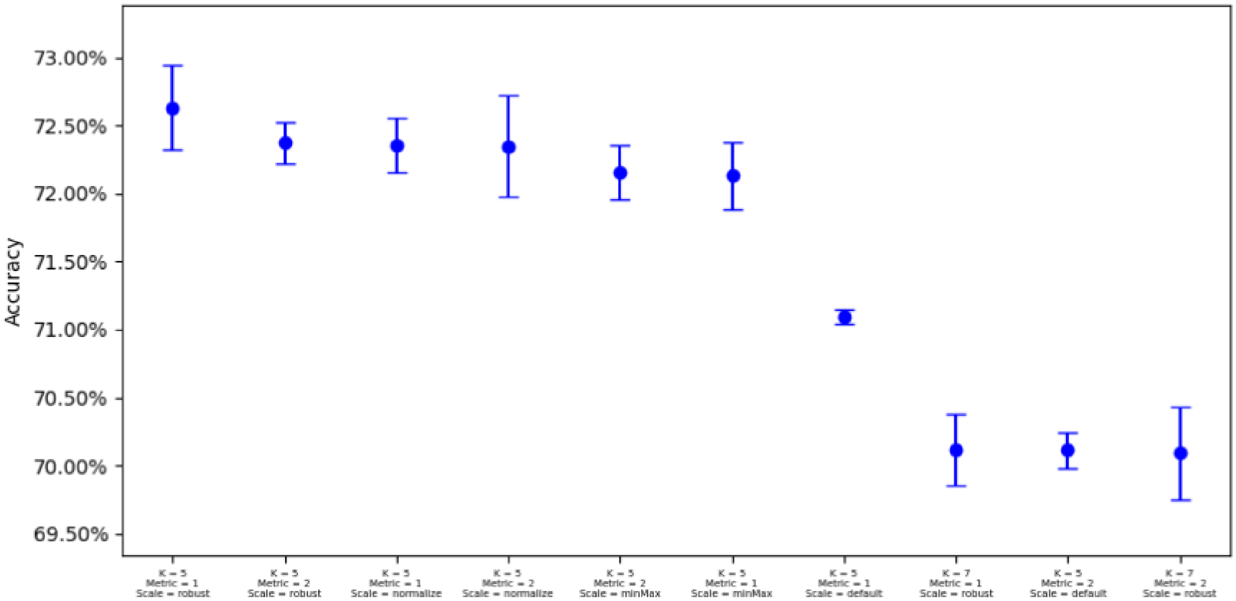


Figure 9: Top 10 KNN models with highest mean cross validation score - news dataset (training)

