A Comparison of Naïve Bayes and Random Forests Algorithms for Speech Accent Recognition using Mel-Frequency Cepstral Coefficients (MFCCs)

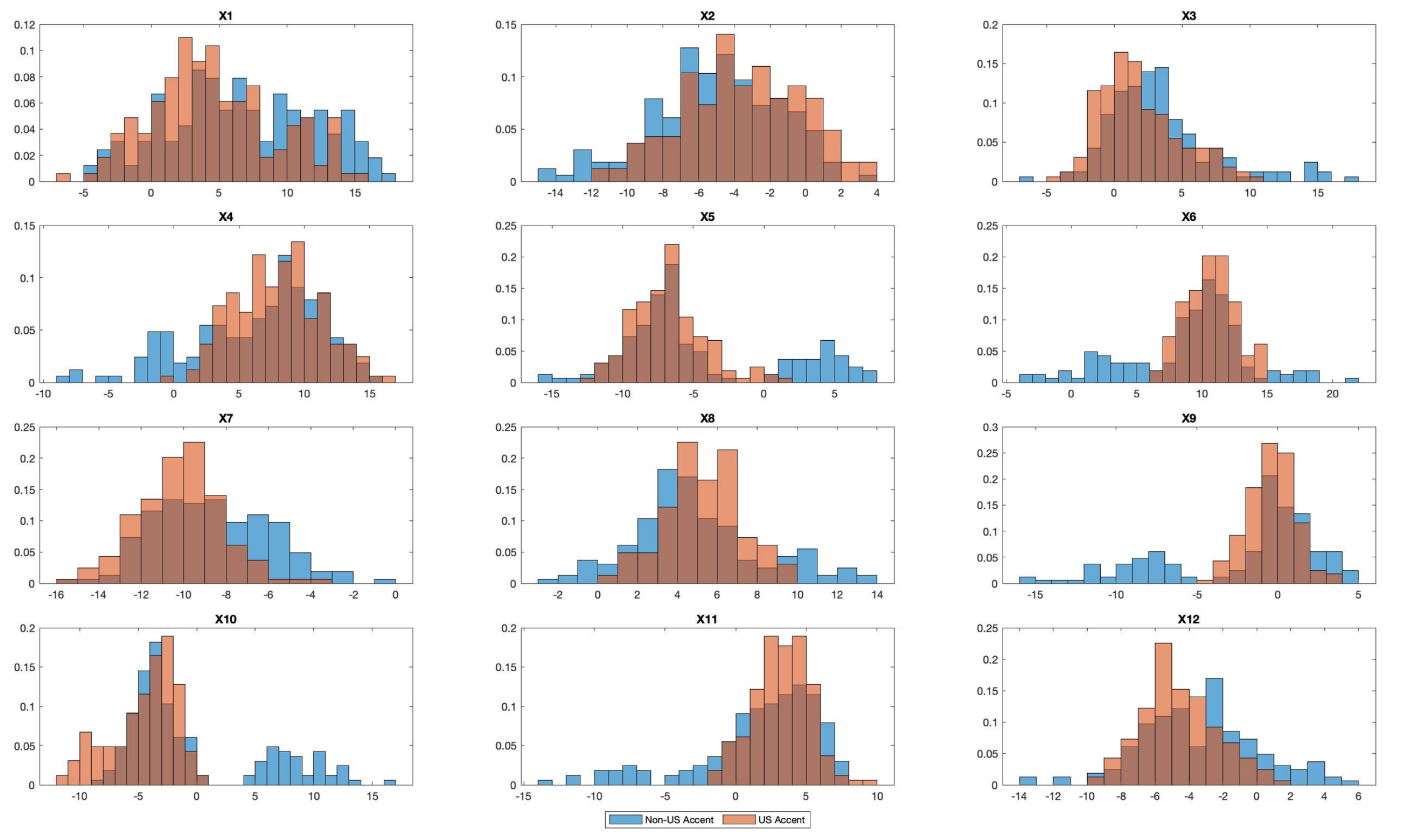
Arman Sarjou

Motivations and Brief Description of the Problem:

- To compare the performance of Naïve Bayes and Random Forests algorithms for identifying the native country of an English language speaker based on their accent
- This type of technology can be used to increase performance of large-scale speech recognition systems which are designed to work with any variant of language (Deshpande, et al., 2005; Biadsy, 2011). Computation time and accuracy of accent recognition plays a key role in performance in this context.
- This study seeks to compare performance of these algorithms to those used in a previous study using this dataset by Ma and Fokoue (2014).

