Introduction

For the purposes of the Ensemble Classifier Case Study I chose a dataset from Kaggle titled "Gender Recognition by Voice and Speech Analysis." The purpose of the data is to identify a voice as either male or female based upon the acoustic properties of the voice and speech patterns. This dataset makes a great use case for considering ensemble classifier approaches first and most obviously because the target attribute is categorical as either male or female. What's more, the fact that the target attribute is binary is also helpful when using ensemble approaches such as AdaBoost and Logistic Regression. The fact that all the descriptive variables are numerical is also helpful in creating unambiguous distance calculations for categorization approaches such as K Nearest Neighbors.

To begin the paper I'll consider five basic categorization algorithms to be used as a baseline for accuracy: K Nearest Neighbors, Naïve Bayes, Linear Discriminant Analysis, Logistic Regression, and Decision Trees. I'll next consider two ensemble approaches to confirm whether I can improve upon the baseline performance of the underlying algorithms: bagging and stacking. With bagging, I'll create an ensemble learner by repeatedly building models using the same algorithm where random samples (with replacement) from the original training set are used to build the model. With stacking, I'll see if there is a way to combine the underlying approaches and allow them to vote in order to determine the overall classification. Because the target attribute is balanced (equal number of male and female voices in the sample), statistical methodology will be used to test hypothesis regarding the accuracy of each model or ensemble learning approach.

Data Description

The dataset consists of 20 descriptive attributes for voices such as frequency, skewness, and flatness. The target attribute is binary as either male or female. The dataset is pulled from Kaggle and can be found by following https://www.kaggle.com/primaryobjects/voicegender. In total, there are 3168 categorized voice samples which are equally distributed between male and female voices. The data consists of three major categories of attributes along with a handful of assorted others. Specifically, we have a number of descriptive attributes related to the frequency of the voice, descriptive attributes around what the researchers call a "fundamental frequency," and descriptive attributes around what is called a "dominant frequency." A full description of the dataset's attributes can be found in Appendix I.

Data Cleaning

Most probably because of the relatively cheap acquisition cost of the data (simply recording voice samples from volunteers), the data is very clean with no missing data points. However, there was some manipulation of the dataset performed prior to running any algorithms to ensure analysis could be completed as expected. First, copies of the dataset with scaled attributes (mean of zero and standard deviation of 1) and also with a binary target attribute (male == 1 while female == 0) for the purposes of the K Nearest Neighbors and Logistic Regression models respectively. I also created three subsets of each of the three copied of the data: one each for the frequency, fundamental frequency, and dominant frequency attributes. The purpose for these subsets of the data is to perform correlation analysis between the attributes and also for use in building ensemble learners using the stacking approach as will be described in more detail below.

Data Analysis

Not surprisingly, the descriptive attributes in each of the frequency, fundamental frequency, and dominant frequency subgroups display high intragroup correlation. If the average dominant frequency for an individual is high, it stands to reason that the 75th percentile of the dominant frequency for this same individual is also high. Comparing the mean frequencies for each of these three groups – frequency, fundamental frequency, and dominant frequency, reveals a high intergroup correlation as well. Figure 1 displays visualizations for these inter and intragroup correlations. Empirically, it holds that there would be correlation in attributes describing a given individual's voice tones and patterns.