Cross validation results - Decision Tree

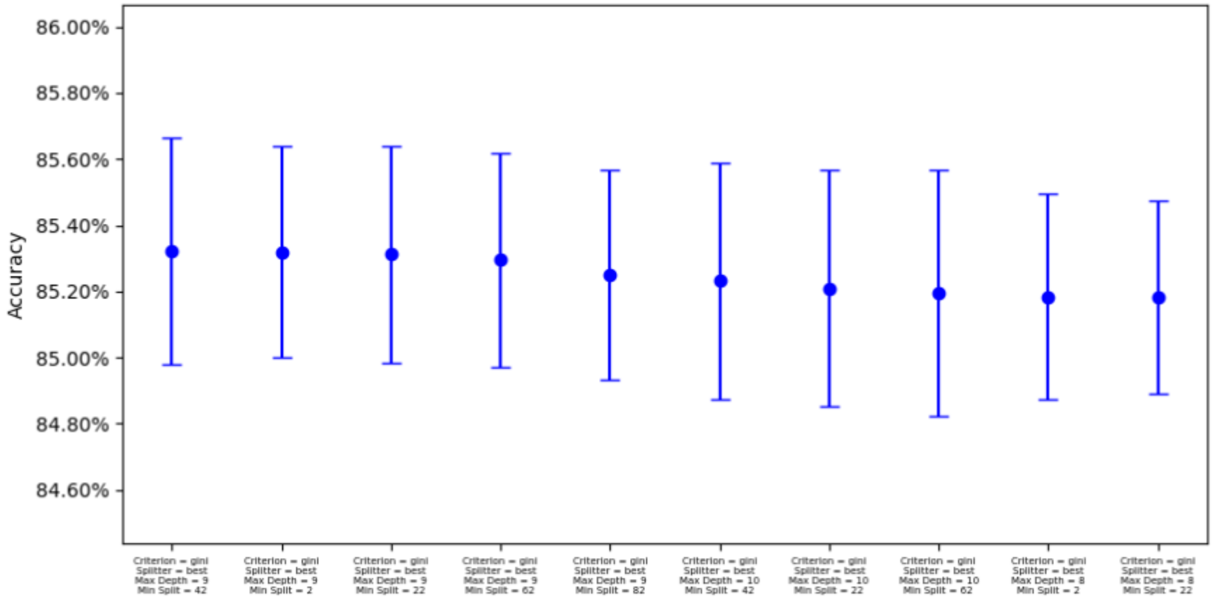


Figure 10: Top 10 decision tree models with highest mean cross validation score - adult dataset (validation)

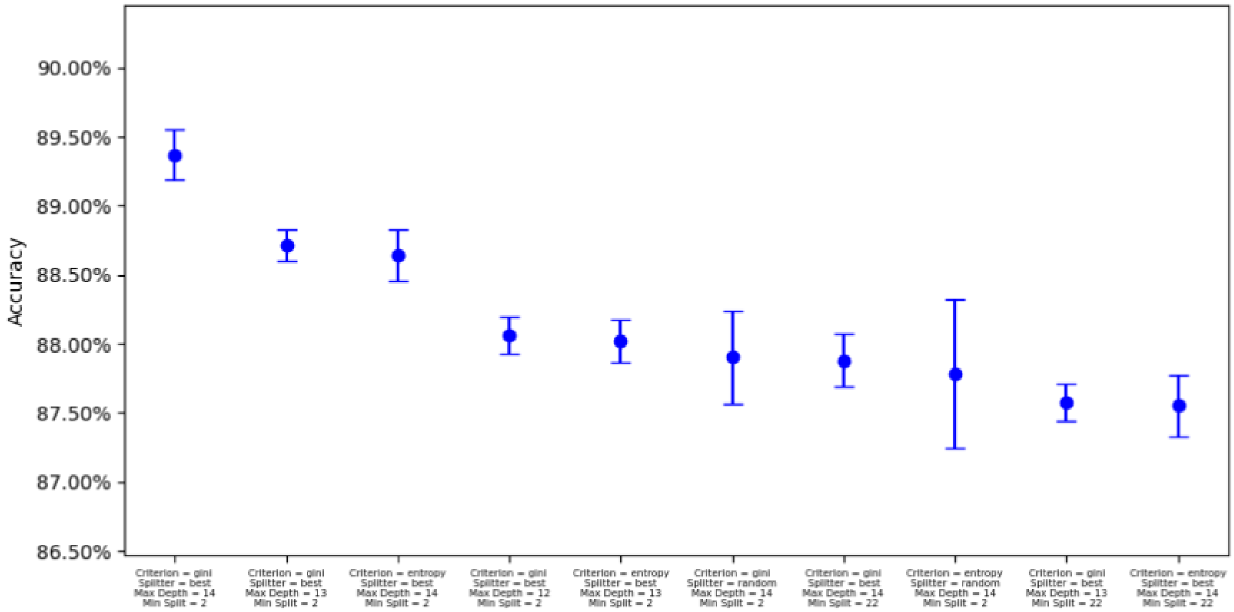


Figure 11: Top 10 decision tree models with highest mean cross validation score - adult dataset (training)

