Classification of Eucalyptus Species for Soil Conservation

Daniel Tai

CS699 - Dr. Jae Young Lee

Spring 2023 Term Project Report

Introduction

A study was conducted over twelve years in New Zealand, to assess different Eucalyptus species' effectiveness in reducing soil erosion simultaneously improving soil fertility. A total of twelve different Eucalyptus species were chosen, planted and observed for their soil conservation and productivity.

The objective of this project is to predict what level of utility the different Eucalyptus species can offer under different environmental conditions. There is a total of five feature selection methods - Chi Square Test, Learning Vector Quantization, Recursive Feature Elimination, Random Forest Importance, Information Gain and another five classification algorithms - Naive Bayes, K-Nearest Neighbours, J48 Decision Tree Algorithm, Rpart Decision Tree Algorithm and Support Vector Machines. Having multiple feature selection methods and classification algorithms allow us to accurately select features without human bias and accurately predict which species can preserve land and to what extent of each species' utility capabilities.

Dataset

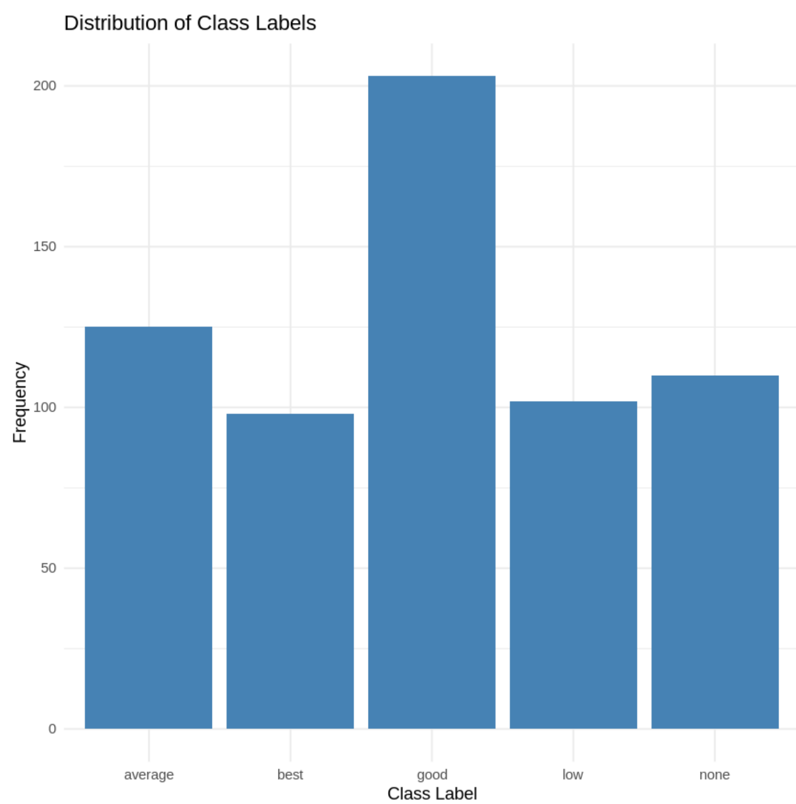
The dataset used for this project was from the twelfth and final year of assessment for the Eucalyptus species and its impact. B.T. Bulloch led the study that determined which Eucalyptus species was useful for soil conservation and soil fertility. The original unedited dataset consists of 738 tuples and 20 different attributes, where the last attribute, Utility is the class attribute. The class attribute consists of five different classifications - None, Low, Average, Good and Best. Different Eucalyptus species will be placed in these classifications depending on the reduced set of features from different feature selection methods and the different classification algorithms. Below is a detailed explanation of each attribute:

1. Abbrev: Site Abbreviation (Cly=Clydebank, Cra=Craggy Range Road, ...)
2. Rep: Number of Experimental Replications
3. Locality: Site location in North Island (Central Hawkes Bay, Northern Hawkes Bay, ...)
4. Map_Ref: Map Location in North Island
5. Latitude: Latitude Approximation South(degrees__minutes) = South(39__38)
6. Altitude: Unit (m)
7. Rainfall: Unit (mm pa)
8. Frosts: Unit (Degrees C)
9. Year: Year of Planting
10. Sp: Species (Enumerated)

11. PMCno: Seedlot Number, a unique number for specific species with specific quantity and quality
12. DBH: Diameter at Breast Height, Unit (cm)
13. Ht: Height, Unit (m)
14. Surv: Survival Percentage
15. Vig: Vigor(Health and Resilience)
16. Ins_res: Insect Resistance
17. Stem_Fm: Stem Form
18. Crown_Fm: Crown Form
19. Brnch_Fm: Branch Form
20. Utility: Determined by None, Low, Average, Good, Best

Data Visualization

It is important to understand the distribution of the class attribute values. A bar plot illustrates the distribution of the values in a relatively even manner with no severe anomalies or outlier labels with the exception of a significant frequency for the class label 'good'. This is a crucial observation and a good starting point for the dataset as an unbalanced dataset may affect the implementation of the classification models.



Data Cleaning

Data cleaning is one of the few crucial steps to ensure our models make accurate predictions. Below are the few key areas of data cleaning.