Initial Analysis of Data including Descriptive Statistics:

- Speaker Accent dataset with 329 samples from the UCI Machine Learning Repository
- Features X1-X12 are generated through MFCC feature extraction of 6 audio clips of 22 speakers (11 male and 11 female) saying various words. 329 audio signals are transformed from time to frequency domain and mapped onto the Mel scale which is built to mimic human hearing perception (Ittichaichareon, et al., 2012). There is negligible noise in the signals due to how the experiment was carried out (Ma & Fokoue, 2014)
- All 12 Features are real (continuous) variables and have been shown in histograms alongside descriptive statistics.
- This dataset originally contained 6 categories (UK, US, ES, IT, GE, FR) which have been grouped to create a binary classification task of US vs Non-US accents in line with work done by Ma and Fokoue (2014) on the same dataset.
- The dataset is well balanced with 49.85% US Accent datapoints and 50.15% Non-US Accent datapoints. 0 = Non-US Accent, 1 = US Accent
- Skewness in X3 and X11 for US Data and no notably high skewness in Non-US data. All features have positive Kurtosis values which indicate that the distributions of the variables have strong peaks and are not perfectly Gaussian
- Distributions for X10, X5 and X9 within Non-US accent data look bimodal which indicates that there is some variation in the distributions of the categories that were merged to form the Non-US Accent category.
- There is little correlation between the features. This is important as Naïve Bayes works on the assumption that all input features are independent (Bishop, 2006)
- Data are all on a similar scale (values between -15 and 20) so normalization hasn't been done prior to training.



Brief Summary of the two Machine Learning (ML) Models with their Pros and Cons:

Random Forests (RF)

- Builds multiple decision trees through bagging to sample the dataset. Select random predictors at each split for each tree.
- Comes to a decision by majority vote across the multiple decision trees for classification (probabilistic approach)
- During testing, each predictor ends up on a specific leaf node of each tree and a decision is made by choosing the class with the highest posterior probability shown by histograms formed at the leaf nodes

Pros:

- Generally produces more accurate predictions than other ML models on a given dataset (Ani et al., 2015)
- Using dual-core systems RF can be parallelized (Mathworks, 2020)

Cons:

- Computationally Expensive, takes longer to train and predict
- Less interpretable as tree ensembles cannot be seen by the user (Liu et al., 2017)
- Can be prone to overfitting (Percy, et al., 2016)

Naïve Bayes (NB)

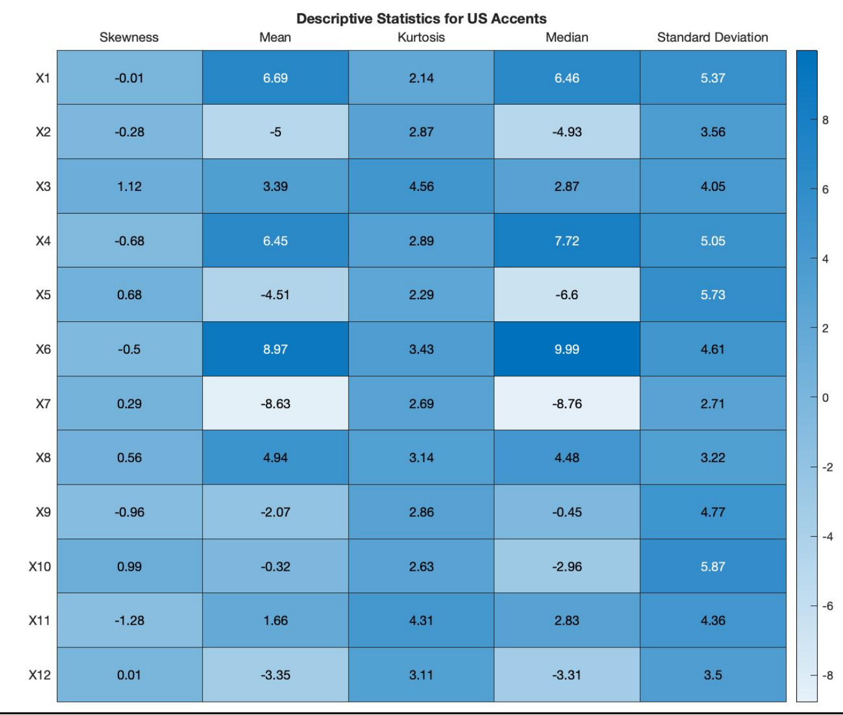
- Uses Bayes Theorem to calculate the probability of a predictor belonging to each class using probability distributions of predictors within classes
- Uses maximum a posteriori rule which selects the most probable class label based on estimated posterior probability for each class
- Named 'Naïve' as it assumes that all predictors are conditionally independent (often not true)