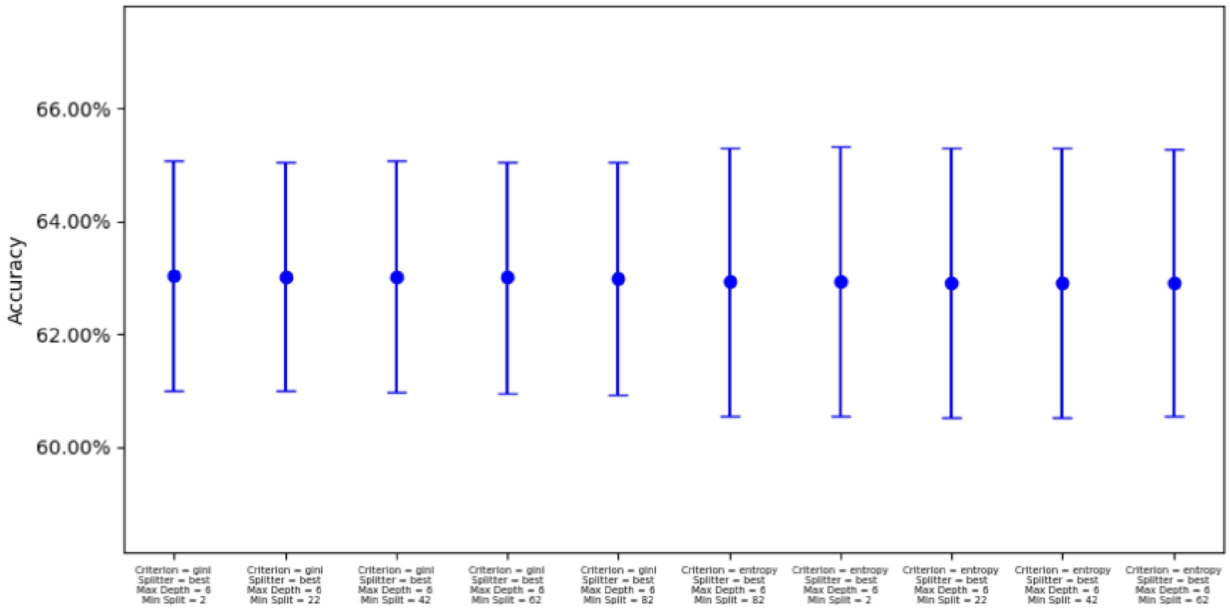


Figure 12: Top 10 decision tree models with highest mean cross validation score - news dataset (validation)

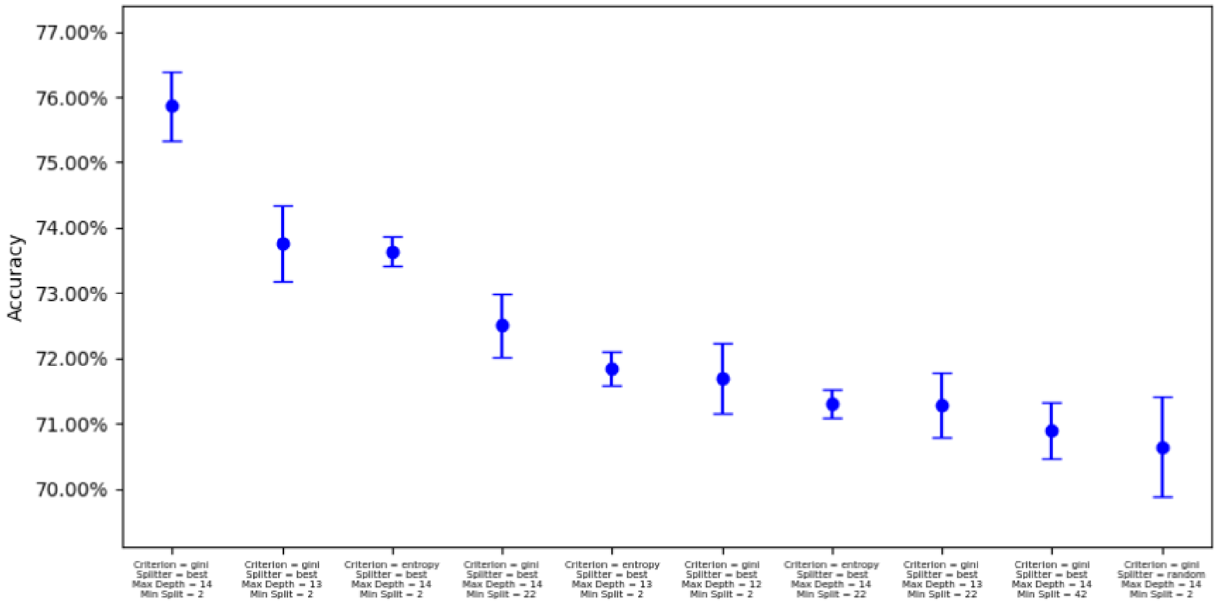


Figure 13: Top 10 decision tree models with highest mean cross validation score - news dataset (training)