1. Missing Data: The initial dataset had invalid entries containing '?' in columns *PMCno*, *Surv*, *Vig*, *Ins_Res*, *Stem_Fm*, *Crown_Fm* and *Brnch_Fm*. As a result, all rows affected by validity were removed. The measures of central tendency of the data(mean, median and mode) was a potential option for replacement. However, given that each row represents a specific specimen subjected to specific conditions in different regions of New Zealand, replacing the missing values could potentially produce inaccuracy at a greater scale.
2. Factor Levels: Previously, it was mentioned that several entries were removed due to missing data. This also affected some of the type factor features. The dataset was initially of type character, we unclassed it and set all strings to factors. This produced unintended levels in certain categorical attributes which have zero entries when we called the 'na.omit' function. Therefore, attributes *Locality*, *Abbrev*, *Sp*, *PMCno*, *Surv*, *Vig*, *Ins_Res*, *Stem_Fm*, *Crown_Fm* and *Brnch_Fm* had to be called in the 'droplevels' function.
3. Outliers: *PMCno* were expected to be a four digit integer, there were several *PMCno* with a value of '1'. All rows affected by this outlier were removed instead of replaced for accuracy purposes, as replacing with most frequent seedlot number could also affect the classification algorithms
4. Irrelevant Data: Two features *Rep* and *Latitude* were completely removed from this project due to the irrelevancy and complexity in predicting the utility level of classification respectively.

Data Preprocessing

Upon inspecting the summary of the cleaned data, a number attributes needed to be converted to numeric as they were continuous variables and categorizing them as factors would be incorrect and can potentially damage the accuracy when using the models for prediction.

Feature Selection Methods

There are five different feature selection methods implemented in this project to best determine a subset of features in predicting whether certain Eucalyptus species with different conditions can be used for soil conservation. There are three types of feature selection methods - filter, wrapper and embedded techniques. In this project, only filter and wrapper techniques were used along with a modified version of a supervised classification algorithm. Below are the five feature selection methods:

1. Chi-Square Test: The Chi-Square test is used to determine the dependency between a feature and the class attribute. A high chi-squared value would indicate that the two variables have a strong relationship and should be preserved. On the other hand, low chi-squared value would indicate their independence and should be discarded. Here, one of the variables will always be the class attribute and the other would be a set of features of which each one is evaluated on their dependency relationship with the class attribute.
2. Learning Vector Quantization: Learning Vector Quantization (L.V.Q.) is a supervised machine learning algorithm from the family of artificial neural networks. For each class label, a set of random prototype vectors are generated. At each iteration, the LVQ model adjusts the vectors to reduce the classification error. The variable importance is calculated for each feature with respect to each class label and an average importance score is determined and a cutoff is set for subsetting the attributes.
3. Recursive Feature Elimination: The Recursive Feature Selection fits a model and ranks the features and recursively eliminates the least important features until a specified amount. In this project, the `predictors` function from the ‘care’ package is used to identify the final set of variables after using the R.F.E. model on the entire dataset with all features.
4. Random Forest Importance: This method is from the ‘FSelector’ package under ‘random.forest.importance’ function. This is different compared to the ‘randomForest’ package which is used for training Random Forest models. It is important to note that a Random Forest model is used here to determine the importance scores but returns a list of features and their ranking instead of the random forest model itself.
5. Information Gain: Information Gain measures the reduction in entropy (uncertainty) of a class attribute against a selected feature. It is usually paired

with decision tree algorithms to calculate the best split cotcome. In this project, a predetermined percentile was set and all features whose information gain scores are less than the percentile will be filtered out.

Classification Algorithms

There are five different feature selection methods implemented in this project to best determine a subset of features in predicting whether certain Eucalyptus species with different conditions can be used for soil conservation. There are three types of feature selection methods - filter, wrapper and embedded techniques. In this project, only filter and wrapper techniques were used along with a modified version of a supervised classification algorithm. Below are the five feature selection methods:

1. Naive Bayes: The Naive Bayes model uses statistical probability to predict a class outcome given a set of features. It assumes feature independence, making it efficient for large datasets. Naive Bayes applies the Bayes Theorem of conditional probability to calculate the probability of each possible class outcome. The outcome with the highest probability is the selected label for prediction.
2. K-Nearest Neighbours: The K-Nearest Neighbours (KNN) is a simple and intuitive classification algorithm that predicts a class label based on the majority class among its K nearest neighbors in the attribute space.In R, the "train" function with "method = 'knn'" and "preProcess" for data preprocessing, "tuneLength" for hyperparameter tuning, and "trainControl" for cross-validation are commonly used for KNN model training and evaluation.
3. J48 Decision Tree: J48 is a machine learning decision tree classification algorithm based on Iterative Dichotomiser 3 to examine the data categorically and continuously. It builds a tree by recursively partitioning the data based on attribute values. In R, the "train" function with "method = 'J48'" and "trainControl" using the "repeatedcv" method, along with "tuneGrid" for hyperparameter tuning, are commonly used.
4. Rpart Decision Tree: Rpart is used for building classification and regression trees using the Recursive Partitioning and Regression Trees (RPART) algorithm. This library implements recursive partitioning and is very easy to use. It is widely used for predictive modeling and data mining tasks, providing interpretable and easy-to-understand tree structures for classification and regression problems.
5. Neural Network: The neural network is a simplified model of the way the human brain processes information. It works by simulating a large number of interconnected processing units that resemble abstract versions of neurons. The

processing units are arranged in layers. It employs an iterative process to optimize the network weights and biases, aiming to minimize the prediction error and maximize model accuracy.