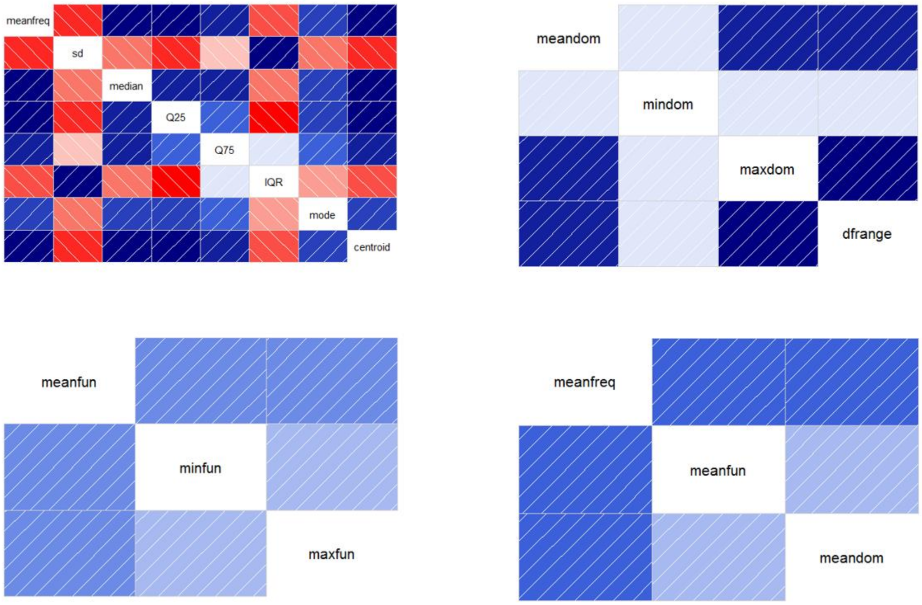

Figure 1 - Correlation between Frequency Variables

Experimental Results

Independent Categorization Approaches

I'll begin the analysis by creating models using specific categorization approaches and using their results as a baseline for evaluating the ensemble methods. 10-fold cross validation was used to evaluate the accuracy of each of the five methods. The K Nearest Neighbor model included a bit more tuning by allowing the *train* function in R to select the optimal number of nearest neighbors (k = 5 is given as the optimal number). This model also used the scaled dataset where the mean of each attribute was zero and the standard deviation was one. Results for the accuracy of each of these three models along with a 95% confidence interval is shown in Figure 2. As an additional interesting benchmark, Random Forest and AdaBoost algorithms were both able to correctly predict 100% of the instances. These two algorithms create ensemble learners by leveraging Bagging and Boosting respectively but are outside the explicit scope of this paper.

Method	**Accuracy**	**95% CI**
Linear Discriminant Analysis	0.9686	(0.9519, 0.9807)
Logistic Regression	0.9765	(0.9615, 0.9868)
Decision Tree	0.9618	(0.9545, 0.9682)
Naïve Bayes	0.9299	(0.9205, 0.9386)
K Nearest Neighbors	0.9842	(0.9792, 0.9883)
Random Forest	1	(0.9988, 1)
AdaBoost	1	(0.9988, 1)

Figure 2 - Accuracy of Individual Categorization Models

Ensemble Learning with Bagging

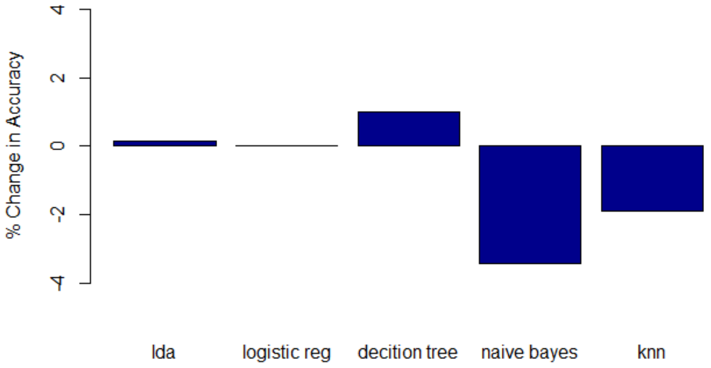

Figure 3 - Impact of Bagging

The first approach used for evaluating ensemble learning techniques is bagging, where an ensemble learner is created by building many models using the same algorithm where random subsets of the original training data are used as the training data. Then, the outcomes of these models are allowed to collectively determine the overall categorization by majority vote. Interestingly, the impact bagging had on our five underlying models was not consistent. In some cases, such as with the Decision Tree model, bagging improved the performance of the underlying model. In other cases, such as Naïve Bayes and the KNN model, bagging actually reduced the accuracy of the model. It's also worth noting that the Naïve Bayes and KNN approaches were our worst and best base algorithms respectively. The change in accuracy as a result of bagging for each underlying model are shown in Figure 3 and each are statistically significant at a level of .01 (with the exception of the logistic regression model as there was no change in accuracy as a result of bagging). Even with the improved accuracy of the decision tree algorithm as a result of bagging, we still reject the null hypothesis that the accuracy of this model as greater than the accuracy of the base KNN model at a level of .05.