	Predicted <=1400 shares	Predicted >1400 shares
Actual <=1400 shares	844	948
Actual >1400 shares	611	1561

Accuracy: 0.606710393541877

Figure 14: Confusion matrix for decision tree model on news test dataset

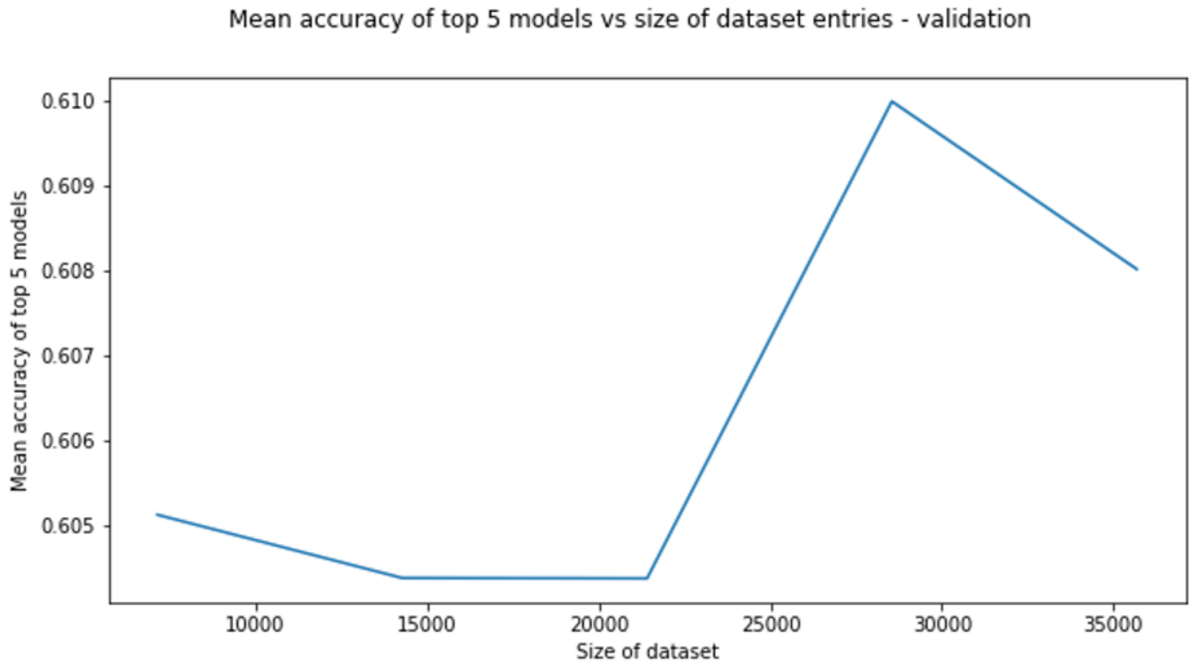


Figure 15: Accuracy for KNN algorithm on News dataset for different subsets with different sizes