Results

- **Chi-Square Test**
    - Naive Bayes

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3636364	0.190751445	0.3265306	0.3636364	0.3440860	0.6278245	0.1662470
best	0.1333333	0.005813953	0.8571429	0.1333333	0.2307692	0.8156331	0.2926092
good	0.7066667	0.295774648	0.5578947	0.7066667	0.6235294	0.7571831	0.3938772
low	0.5200000	0.125000000	0.3513514	0.5200000	0.4193548	0.8593750	0.3353363
none	0.6428571	0.058201058	0.6206897	0.6428571	0.6315789	0.9130763	0.5760132
Weighted averages	0.4732987	0.1351082208	0.54272186	0.4732987	0.44986366	0.7946184	0.35281658

Confusion Matrix					
Prediction	average	best	good	low	none
average	16	10	15	5	3
best	0	6	1	0	0
good	13	26	53	1	2
low	11	2	6	13	5
none	4	1	0	6	18

- KNN Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.1363636	0.06358382	0.3529412	0.1363636	0.1967213	0.4799531	0.1088983
best	0.2222222	0.02906977	0.6666667	0.2222222	0.3333333	0.7266150	0.3087091
good	0.8800000	0.48591549	0.4888889	0.8800000	0.6285714	0.4998686	0.3865364
low	0.2400000	0.06770833	0.3157895	0.2400000	0.2727273	0.6954837	0.1946146
none	0.6071429	0.07407407	0.5483871	0.6071429	0.5762712	0.7254167	0.5106882
Weighted averages	0.41714574	0.144070296	0.47453468	0.41714574	0.4015249	0.62546742	0.30188932

Confusion Matrix					
Prediction	average	best	good	low	none
average	6	3	2	4	2
best	0	10	3	2	0
good	27	28	66	7	7
low	7	2	2	6	2
none	4	2	2	6	17

- J48 Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3409091	0.13872832	0.3846154	0.3409091	0.3614458	0.4750235	0.2117127
best	0.2444444	0.04651163	0.5789474	0.2444444	0.3437500	0.7818475	0.2839088
good	0.6533333	0.30281690	0.5326087	0.6533333	0.5868263	0.5285733	0.3373139
low	0.6400000	0.14062500	0.3720930	0.6400000	0.4705882	0.7012472	0.3999803
none	0.7500000	0.01587302	0.8750000	0.7500000	0.8076923	0.6778125	0.7846877
Weighted averages	0.52573736	0.128910974	0.5486529	0.52573736	0.51406052	0.6329008	0.40352068

Confusion Matrix					
Prediction	average	best	good	low	none
average	15	5	14	4	1
best	0	11	8	0	0
good	14	27	49	2	0
low	15	2	4	16	6
none	0	0	0	3	21

- rpart Decision Tree Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3636364	0.12138728	0.4324324	0.3636364	0.3950617	0.5578873	0.2589849
best	0.4222222	0.07558140	0.5937500	0.4222222	0.4935065	0.8175711	0.3963596
good	0.6266667	0.27464789	0.5465116	0.6266667	0.5838509	0.5154362	0.3422597
low	0.6800000	0.11458333	0.4358974	0.6800000	0.5312500	0.6354875	0.4701613
none	0.7500000	0.01058201	0.9130435	0.7500000	0.8235294	0.7923958	0.8052582
Weighted averages	0.56850506	0.119356382	0.58432698	0.56850506	0.5654397	0.66375558	0.45460474

Confusion Matrix					
Prediction	average	best	good	low	none
average	16	5	10	5	1
best	0	19	13	0	0
good	17	21	47	1	0
low	11	0	5	17	6
none	0	0	0	2	21

- Neural Network Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3636364	0.08092486	0.5333333	0.3636364	0.4324324	0.5685446009389	0.3293149497081
best	0.3555556	0.09883721	0.4848485	0.3555556	0.4102564	0.8162790697674	0.2898421251487
good	0.6133333	0.24647887	0.5679012	0.6133333	0.5897436	0.6044403573305	0.3607087219593
low	0.7200000	0.14062500	0.4000000	0.7200000	0.5142857	0.7870370370370	0.4562573669363
none	0.6785714	0.04761905	0.6785714	0.6785714	0.6785714	0.8310416666666	0.6309523809523
Weighted averages	0.54621934	0.122896998	0.53293088	0.54621934	0.5250579	0.7214685463	0.4134151089

Confusion Matrix					
Prediction	average	best	good	low	none
average	16	2	11	1	0
best	2	16	15	0	0
good	9	26	46	0	0
low	14	1	3	18	9
none	3	0	0	6	19

- **Learning Vector Quantization**
  - Naive Bayes

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3863636	0.19653179	0.3333333	0.3863636	0.3578947	0.6831319	0.1800028
best	0.0000000	0.00000000	NaN	0.0000000	NaN	0.8087855	NaN
good	0.7600000	0.33098592	0.5480769	0.7600000	0.6368715	0.7430986	0.4084044
low	0.4400000	0.09895833	0.3666667	0.4400000	0.4000000	0.8908333	0.3154617
none	0.8571429	0.04232804	0.7500000	0.8571429	0.8000000	0.9550265	0.7703853
Weighted averages	0.4887013	0.133760816	0.5549145333	0.4887013	0.54869155	0.81617516	0.41856355

Confusion Matrix					
Prediction	average	best	good	low	none
average	17	10	15	6	3
best	0	0	0	0	0
good	12	35	57	0	0
low	15	0	3	11	1
none	0	0	0	0	24

- KNN Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3863636	0.12138728	0.4473684	0.3863636	0.4146341	0.4883099	0.2803097
best	0.2444444	0.00000000	1.0000000	0.2444444	0.3928571	0.8173127	0.4517734
good	0.8133333	0.31690141	0.5754717	0.8133333	0.6740331	0.6171834	0.4723023
low	0.5600000	0.11458333	0.3888889	0.5600000	0.4590164	0.7933673	0.3822932
none	0.7142857	0.03174603	0.7692308	0.7142857	0.7407407	0.7929167	0.7045867
Weighted averages	0.5436854	0.11692361	0.63619196	0.5436854	0.53625628	0.701818	0.45825306

