Machine learning Project Supplementary Material File

Mussa Yousef

Contents:

1) Glossary of Technical Terminologies used in order of occurrence:

	Terminology	Definition
1.	Computer visions	The study of how computer can contextualise, read and understand images digitally [1]
2.	Classification	Machine Learning- simplifies real like decision by building respective models; a type of method which computers use to predict is called 'classification' where the model is able to predict using class labels. [2]
3.	Algorithm	A mathematical sequence which implements instructions to the computer which solves machine learning problems and is an integral part in prediction.3]
4.	Ensemble	Ensemble model builds models and iterates the process to improve the model by bagging results and using the majority outcome as the official prediction for the Machine Learning model. *[4]*
5.	Physicochemical	Relating to the wine dataset used in this project the physicochemical is the combination of both physical and chemical elements within the structure of red wine. [5]
6.	Hyperparameters	A hyper-parameter is a parameters value which has been inspected and externally chosen to best optimise a model. A standard parameter us usually defined via training. [6]
7.	Correlation	Correlation is the statistic the investigates the relationship between two variables and how the move in relation to one another.[7]
8.	Recall	Recall is conserved to be the proportion of how many actual positives was identified correctly [8].
9.	F-score	In statistical analysis F-score is the accuracy of the test data being compared to the training data.[10]
10.	Regression	Regression like classification is a method which computers use to predict however this type is used to predict continuous outcomes variable (y) based on 1 or more predictors (x).
11.	Node/Root node	The root node is the initial node which the tree beings at and breaks into all possible outcomes. Each lead to additional nodes, which branch off into other possibilities. [12]
12.	Computational cost	This is the complexity cost of an algorithm is the consideration of the number of resources required to run the algorithm i.e., Time and Memory used. [13]
13.	Accuracy	Accuracy is the number of correctly predicted data points from all predicted variables.[14]
14.	Optimisation	Optimisation is essentially the process of making sure the Machine learning model uses the most efficient number of resources while performing at its best given the amount of information being processed. [15]
15.	Misclassification	Misclassification is the calculation of error which refers to the number of predictions which has been wrongly predicted as another label.
16.	Bayesian/Bayesian optimisation	Bayesian optimisation which is widely used in this report refers to the approach that uses Bayes Theorem to direct the search in order to find the minimum/maximum of an objective function. [17]
17.	Out-of-bag/Out-of-bag Error	Out-of-bag error, also called out-of-bag estimate, is a method of measuring the prediction error of the ensemble method of random forests [18]
18.	Aggregate/s	Bootstrap aggregating, also sometimes referred to as bagging, is an ensemble algorithm which is designed to enhance accuracy by iterating through the model via 'n' times while bagging outcomes in a tally and choosing the majority as the final outcome.
19.	Grid search	Grid-search is the process of looking through data and configuring optimal hyperparameters via indicators. [20]
20.	Minimum leaf size	This is the limit to split a node.[21]
21.	Loop search	The process of iterating a programming structure that repeats a sequence of instruction until a specific condition is met. [22]
22.	Confusion matrix	Summarises the performance of a classification algorithm by calculating the outcome of Actual and Predicted variables. [23]
23.	Cross validation	Cross-validation is the technique used to train the data set. Cross-validation is largely used in settings where the data is used to validate itself by splitting into folds and training on parts of the data while testing on another and vice versa. This process allows us to estimate the accuracy of the performance.[24]
24.	Skewed	Skewed data is common in data science, skew is the degree of distortion from normal distribution.
25.	Precision	Precision is the proportion of positive identifications which was actually correctly predicted.[9]
26.	PCA (Principal component analysis)	Principal Component Analysis (PCA) is a statistical procedure that converts correlated and uncorrelated variables. PCA is common tool to explore data analysis. [26]