Pros:

- Quicker and more easily scaled than other ML methods such as SVM and Random Forests
- Easily explainable
- Relatively insensitive to noise (Webb, 2016)
- Assumptions about independence of variables do not strongly affect the accuracy of predictions (Dong, et al., 2014)

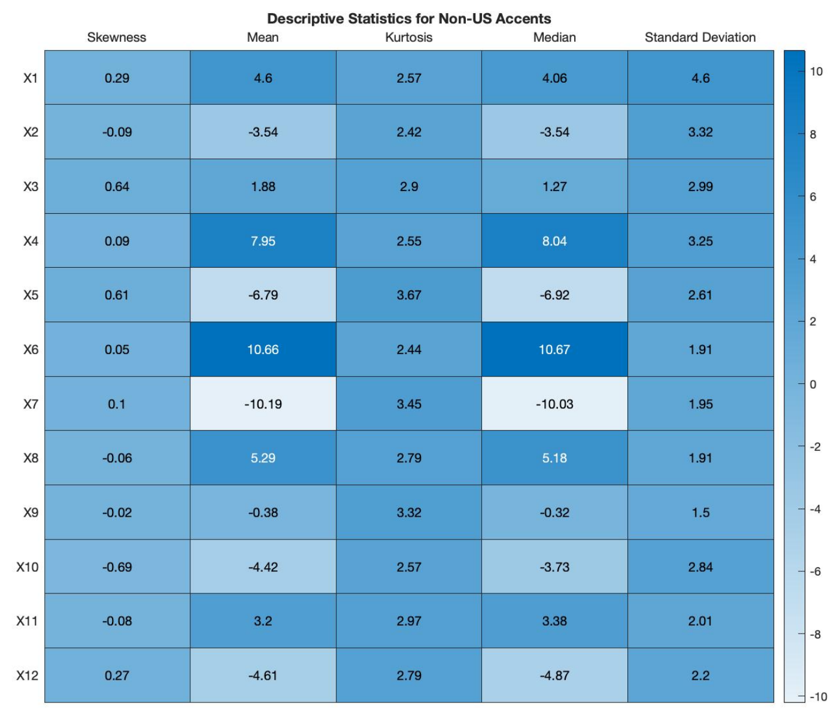
Cons:

- Performs relatively poorly with regards to accuracy when compared to other methods



Descriptive Statistics for US Accents

	Skewness	Mean	Kurtosis	Median	Standard Deviation
X1	-0.01	6.69	2.14	6.46	5.37
X2	-0.28	-5	2.87	-4.93	3.56
X3	1.12	3.39	4.56	2.87	4.05
X4	-0.68	6.45	2.89	7.72	5.05
X5	0.68	-4.51	2.29	-6.6	5.73
X6	-0.5	8.97	3.43	9.99	4.61
X7	0.29	-8.63	2.69	-8.76	2.71
X8	0.56	4.94	3.14	4.48	3.22
X9	-0.96	-2.07	2.86	-0.45	4.77
X10	0.99	-0.32	2.63	-2.96	5.87
X11	-1.28	1.66	4.31	2.83	4.36
X12	0.01	-3.35	3.11	-3.31	3.5