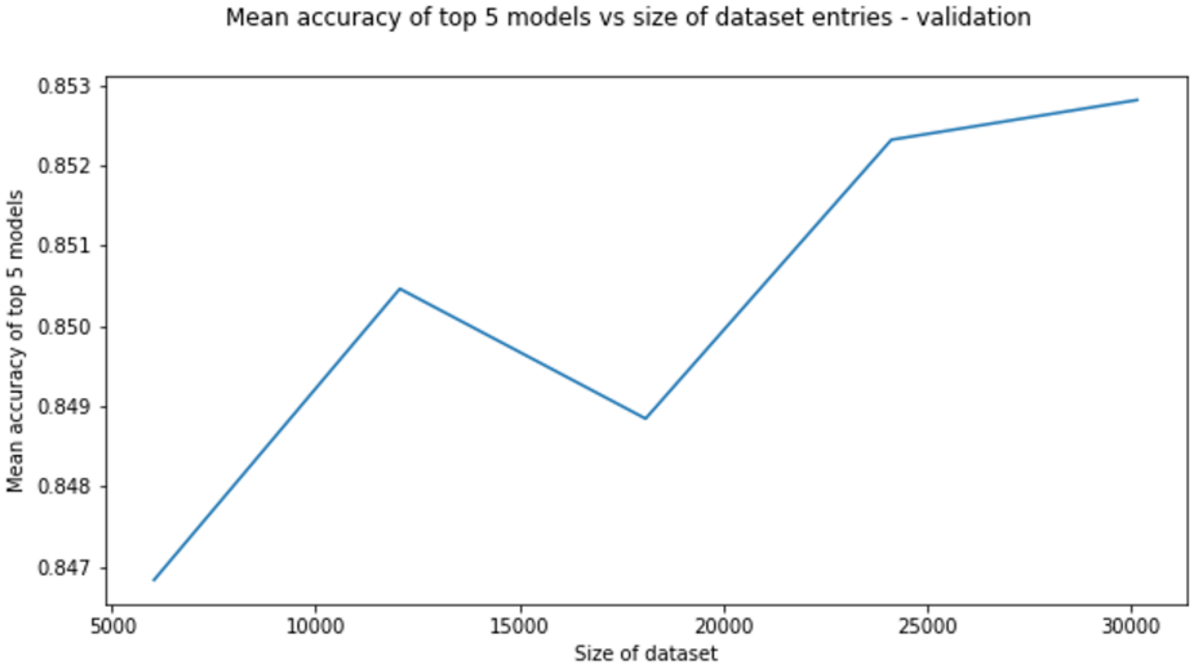


Figure 16: Accuracy for decision tree algorithm on Adult dataset for different subsets with different sizes

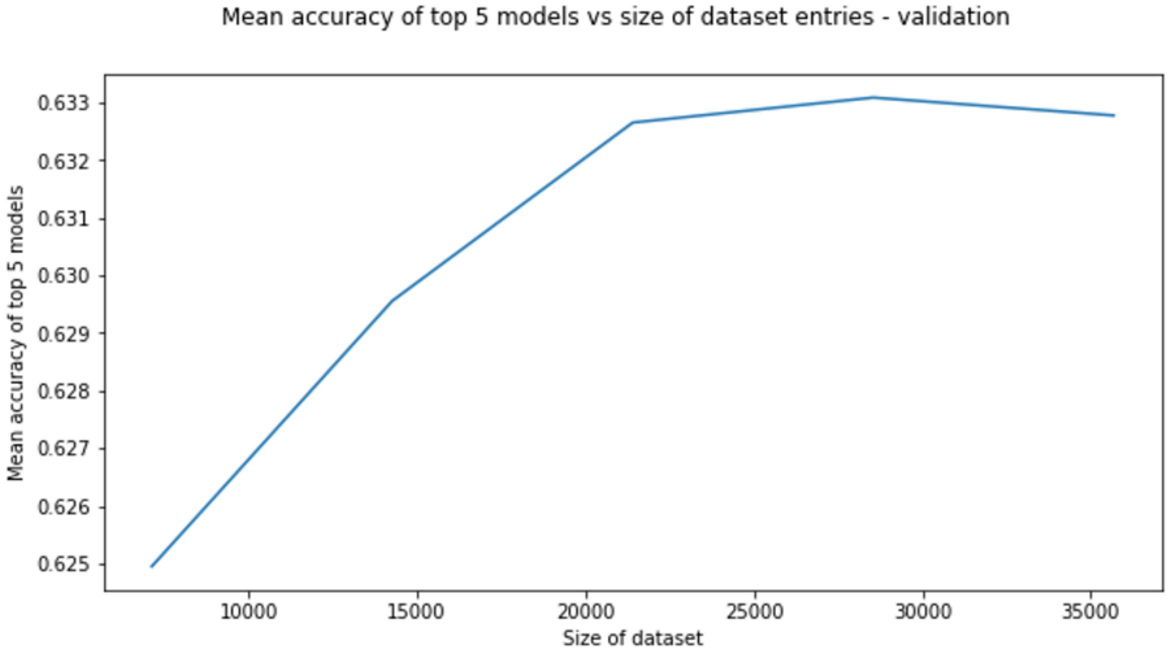


Figure 17: Accuracy for decision tree algorithm on News dataset for different subsets with different sizes