Reference sourced to definition

[1] https://deepai.org/machine-learning-glossary-and-terms/computer-vision
[2] https://machinelearningmastery.com/types-of-classification-in-machine-learning/
[3] https://dictionary.cambridge.org/dictionary/english/algorithm
[4] https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2
[5] https://www.collinsdictionary.com/dictionary/english/physicochemical
[6] https://en.wikipedia.org/wiki/Hyperparameter_(machine_learning)
[7] https://www.investopedia.com/terms/c/correlation.asp
[8] https://en.wikipedia.org/wiki/Precision_and_recall
[9] https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall
[10] https://en.wikipedia.org/wiki/F-score
[11] http://www.sthda.com/english/wiki/regression-analysis-essentials-for-machine-learning
[12] https://towardsdatascience.com/machine-learning-basics-descision-tree-from-scratch-part-i-4251bfa1b45c
[13] https://en.wikipedia.org/wiki/Computational_complexity
[14] https://deepai.org/machine-learning-glossary-and-terms/accuracy-error-rate
[15] https://towardsdatascience.com/demystifying-optimizations-for-machine-learning-c6c6405d3eea
[16] https://www.researchgate.net/post/What-is-misclassification-error-and-what-algorithm-is-best-for-this
[17] https://machinelearningmastery.com/what-is-bayesian-optimization/
[18] https://en.wikipedia.org/wiki/Out-of-bag_error
[19] https://en.wikipedia.org/wiki/Bootstrap_aggregating
[20] https://elutins.medium.com/grid-searching-in-machine-learning-quick-explanation-and-python-implementation-550552200596
[21] https://zyxo.wordpress.com/2011/07/04/how-to-use-the-settings-to-control-the-size-of-decision-trees/
[22] https://techterms.com/definition/loop
[23] https://machinelearningmastery.com/confusion-matrix-machine-learning/
[24] https://www.techopedia.com/definition/32064/cross-validation
[25] https://medium.com/@ODSC/transforming-skewed-data-for-machine-learning-90e6cc364b0
[26] https://www.geeksforgeeks.org/ml-principal-component-analysispca/

2) Intermediate results including any negative results

- **We used feature importance when building and training our decision tree model, prior to this we looked at how the two highest correlated features influenced predictions. By plotting a scatter graph, we were able to visualise the dispersed prediction and investigate If there were any clusters.**

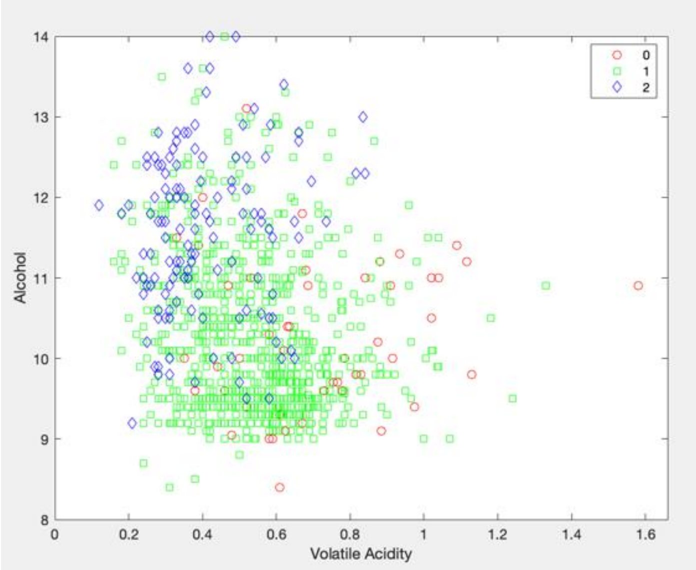

Fig.1: Scatter graph of label express by the distribution of the two highest correlated features.

- **Although generally there wasn't any overall correlations; inspecting the graph we find most good quality red wine (2) have an alcohol measurement between 9-13 while narrowly all have a very low volatile acidity.**

When manually tuning the decision tree we used 10-fold cross validation. Using this method, we found our results to be static. Using the loop, we generalised the best results by find the lowest point when looking at these individually. However, a huge drawback of this is that merging the optimal hyperparameters doesn't necessarily mean that the model built would provide best performance.