Descriptive Statistics for Non-US Accents

	Skewness	Mean	Kurtosis	Median	Standard Deviation
X1	0.29	4.6	2.57	4.06	4.6
X2	-0.09	-3.54	2.42	-3.54	3.32
X3	0.64	1.88	2.9	1.27	2.99
X4	0.09	7.95	2.55	8.04	3.25
X5	0.61	-6.79	3.67	-6.92	2.61
X6	0.05	10.66	2.44	10.67	1.91
X7	0.1	-10.19	3.45	-10.03	1.95
X8	-0.06	5.29	2.79	5.18	1.91
X9	-0.02	-0.38	3.32	-0.32	1.5
X10	-0.69	-4.42	2.57	-3.73	2.84
X11	-0.08	3.2	2.97	3.38	2.01
X12	0.27	-4.61	2.79	-4.87	2.2

10
8
6
4
2
0
-2
-4
-6
-8
-10

Hypothesis Statements

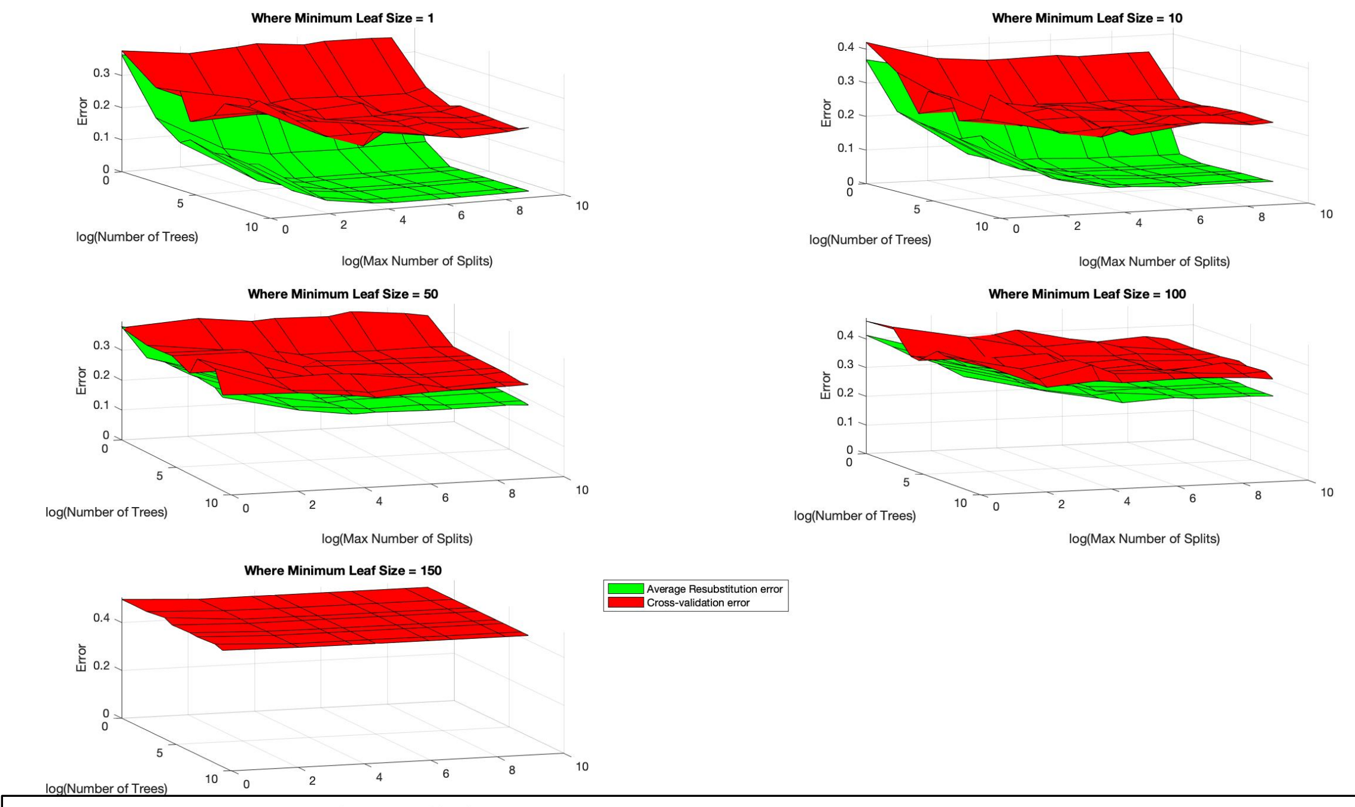
- RF classifier will perform better than NB classifier in terms of accuracy of accent detection (Ahmad, et al., 2015)
- NB classifier will perform better than RF in terms of time taken to train and test the model (Liu, et al., 2017)
- Using the same dataset with 12 MFCCs, Ma and Fokoue (2014) achieve an average accuracy of 80.97% for an SVM (Polynomial Kernel) classifier and 85.48% for a k-NN classifier.
- Ahmad, et al (2015) implies that, using an MFCC audio dataset of a similar size, RF and NB classifiers will perform worse than SVM (Polynomial Kernel) and k-NN classifier. It is expected that RF and NB classifiers will perform worse than SVM (Polynomial Kernel) and k-NN classifiers from Ma and Fokoue (2014)