Confusion Matrix					
Prediction	average	best	good	low	none
average	17	3	11	5	2
best	0	11	0	0	0
good	12	31	61	0	2
low	15	0	3	14	4
none	0	0	0	6	20

- J48 Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3409091	0.13294798	0.3947368	0.3409091	0.3658537	0.5463850	0.2199952
best	0.2666667	0.04651163	0.6000000	0.2666667	0.3692308	0.8021318	0.3085681
good	0.6400000	0.31690141	0.5161290	0.6400000	0.5714286	0.5518917	0.3104971
low	0.5600000	0.08854167	0.4516129	0.5600000	0.5000000	0.7282691	0.4301566
none	0.8571429	0.05820106	0.6857143	0.8571429	0.7619048	0.7089583	0.7282078
Weighted averages	0.53294374	0.12862075	0.5296386	0.53294374	0.51368358	0.66752718	0.39948496

Confusion Matrix					
Prediction	average	best	good	low	none
average	15	4	16	3	0
best	1	12	7	0	0
good	15	28	48	0	2
low	12	0	3	14	2
none	1	1	1	8	24

- rpart Decision Tree Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3636364	0.12716763	0.4210526	0.3636364	0.3902439	0.5427700	0.2501525
best	0.3555556	0.06976744	0.5714286	0.3555556	0.4383562	0.7593669	0.3456247
good	0.6133333	0.30281690	0.5168539	0.6133333	0.5609756	0.5730426	0.3002337
low	0.4400000	0.07291667	0.4400000	0.4400000	0.4400000	0.6947279	0.3670833
none	0.8214286	0.07407407	0.6216216	0.8214286	0.7076923	0.7171875	0.6661921
Weighted averages	0.51879078	0.129348542	0.51419134	0.51879078	0.5074536	0.65741898	0.38585726

Confusion Matrix					
Prediction	average	best	good	low	none
average	16	4	14	4	0
best	1	16	11	0	0
good	16	22	46	2	3
low	10	0	2	11	2
none	1	3	2	8	23

- Neural Network Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.2727273	0.09826590	0.4137931	0.2727273	0.3287671	0.5389671361502	0.2061440298277
best	0.3333333	0.02906977	0.7500000	0.3333333	0.4615385	0.8268733850129	0.4264541255951
good	0.7866667	0.31690141	0.5673077	0.7866667	0.6592179	0.5488702049395	0.4471979536773
low	0.6000000	0.11979167	0.3947368	0.6000000	0.4761905	0.7694633408919	0.4033963805747
none	0.7142857	0.03174603	0.7692308	0.7142857	0.7407407	0.8352083333333	0.7045867409340
Weighted averages	0.5414026	0.119154956	0.57901368	0.5414026	0.53329094	0.7038764801	0.4375558461

Confusion Matrix					
Prediction	average	best	good	low	none
average	12	3	9	3	2
best	0	15	5	0	0
good	17	27	59	1	0
low	15	0	2	15	6
none	0	0	0	6	20

- **Recursive Feature Elimination**
    - Naive Bayes

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.4090909	0.16184971	0.3913043	0.4090909	0.4000000	0.6879926	0.2432166
best	0.0000000	0.00000000	NaN	0.0000000	NaN	0.8620155	NaN
good	0.7466667	0.35915493	0.5233645	0.7466667	0.6153846	0.7528638	0.3686137
low	0.4800000	0.11979167	0.3428571	0.4800000	0.4000000	0.8827083	0.3126833
none	0.7500000	0.04232804	0.7241379	0.7500000	0.7368421	0.9676871	0.6972106
Weighted averages	0.47715152	0.13662487	0.49541595	0.47715152	0.538056675	0.83065346	0.40543105

Confusion Matrix					
Prediction	average	best	good	low	none
average	18	6	14	5	3
best	0	0	0	0	0
good	13	37	56	0	1
low	13	2	5	12	3
none	0	0	0	8	21

- KNN Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3181818	0.11560694	0.4117647	0.3181818	0.3589744	0.5507042	0.2240629
best	0.3333333	0.04651163	0.6521739	0.3333333	0.4411765	0.8286822	0.3777615
good	0.6933333	0.29577465	0.5531915	0.6933333	0.6153846	0.5511692	0.3815571
low	0.6400000	0.12500000	0.4000000	0.6400000	0.4923077	0.7342215	0.4240443
none	0.6071429	0.04761905	0.6538462	0.6071429	0.6296296	0.6636458	0.5775973
Weighted averages	0.51839826	0.126102454	0.53419526	0.51839826	0.50749456	0.66568458	0.39700462

Confusion Matrix					
Prediction	average	best	good	low	none
average	14	4	12	3	1
best	0	15	8	0	0
good	15	25	52	0	2
low	12	1	3	16	8
none	3	0	0	6	17

- J48 Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.2272727	0.10404624	0.3571429	0.2272727	0.2777778	0.6123944	0.1477895
best	0.2888889	0.04651163	0.6190476	0.2888889	0.3939394	0.7109173	0.3323722
good	0.7066667	0.34507042	0.5196078	0.7066667	0.5988701	0.5167499	0.3445480
low	0.6400000	0.15625000	0.3478261	0.6400000	0.4507042	0.6232993	0.3778893
none	0.5714286	0.02116402	0.8000000	0.5714286	0.6666667	0.6270833	0.6377249
Weighted averages	0.48685138	0.134608462	0.52872488	0.48685138	0.47759164	0.61808884	0.36806478

Confusion Matrix					
Prediction	average	best	good	low	none
average	10	3	10	5	0
best	0	13	8	0	0
good	19	29	53	1	0
low	14	0	4	16	12
none	1	0	0	3	16