Ensemble Learning with Stacking

We can also create ensemble learners by combining multiple "weak" learners such as those used to establish our baseline for accuracy and allowing them to collectively make a categorization. Ultimately, the goal will be to improve upon the accuracy found with our best performing individual algorithm which was K Nearest Neighbors with an accuracy of 0.9842. The simplest way of combining the algorithms would be allowing the overall categorization to made by a majority vote. When this is done the accuracy of the combined model is 0.9751 which is not an improvement over the KNN model alone. However, we know that the underlying models have varying levels of accuracy, so why allow them each the same vote? A smarter approach would be to weight each algorithm and make the overall categorization based on a the weighted score from the results of each algorithm per instance. We could solve for these coefficients (weights) directly, but in this case the system is computationally singular so this is not possible. Instead, we must estimate the weights using a method such as linear regression. Using R's lm() function, Figure 4 displays the weights attributed to each of our underlying models. The coefficients returned by the linear regression model are consistent with what we would expect based on the accuracy of each model. For example, the highest weight is assigned to our best performing model (KNN) while the lowest weight is assigned to our worst performing model (Naïve Bayes). A coefficient of NA for the logistic regression model indicates that this variable is linearly related to the other variables and so has been removed from the formula.

Model	Intercept	Decision Tree	Naïve Bayes	KNN	LDA	Logistic Regression
Coefficient	0.01386527	0.12060637	0.00608915	0.68434159	0.17830023	NA

Figure 4 – Coefficients assigned to each attribute

When we allow the weights to be assigned in this way and make assign categorizations based on these weights, we get a total accuracy that matches the performance of the KNN approach of 0.9842. What this means is that the underlying models weren't wrong in enough "different ways" for their miscateogrizations to cancel each other out when taken together. In this case, the weight assigned to the KNN model is so great that the rest of the models must correctly categorize any instance miscategorized by the KNN model in order for the instance to be correctly categorized by the ensemble.

As we've seen, variability in the underlying models is a necessary condition for ensemble learning approaches to improve upon base classifiers. With this fact in mind, another way to think about stacking is to allow the different algorithms to consider different *aspects* of the data prior to considering their results. Recall how this dataset consists of three similar types of frequency: frequency, dominant frequency, and fundamental frequency. If we let our underlying models consider these data elements independently, this may allow our ensemble learner to understand different aspects of the data and then combine these understandings for an improved overall prediction. For the purposes of this exercise, I built KNN and Decision Tree models based on each of these three aspects of the original dataset. I next made predictions for the overall dataset based on each of the underlying models and allowed a categorization to made based on simple majority vote. Figure 5 displays the results of this analysis.

Model	**Data Subset**	**Accuracy**	**95% CI**
Decision Tree	Frequency	0.9113	(0.9009, 0.921)
	Fundamental Frequency	0.9605	(0.9532, 0.9671)
	Dominant Frequency	0.6859	(0.6694, 0.7021)
Combined Decision Tree models		0.941	(0.9322, 0.9489)
KNN	Frequency	0.9612	(0.9539, 0.9676)
	Fundamental Frequency	0.9669	(0.96, 0.9728)
	Dominant Frequency	0.7778	(0.7629, 0.7922)
Combined KNN models		0.976	(0.9701, 0.9811)
Combined All models		0.976	(0.9701, 0.9811)