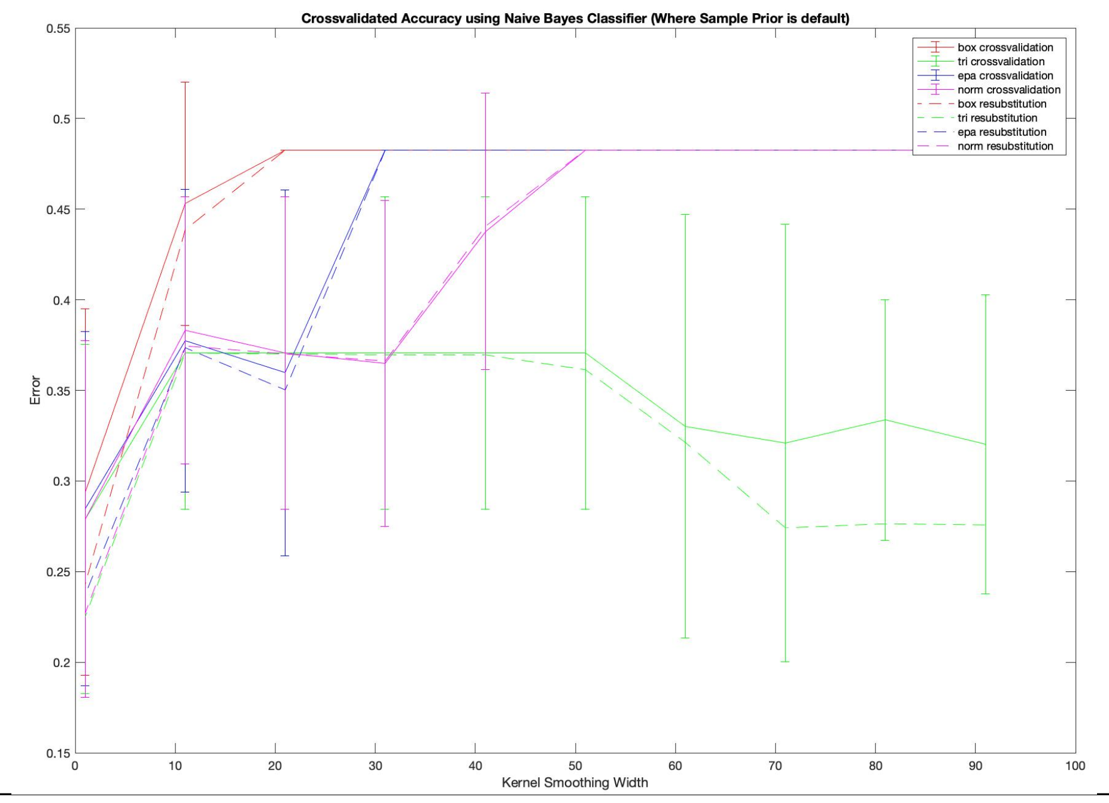
Description and explanation of training and evaluation methodology:

- Held 30% (99 samples) of the dataset for model testing and 70% (230 samples) of the dataset for training
- When using the same dataset, Ma and Fokoue (2014) used hold-out cross validation based on random stratified sampling iterated over 500 times. The accuracy in their study is then calculated as the average accuracy of these 500 iterations. Due to the relatively small size of the dataset, this type of cross-validation is unsuitable and a K-Fold cross-validation approach is more appropriate (Kohavi, 1995). Both were tested to observe variations in performance
- fitcensemble (used for generation of RF) performs bootstrapping internally whereas fitcnb (used for generation of NB classifier) does not. It was decided that it would be excessive to perform bootstrapping on the Naïve Bayes classifier as well as K-Fold Cross Validation
- Carried out manual Grid Search to optimize hyperparameters for each model. Selected best models based on the cross-validated loss to ensure that there is little overfitting (which would lead to poor accuracy in testing). Record standard deviation of cross-validated and average resubstitution error to evaluate stability of models.
- During testing, calculated F1 score, Accuracy, Precision, Recall and AUC to compare and evaluate the models. Like Ma and Fokoue (2014), the average time to carry out cross-validation is used to compare computational cost.