- rpart Decision Tree Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.5681818	0.19653179	0.4237288	0.5681818	0.4854369	0.5653991	0.3358374
best	0.2888889	0.01744186	0.8125000	0.2888889	0.4262295	0.8272610	0.4211119
good	0.6266667	0.24647887	0.5731707	0.6266667	0.5987261	0.5432869	0.3729058
low	0.5200000	0.06250000	0.5200000	0.5200000	0.5200000	0.6876417	0.4575000
none	0.9285714	0.04761905	0.7428571	0.9285714	0.8253968	0.7518750	0.8029576
Weighted averages	0.58646176	0.114114314	0.61445132	0.58646176	0.57115786	0.67509274	0.47806254

Confusion Matrix					
Prediction	average	best	good	low	none
average	25	8	23	3	0
best	0	13	3	0	0
good	10	24	47	1	0
low	9	0	2	13	2
none	1	0	0	8	26

- Neural Network Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3863636	0.06358382	0.6071429	0.3863636	0.4722222	0.5249765258215	0.3871202137649
best	0.3555556	0.06395349	0.5925926	0.3555556	0.4444444	0.8854005167958	0.3581808775731
good	0.7466667	0.27464789	0.5894737	0.7466667	0.6588235	0.5881502890173	0.4524726814987
low	0.6000000	0.10416667	0.4285714	0.6000000	0.5000000	0.7843915343915	0.4304142310105
none	0.7500000	0.05820106	0.6562500	0.7500000	0.7000000	0.820625	0.6540771203607
Weighted averages	0.56771718	0.112910586	0.57480612	0.56771718	0.55509802	0.7207087732	0.4564530248

Confusion Matrix					
Prediction	average	best	good	low	none
average	17	1	6	3	1
best	0	16	11	0	0
good	11	28	56	0	0
low	12	0	2	15	6
none	4	0	0	7	21

- **Random Forest Importance**
    - Naive Bayes

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.4090909	0.17919075	0.3673469	0.4090909	0.3870968	0.6851025	0.2210731
best	0.0000000	0.00000000	NaN	0.0000000	NaN	0.8454780	NaN
good	0.7066667	0.33802817	0.5247525	0.7066667	0.6022727	0.7769953	0.3514680
low	0.5600000	0.10416667	0.4117647	0.5600000	0.4745763	0.8893750	0.4003702
none	0.7857143	0.05820106	0.6666667	0.7857143	0.7213115	0.9380197	0.6791801
Weighted averages	0.49229438	0.13591733	0.4926327	0.49229438	0.546314325	0.8269941	0.41302285

Confusion Matrix					
Prediction	average	best	good	low	none
average	18	9	19	3	0
best	0	0	0	0	0
good	12	34	53	1	1
low	10	2	3	14	5
none	4	0	0	7	22

- KNN Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.2954545	0.19653179	0.2765957	0.2954545	0.2857143	0.4874648	0.09655443
best	0.2666667	0.05813953	0.5454545	0.2666667	0.3582090	0.7771964	0.28009439
good	0.6933333	0.40845070	0.4727273	0.6933333	0.5621622	0.5193773	0.27098955
low	0.2000000	0.03125000	0.4545455	0.2000000	0.2777778	0.8236017	0.24560344
none	0.5000000	0.06878307	0.5185185	0.5000000	0.5090909	0.7751042	0.43797270
Weighted averages	0.3910909	0.152631018	0.4535683	0.3910909	0.39859084	0.67654888	0.266242902

Confusion Matrix					
Prediction	average	best	good	low	none
average	13	8	12	9	5
best	0	12	9	1	0
good	26	25	52	2	5
low	2	0	0	5	4
none	3	0	2	8	14

- J48 Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.4090909	0.15606936	0.4000000	0.4090909	0.4044944	0.6178404	0.2509207
best	0.3333333	0.05232558	0.6250000	0.3333333	0.4347826	0.7640827	0.3632491
good	0.5733333	0.30281690	0.5000000	0.5733333	0.5341615	0.5001314	0.2630169
low	0.7200000	0.12500000	0.4285714	0.7200000	0.5373134	0.6575019	0.4808326
none	0.6071429	0.01587302	0.8500000	0.6071429	0.7083333	0.6746875	0.6852476
Weighted averages	0.52858008	0.130416972	0.56071428	0.52858008	0.52381704	0.64284878	0.40865338

Confusion Matrix					
Prediction	average	best	good	low	none
average	18	3	22	2	0
best	1	15	8	0	0
good	16	25	43	2	0
low	9	2	2	18	11
none	0	0	0	3	17

- rpart Decision Tree Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3636364	0.10404624	0.4705882	0.3636364	0.4102564	0.6058216	0.2871260
best	0.2888889	0.01744186	0.8125000	0.2888889	0.4262295	0.8285530	0.4211119
good	0.7866667	0.33802817	0.5514019	0.7866667	0.6483516	0.5202312	0.4267594
low	0.5200000	0.06250000	0.5200000	0.5200000	0.5200000	0.6655329	0.4575000
none	0.9285714	0.04761905	0.7428571	0.9285714	0.8253968	0.7518750	0.8029576
Weighted averages	0.57755268	0.113927064	0.61946944	0.57755268	0.56604686	0.67440274	0.47909098

Confusion Matrix					
Prediction	average	best	good	low	none
average	16	4	11	3	0
best	0	13	3	0	0
good	19	28	59	1	0
low	8	0	2	13	2
none	1	0	0	8	26