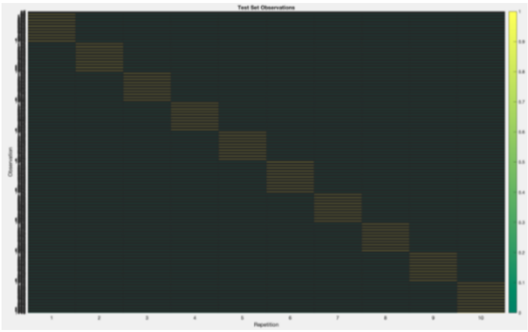

Fig.3: Visualisation of the 10-fold cross validation

Fig.4: Graph visualising the Cross-Validation Error vs Max Num Split

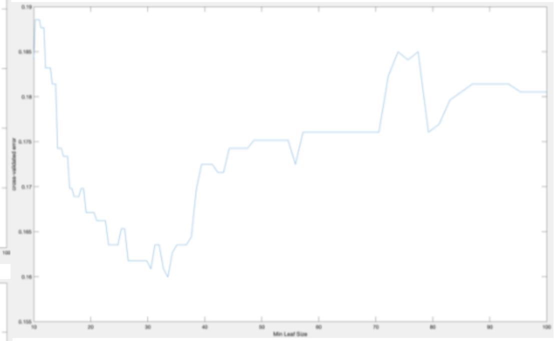

Fig.5: Graph visualising the Cross-Validation Error vs Min Leaf Size

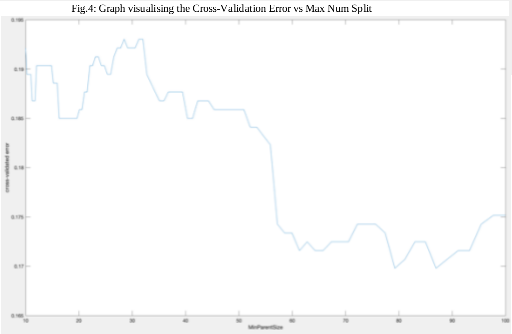

Fig.6: Graph visualising the Cross-Validation Error vs Minimum Parent Size

We felt the manual grid searched decision tree may have performed better if a combination of parameters were evaluated regards of their performance individually. Reflecting on results this explains the outperformance when comparing to the Bayesian optimised model.

Evaluating the Random Forest, we used the Out-of-bag error to evaluate the initial general performance of using a specific hyperparameter. Generally evaluating the ensemble of decision trees we found that fundamentally after 50 Trees we see the OOB error rate oscillate around 0.15, this indicates the optimal number of trees would be greater than 50.

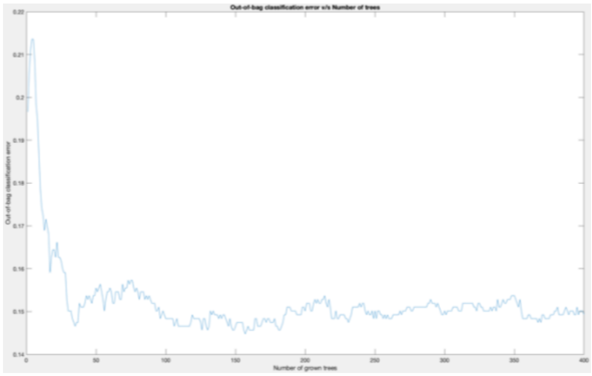

Fig.7: Graph visualising the Out-of-bag error vs Number of trees

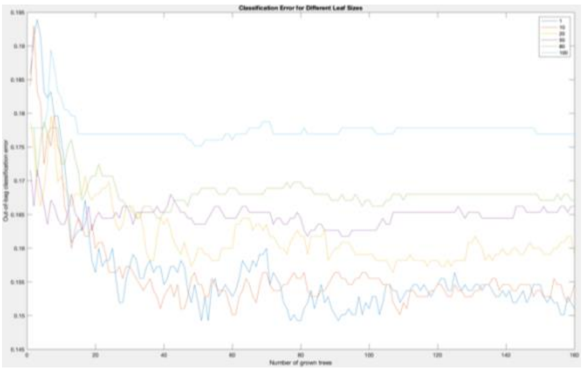

Fig.8: Graph visualising the Out-of-bag error vs Number of grown trees

Now dissecting the out-of-bag process we examined the parameter of the different sized leaf's as we grow the number of ensembled trees. We immediately comprehend that the higher the leaf size the greater the out-of-bag error. Therefore, limiting the leaf size was imperative when grid searching.

Machine learning Project Supplementary Material File
Mussa Yousef

3) **Brief description of main implementation (Schematic)**