Choice of Parameters and Experimental Results:

- 10-Fold Cross-Validation is used instead of the iterated randomly stratified hold-out sampling which was suggested by Ma and Fokoue (2014). Optimization using the hold-out sampling method took over 38 hours for the Naïve Bayes (360 different models) and could not be completed for the Random Forests Method (125 different models) due to time constraints. Experiments were run on an Intel i7, 16GB RAM Macbook Pro.
- Naïve Bayes hyperparameters: Distribution Type (Kernel or Normal), Kernel Smoothing Width, Kernel Smoothing type
- Random Forests hyperparameters: Number of Trees, Maximum Number of Splits (Tree Depth), Minimum Leaf Size (Number of samples at each node)
- Optimisation for Naïve Bayes took ~30 minutes whereas Random Forest took ~10 Hours
- Best Naïve Bayes model has the following parameters: Kernel Distribution Type, Kernel Width: 1, Normal Kernel
- Best Random Forest model has the following parameters: Minimum Leaf Size: 1, Number of Trees: 50, Maximum Number of Splits: 5000

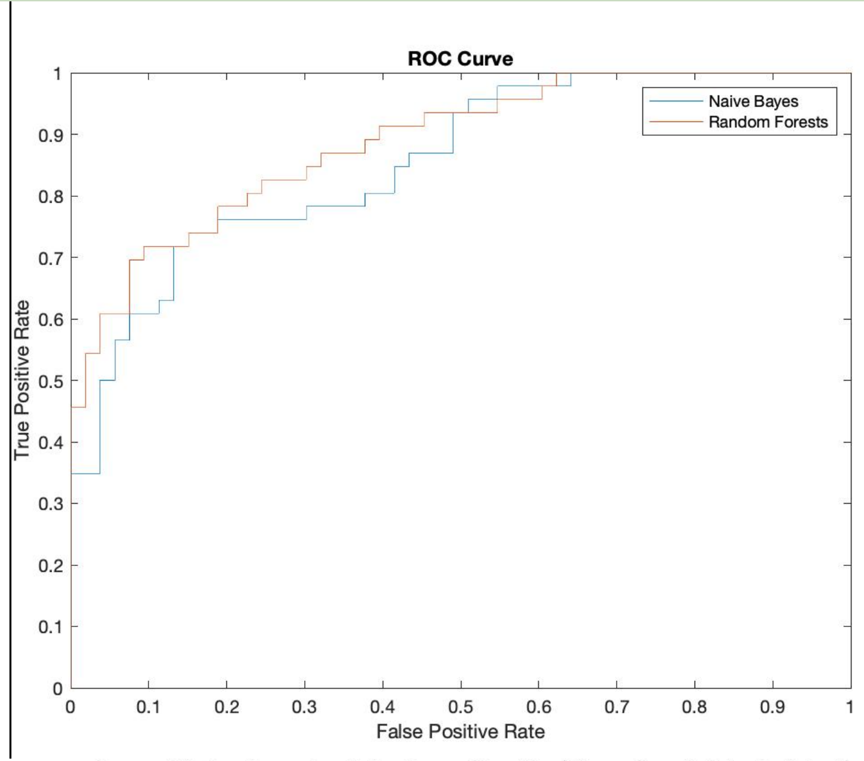


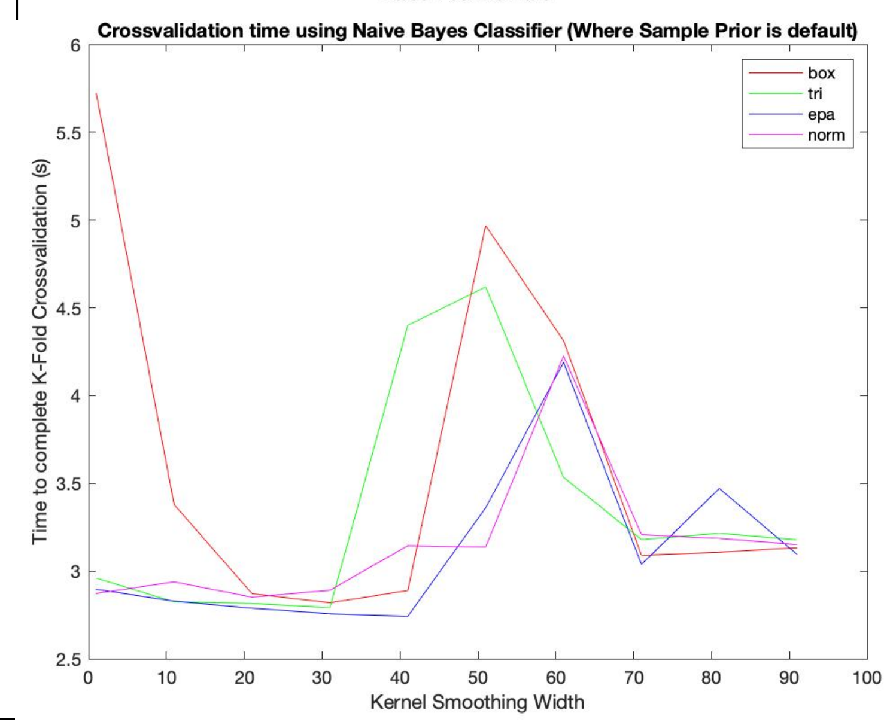


Method	Resubstitution Accuracy (1– Error)	Testing Accuracy (1 – Error)	Testing AUC	Testing F-Score	Testing Precision	Testing Recall	Time to perform 10-Fold Crossvalidation (s)
Random Forests	99.1%	78.5%	0.887	76.9%	77.8%	76.1%	3.54
Naïve Bayes	77.0%	72.3%	0.855	64.8%	92.0%	50.0%	2.87

Analysis and Critical Evaluation of Results :

- RF model has a moderately higher accuracy of prediction on the testing data (78.5%) than the NB model (72.3%). Both final models confirm the hypothesis that NB and RF classifiers will perform worse than SVM and k-NN for this MFCC dataset. However, the RF model has a much higher resubstitution accuracy (99.1%) than the NB model (77.0%). The difference between these values indicates that the RF model is a higher variance model but manages the variance-bias trade off better than the NB (seen by higher testing accuracy). The NB model has lower variance with relatively higher bias which means that it will perform worse on this dataset but may also perform similarly on other MFCC datasets.
- During Grid Search, the best performing RF model had a Standard Deviation in training error of 0.0069 and 0.1002 in cross-validation error. The best performing NB model had a Standard Deviation in training error of 0.0070 and 0.00984 in cross-validation error. This indicates the final models are relatively stable.
- For RF, as Tree Depth increases the stability of the model increases as seen by general decreases in standard deviation of error. Additionally, as Tree Depth increases, the average resubstitution/cross-validation error difference decreases at all leaf sizes. Increased Tree Depth creates more complex models that have higher variance. Too much depth can cause overfitting (Liu, et al., 2017). Highest stability is found with shallow trees, but this causes larger cross-validation error (Liu, et al., 2017)