- Neural Network Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.4090909	0.08092486	0.5625000	0.4090909	0.4736842	0.5630046948356	0.3721192961804
best	0.4666667	0.09883721	0.5526316	0.4666667	0.5060241	0.8645994832041	0.3923726006430
good	0.6133333	0.21830986	0.5974026	0.6133333	0.6052632	0.5704151339989	0.3926343826377
low	0.6800000	0.11458333	0.4358974	0.6800000	0.5312500	0.7834467120181	0.4701613448540
none	0.7500000	0.05291005	0.6774194	0.7500000	0.7118644	0.8360416666666	0.6678230711206
Weighted averages	0.58381818	0.113113062	0.5651702	0.58381818	0.56561718	0.7235015381	0.4590221391

Confusion Matrix					
Prediction	average	best	good	low	none
average	18	2	10	2	0
best	0	21	17	0	0
good	11	20	46	0	0
low	11	2	2	17	7
none	4	0	0	6	21

- **Information Gain**
    - Naive Bayes

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.4545455	0.17341040	0.4000000	0.4545455	0.4255319	0.6336048	0.2684239
best	0.0000000	0.00000000	NaN	0.0000000	NaN	0.7678295	NaN
good	0.6933333	0.40140845	0.4770642	0.6933333	0.5652174	0.7078873	0.2776648
low	0.4800000	0.08333333	0.4285714	0.4800000	0.4528302	0.8627083	0.3777778
none	0.7500000	0.04761905	0.7000000	0.7500000	0.7241379	0.9263039	0.6821835
Weighted averages	0.47557576	0.141154246	0.5014089	0.47557576	0.54192935	0.77966676	0.4015125

Confusion Matrix					
Prediction	average	best	good	low	none
average	20	3	20	6	1
best	0	0	0	0	0
good	13	40	52	2	2
low	8	2	2	12	4
none	3	0	1	5	21

- KNN Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3863636	0.19653179	0.3333333	0.3863636	0.3578947	0.5550704	0.1800028
best	0.1333333	0.05232558	0.4000000	0.1333333	0.2000000	0.6868217	0.1294720
good	0.5733333	0.38028169	0.4432990	0.5733333	0.5000000	0.5543221	0.1846596
low	0.4000000	0.09895833	0.3448276	0.4000000	0.3703704	0.7487717	0.2824680
none	0.3928571	0.07407407	0.4400000	0.3928571	0.4150943	0.6929167	0.3347222
Weighted averages	0.37717746	0.160434292	0.39229198	0.37717746	0.36867188	0.64758052	0.22226492

Confusion Matrix					
Prediction	average	best	good	low	none
average	17	6	19	7	2
best	0	6	8	0	1
good	16	30	43	3	5
low	7	1	2	10	9
none	4	2	3	5	11

- J48 Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.4090909	0.14450867	0.4186047	0.4090909	0.4137931	0.4985915	0.2668709
best	0.1555556	0.04651163	0.4666667	0.1555556	0.2333333	0.7881137	0.1742812
good	0.6400000	0.32394366	0.5106383	0.6400000	0.5680473	0.5874277	0.3033352
low	0.6800000	0.12500000	0.4146341	0.6800000	0.5151515	0.6813114	0.4526529
none	0.6428571	0.03174603	0.7500000	0.6428571	0.6923077	0.6787500	0.6531995
Weighted averages	0.50550072	0.134341998	0.51210876	0.50550072	0.48452658	0.64683886	0.37006794

Confusion Matrix					
Prediction	average	best	good	low	none
average	18	2	15	6	2
best	1	7	7	0	0
good	11	34	48	0	1
low	12	2	3	17	7
none	2	0	2	2	18

- rpart Decision Tree Model

	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.3636364	0.07514451	0.5517241	0.3636364	0.4383562	0.5625352	0.3408827
best	0.2000000	0.06395349	0.4500000	0.2000000	0.2769231	0.7801680	0.1906820
good	0.7200000	0.38028169	0.5000000	0.7200000	0.5901639	0.5592486	0.3231236
low	0.4000000	0.07812500	0.4000000	0.4000000	0.4000000	0.7020975	0.3218750
none	0.7857143	0.06878307	0.6285714	0.7857143	0.6984127	0.7283333	0.6534580
Weighted averages	0.49387014	0.133257552	0.5060591	0.49387014	0.48077118	0.66647652	0.36600426

Confusion Matrix					
Prediction	average	best	good	low	none
average	16	1	7	5	0
best	2	9	9	0	0
good	17	32	54	2	3
low	9	0	3	10	3
none	0	3	2	8	22