Figure 5 – Results of Stacking

Experimental Analysis

What we've seen is that each of our individual models perform relatively well in predicting the gender of the volunteer based on the acoustic properties of their voice with our worst model – the Naïve Bayes algorithm – correctly predicting 93% of the instances. While this is a good indicator off hand, it may have also have led to some of the difficulty we saw with improving upon the base models – particularly with stacking – as there simply was not much improvement to be made. We also saw some concerning issues in the underlying data that lead to difficulty with our regression models. As noted in the Data Analysis section, many of the attributes in our dataset are highly correlated with each other. As we know, the basic "naïve" assumption of the Naïve Bayes model is that the descriptive characteristics are independent. Because this is clearly not the case in our dataset, it holds that the

Naïve Bayes approach will perform poorly. In our initial stacking approach that attempted to assign weights to the predictions of all of the underlying models, we found that the Logistic Regression model was left out of the equation due to the fact that it was linearly related to other models in the formula (notably Linear Discriminant Analysis) and was therefore assigned no weight. Without such collinearity in the data, we may have seen the weights spread more evenly across the five algorithms which would have increased the likelihood that the approach would have correctly categorized instances that had been miscategorized by the K Nearest Neighbors approach.

We also discussed the impact of Bagging on each underlying categorization approach to create an ensemble learning. Interestingly, we didn't always see improvement in the accuracy of each model due to bagging and sometimes even saw a decrease in the performance of the models. Why might this be? It holds that the accuracy of the decision tree model would increase with bagging, as the Random Forest algorithm (which is a variation of a Decision Trees with bagging) was able to correctly predict 100% of the instances. We also know that the Naïve Bayes model suffered from collinearity in the underlying data, so it would seem to hold that the impact of problem would increase as the number of faulty models grow.

When we think about the fact that our "smart" Decision Tree was able to correctly identify 96% of the instances, it's remarkable that both the Random Forest and the AdaBoost algorithms – which both use a variation of a "weak" Decision Tree as part of their underlying algorithm – were able to correctly classify 100% of the instances with an ensemble of weak learners.

Conclusion

Overall, we found that while we were able to improve the accuracy of individual models with Bagging and Boosting using Random Forest and AdaBoost, we were unable to directly apply the Bagging and Stacking methods to improve upon the accuracy of our best base categorization method. We noted that the Bagging methodology had a positive effect on the accuracy of some categorization approaches while it decreased the accuracy of others. We also explored different methodologies of stacking by combining the results of our underlying algorithms and/or allowing different algorithms to consider different elements of the larger dataset prior to making a final categorization. Even with all these things considered, this paper has only brushed the surface of the different variations available through ensemble learning. Considerations for future work should include continued exploration of stacking approaches, letting different algorithms consider different elements of the underlying data. Instead of making an overall categorization from our stacking approach based on simple majority voting, we could also make this categorization based on weighted inputs from the underlying categorizations similar to the weights shown in Figure 4. We could also spend more time fine tuning the underlying base categorization approaches with methods such as grid search. Overall, however, we've been able to effectively demonstrate the power of ensemble approaches while also highlighting many of the reasons that building such ensemble learners is an art as much as it is a science.

Appendix I – Data Dictionary

Label	Description
meanfreq	mean frequency (in kHz)
sd	standard deviation of frequency
median	median frequency (in kHz)
Q25	first quantile (in kHz)
Q75	third quantile (in kHz)
IQR	interquantile range (in kHz)
skew	skewness (see note in specprop description)
kurt	kurtosis (see note in specprop description)
sp.ent	spectral entropy
sfm	spectral flatness
mode	mode frequency
centroid	frequency centroid (see specprop)
peakf	peak frequency (frequency with highest energy)
meanfun	average of fundamental frequency measured across acoustic signal
minfun	minimum fundamental frequency measured across acoustic signal
maxfun	maximum fundamental frequency measured across acoustic signal
meandom	average of dominant frequency measured across acoustic signal
mindom	minimum of dominant frequency measured across acoustic signal
maxdom	maximum of dominant frequency measured across acoustic signal
dfrange	range of dominant frequency measured across acoustic signal
modindx	modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
label	male or female