- As number of trees increase, there is a general decrease in error as well as less dramatic increases in the difference in the average resubstitution/cross-validation error difference. This indicates that increasing number of trees helps the RF algorithm to manage the variance-bias trade off better through a larger committee of uncorrelated trees during voting. Stability increases with increase in number of trees due to central limit theorem but comes with increase in computational cost (Liu, et al., 2017).
- When observing time to perform cross-validation, there is a clear relationship between the Tree Depth and the average time for cross-validation with leaf size and number of Trees having a minimal effect in comparison. This varies from Naive Bayes where there is a variation in time taken to cross-validate between different filter types.
- NB consistently produces results a relatively smaller resubstitution/cross-validation error difference where kernel width is less than 50. NB models with triangular smoothing kernel show a much larger difference between resubsitution error and cross-validation error at Kernel Width greater than 60. This is could be due to the high kurtosis of the data distributions which makes the triangle a better representation of the distributions
- NB model has a lower AUC value (0.855) in testing and lower ROC values than RF in testing. NB outperforms RF over some thresholds. RF model provides strong in-sample results.
- NB is much better than RF at not returning false positive results (stating someone has a US accent when they do not) as seen by precision scores (92%) in testing. RF performs better in not returning false negative results (stating someone doesn't have a US accent when they do) as seen by recall results in testing (76.1%). In the US, a NB classifier may be more appropriate to ensure those who aren't native can receive relevant adjustments. The balance of precision and recall seen in F-Score where RF performs 12.1% better.
- When Sample Priors are adjusted in NB, it is possible to achieve improvements in resubstitution and cross-validation error with decreases in Recall and F1 scores. When priors are at extreme values (eg [0.9, 0.1]), models predicts only one class which produces 0 values in F1, Recall and Precision scores
- Both final models take a similar time to perform 10-Fold cross-validation and only perform worse than the k-NN classifier in Ma and Fokoue (2014) although this is would be an unfair comparison to make due to different computational equipment being used.