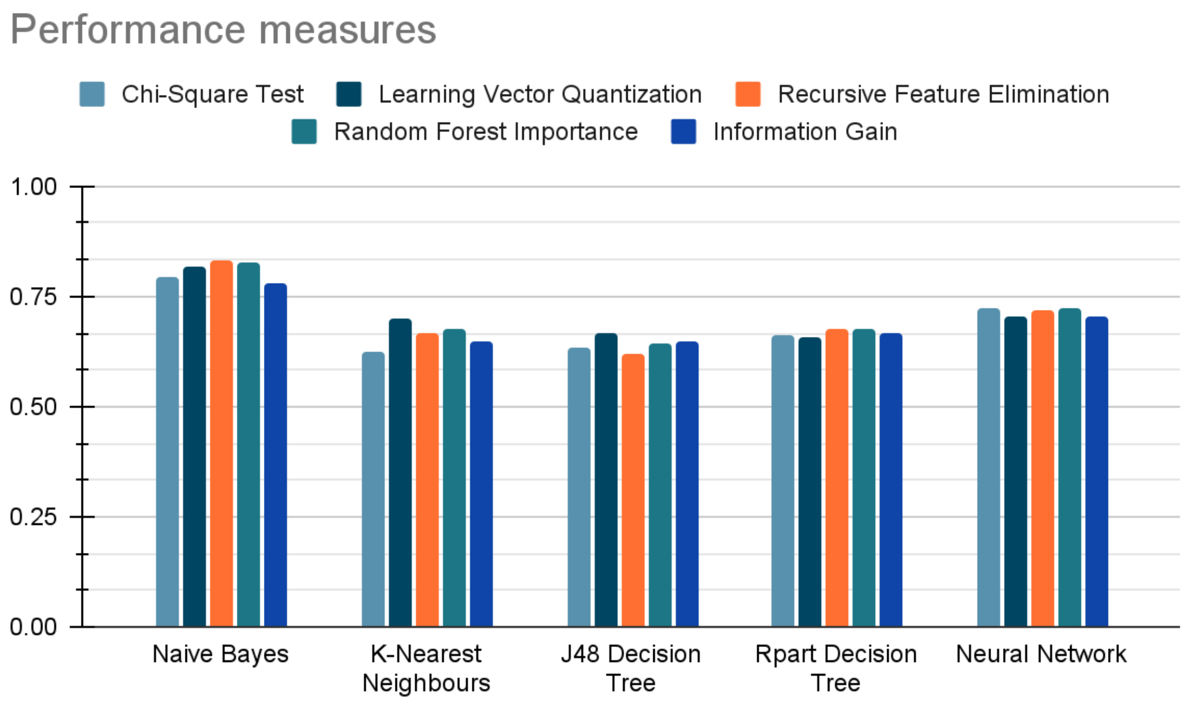
- Neural Network Model

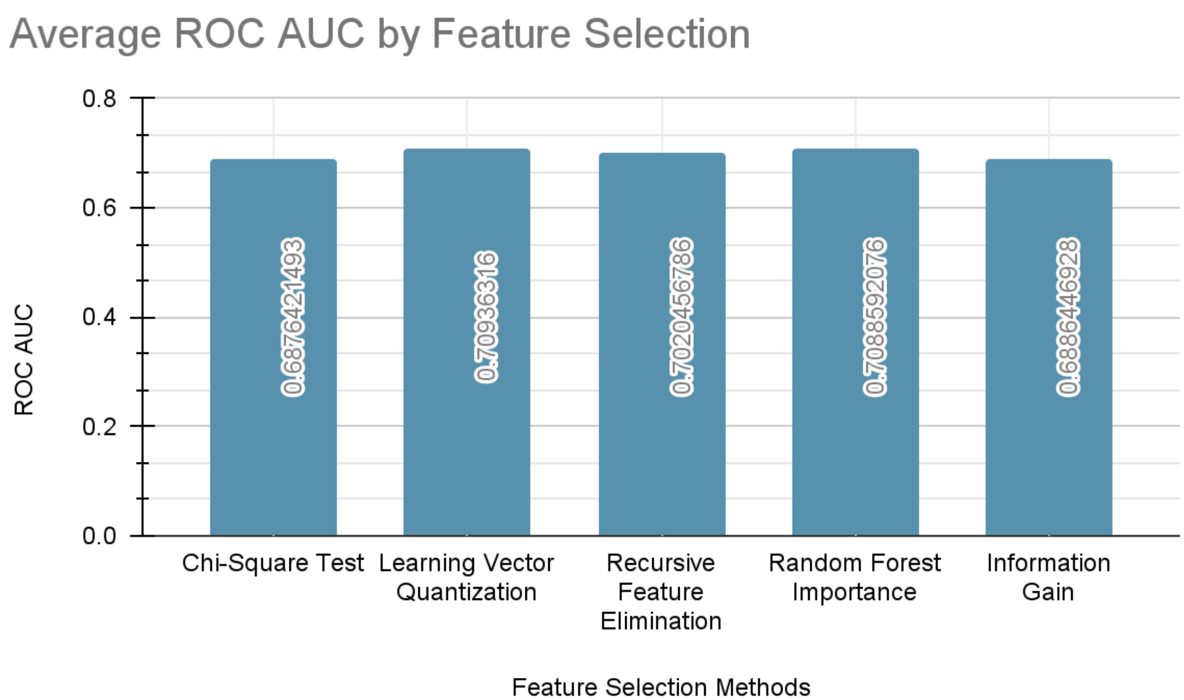
	TP	FP	precision	recall	F-measure	ROC	MCC
average	0.4545455	0.12138728	0.4878049	0.4545455	0.4705882	0.5939906103286	0.3421776167463
best	0.2222222	0.04651163	0.5555556	0.2222222	0.3174603	0.7758397932816	0.2582888372922
good	0.6400000	0.30985915	0.5217391	0.6400000	0.5748503	0.5601681555438	0.3177057992747
low	0.6000000	0.11979167	0.3947368	0.6000000	0.4761905	0.7705971277399	0.4033963805747
none	0.6428571	0.05291005	0.6428571	0.6428571	0.6428571	0.8127083333333	0.5899470899470
Weighted averages	0.51192496	0.130091956	0.5205387	0.51192496	0.49638928	0.702660804	0.3823031448

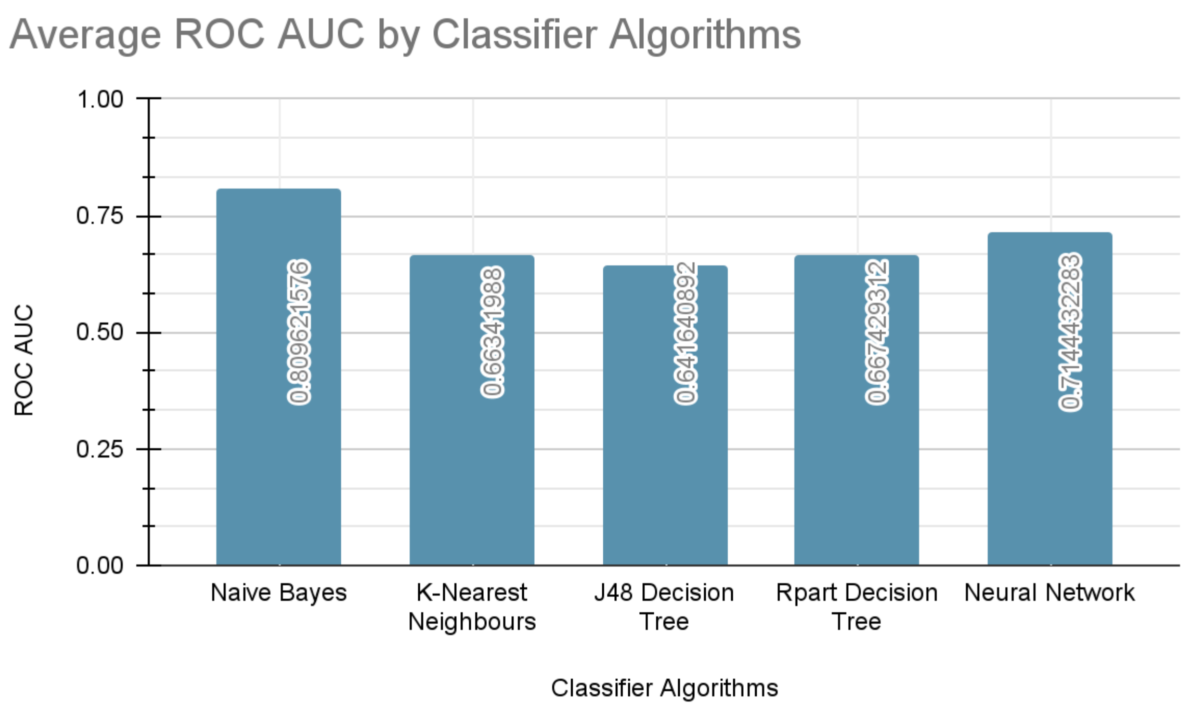
Confusion Matrix					
Prediction	average	best	good	low	none
average	20	3	14	3	1
best	0	10	8	0	0
good	12	29	48	1	2
low	10	1	5	15	7
none	2	2	0	6	18

Best Model:

The ROC Performance Results of the 25 Classification Models					
	Chi-Square Test	Learning Vector Quantization	Recursive Feature Elimination	Random Forest Importance	Information Gain
Naive Bayes	0.7946184	0.81617516	**0.83065346**	0.8269941	0.77966676
K-Nearest Neighbors	0.62546742	0.701818	0.66568458	0.67654888	0.64758052
J48 Decision Tree	0.6329008	0.66752718	0.61808884	0.64284878	0.64683886
Rpart Decision Tree	0.66375558	0.65741898	0.67509274	0.67440274	0.66647652
Neural Network	0.7214685463	0.7038764801	0.7207087732	0.7235015381	0.702660804


Performance measures
Chi-Square Test
Learning Vector Quantization
Recursive Feature Elimination
Random Forest Importance
Information Gain
1.00
0.75
0.50
0.25
0.00
Naive Bayes
K-Nearest Neighbours
J48 Decision Tree
Rpart Decision Tree
Neural Network


Average ROC AUC by Feature Selection
ROC AUC
0.8
0.6
0.4
0.2
0.0
0.6876421493
0.70936316
0.7020456786
0.7088592076
0.6886446928
Chi-Square Test
Learning Vector Quantization
Recursive Feature Elimination
Random Forest Importance
Information Gain
Feature Selection Methods



Best Model

Overall, the best model is the Recursive Feature Elimination attribute selection method combined with the Naive Bayes classification algorithm with the highest ROC AUC of 0.83065346. Based on the bar charts, we can say that Learning Vector Quantization is the best attribute selection as it has the highest average ROC AUC value, However, both the Recursive Feature Elimination and Random Forest Importance have equally close averages of ROC AUC. The Recursive Feature Elimination and Learning Vector Quantization both apply elimination in their feature selection techniques and have their respective methods of reducing the features to a subset after iterative or recursive eliminations. The best classifier algorithm to have the highest average ROC AUC is the Naive Bayes which is also our recommended classification algorithm for the best model. The Naive Bayes classifier originates from the Naive Bayes Theorem on conditional probability where the class label with the highest probability is the selected label for prediction.

Conclusion

To summarize the report, we have explored five different attribute selection methods and five different classifier algorithms. The attribute selection methods were used to improve the overall performance of classifier algorithms when working with high dimensional datasets. The initial dataset had 20 attributes (including class attribute), but with these different feature selection methods, we were able to reduce them to less than half of the total feature count. The variety in classifier algorithms was to ensure that predictions could be made from different forms of measurements. While a variety of performance measures were noted in this project, the ROC AUC was the best source of measurement for selecting the best combination of attribute selection and classification. This project is a perfect example of how data mining can be used to identify and analyze patterns. By observing only the attributes, there are different potential patterns that can be extracted (a: Survival and Insect Resistance, b: Vigor, Rainfall, Frosts and DBH and many more) to predict the level of utility of which the Eucalyptus species can be used for soil preservation and fertility. In addition to that, this project allowed our team to understand the entire process of data mining. Starting from deciding our dataset all the way to developing classification algorithms to analyzing the results and performance and coming to a conclusion.

Dataset Source:
https://www.openml.org/search?type=data&status=active&id=188

R Code:
https://colab.research.google.com/drive/1EoH5PfIj349ZmKAM1qI-LC7nqytq2JJs?usp=sharing