Future work and lessons learnt:

- Both models produce relatively stable final models, RF takes much longer to optimize using manual grid search which can prove a problem where computational cost is important.
- NB performs well for its simplicity and could be a solution in a scenario where scalability, ease of optimization and explainability is key
- Observe how the two models perform in a multiclassification problem (not binary)
- Observe the effect of Feature Selection in NB for model performance and number of features randomly selected at each split for RF
- Observe the effect of varying training data sizes (Ahmad et al,. 2015)
- Add noise to dataset to see how each model handles introduction of noise (Ma & Fokoue, 2014)

Ahmad, J. et al., 2015. Gender Identification using MFCC for Telephone Applications - A Comparative Study. International Journal of Computer Science and Electronics Engineering, 3(5).

Ani, R., Augustine, A., Akhil, N. & Gopkumar, D. O., 2016. Random Forest Ensemble Classifier to Predict Coronary Heart Disease using Risk Factors. Advances in Intelligent Systems and Computing, Volume 397, pp. 701-710.

Biadsy, F., 2011. Automatic Dialect and Accent Recognition and Its Application to Speech Recognition. New York: Columbia University.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Cambridge : Springer.

Dong, L., Li, X. & Xie, G., 2014. Nonlinear Methodologies for Identifying Seismic Event and Nuclear Explosion Using Random Forest, Support Vector Machine and Naive Bayes Classification. Abstract and Applied Analysis, p. 8.

Deshpande, S., Chikkerur, S. & Govindaraju, V., 2005. Accent Classification in Speech. Fourth IEEE Workshop on Automatic Identification Advanced Technologies, 11(09), pp. 139-143.

Ittichaichareon, C., Suksri, S. & Yingthawornsuk, T., 2012. Speech Recognition using MFCC. Pattaya, International Conference on Computer Graphics, Simulation and Modelling.

Kohavi, R., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Stanford, International Joint Conference on Artificial Intelligence.

Liu, B. C. H., Chamberlain, B. P., Little, D. A. & Cardoso, A., 2017. Generalising Random Forest Parameter Optimisation to Include Stability and Cost. Machinee Learning and Knowledge Discovery in Databases, Volume 10536, pp. 4-9.

Mathworks, 2020. Statistics and Machine Learning Toolbox. [Online]
Available at: https://www.mathworks.com/help/stats [Accessed 03 December 2020].

Ma, Z. & Fokoue, E., 2014. A Comparison of Classifiers in Performing Speaker Accent Recognition using MFCCs. Open Journal of Statistics, Volume 4, pp. 258-266.

Ma, Z. & Fokoue, E., 2014. Accent Recognition for Noisy Audio Signals. Serdica Journal of Computing, Volume 2, pp. 169-182.

Percy, C., Franca, M., Dragicevic, S. & Garcez, A. d., 2016. Predicting online gambling self-exclusion: an analysis of the performance of supervised machine learning models. International Gambling Studies, 16(2), pp. 193-210.

Webb, G. I., Keogh, E. & Miikkulainen, R., 2016. Naive Bayes. In: Encyclopedia of Machine Learning. Boston: Springer.