Statistical Learning Methods for Facial Demographic Analysis: An Exploration with the MORPH-II Dataset

D. Johnston – Brigham Young University
2019 UNCW REU Program
July 22nd, 2019

ABSTRACT. In this report, we explore the MORPH-II Dataset numerically and graphically. We gather data on the demographics of the dataset then subset the data to train, test, and compare various statistical learning models for classification and regression, including Polynomial Regression, Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN), Decision Tree (DT), Bagging, Random Forest (RF), Boosting, and Support Vector Machine (SVM). Results for each model are compared; SVM achieved the highest accuracy for gender classification, while Boosting attained the lowest mean squared error in predicting age.

We further explore the use of Latent Class Analysis (LCA) as an unsupervised method for grouping individuals into classes based on chemical exposure. We conclude there is little to no relationship between these classes and age or conviction status of each individual.

Finally, we return to age and gender prediction models and utilize Principal Component Analysis (PCA) and Kernel PCA (KPCA) to reduce dimensionality and enable our models to train on a large proportion of the data. Improvement is seen across all models.

Keywords: MORPH-II, Facial Demographic Analysis, Gender classification, age prediction

1. Introduction

The MORPH-II dataset is one of the largest longitudinal morphological face database available to the public (Ricanek and Tesafaye, 2006). It is comprised of mug shots taken of arrested persons over a period of five years and has been used for advancement in facial recognition in many settings. For this report, we used an album with over 55,000 entries of more than 13,000 unique persons.

The dataset contains information such as gender, race, and age for each image, as well as 2,500 Bio-Inspired Features (BIF) taken from the images that we will use to predict age and gender. Prediction models used include Polynomial Regression, Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN), Decision Tree (DT), Bagging, Random Forest (RF), Boosting, and Support Vector Machine (SVM). We will compare the performance of these models using Mean Squared Error (MSE) as a metric for regression to predict age, and accuracy, specificity, sensitivity, standard error, and time elapsed as metrics for gender classification. We will compare results using both 5-Fold Cross-Validation and Leave-One-Out Cross-Validation.

Following this analysis, we will explore Latent Class Analysis (LCA) as a means to classify individuals based on given chemical measures and determine whether a relationship exists between these classes and age or whether the individual was convicted.

Received by the editors May 8, 2020.

Finally, we will return to the gender classification and age regression models. We perform similar tests as before, using Principal Component Analysis (PCA) and Kernel PCA (KPCA) to reduce dimensionality in the data and compare our results.

2. The Data

2.1. Uncleaned Data

Initially, we were given data that had not been processed or cleaned. As we gathered initial exploratory data, we found discrepancies in the numbers. There was one case where someone had multiple images in the dataset, but they weren't all recorded as the same gender. There were many other entries that attached multiple races to one person. Table 2.1 breaks down the number of unique individuals by race and gender before cleaning the data.

TABLE 2.1. Before Cleaning: Number of Unique Individuals by Gender and Race

	Black	**W**hite	**A**sian	**Hi**spanic	**O**ther	**Total**
Male	8,837	2,070	49	517	15	11,488
Female	1,494	634	6	30	5	2,169
Total	10,331	2,704	55	547	20	**13,657**

We see a total of 13,657 individuals when summing number of unique individuals in each subgroup. However, when we found the total number of unique individuals, it amounted to only 13,617. We can see the there were many people assigned to multiple subgroups in the data. Figures 2.1 and 2.2 illustrate the distribution of this data by gender and race.

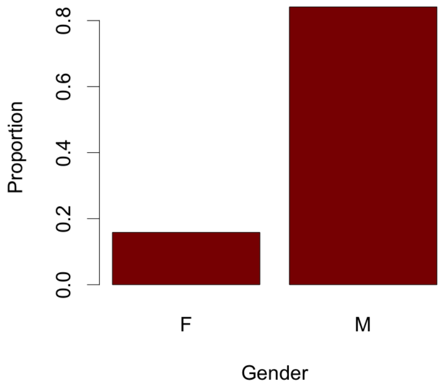

FIGURE 2.1. Distribution by Gender

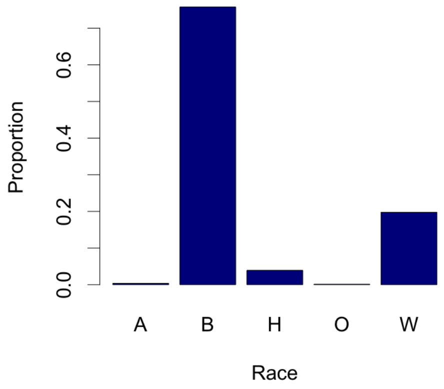

FIGURE 2.2. Distribution by Race

These graphs make more apparent the uneven distribution inherent in the data. The number of individuals is very male-dominant, and the distribution by race is characterized by a large majority of Black people, followed by a significant portion of White people, a small number of Hispanic people, and very few individuals in the Asian or Other categories.

Examining the distribution of age, we find relatively similar results throughout each subgroup. By gender, we have close to the same range of ages represented, with similar numbers across a five-number summary. As we analyze the results by race, we see younger individuals on average among the Hispanic, Other, and Asian subgroups than among the White and Black subgroups. However, this could be due to the relatively smaller sample size of Hispanics, Others, and Asians in the data. A summary of the results follows in Table 2.2.

TABLE 2.2. Numerical Summary of Age

	Min	1st Q	Median	Mean	3rd Q	Max
All	16.00	23.00	33.00	32.62	41.00	77.00
Female	16.00	25.00	34.00	33.33	41.00	75.00
Male	16.00	23.00	32.00	32.49	41.00	77.00
Black	16.00	23.00	32.00	32.40	41.00	71.00
White	16.00	25.00	35.00	34.58	42.00	77.00
Asian	16.00	20.00	23.00	23.03	26.00	51.00
Hispanic	16.00	20.00	25.00	26.67	32.00	65.00
Other	19.00	27.75	39.50	37.18	48.00	53.00

Despite some differences in the age distribution among different race subgroups, the same general trend is followed in each. More arrests occur among a younger population.

Figure 2.3 illustrates the downward trend of frequency of arrests with age. The age group with the most arrests is the 16-20 years old group. The numbers generally decline as age increases, with very few arrests of individuals over the age of 60.

2.2. Cleaned Data

As we began our second project, we were given data that had been cleaned and pre-processed already. As we expected, the discrepancies found in the uncleaned data from Project 1 had been dealt with, and the numbers for the distribution by gender and race actually made sense in the context of the total number of unique individuals. Comparing Table 2.3 to Table 2.1, we can see at least 40 corrections were made, likely more.

TABLE 2.3. After Cleaning: Number of Unique Individuals by Gender and Race

	Black	**W**hite	**A**sian	**H**ispanic	**O**ther	**Total**
Male	8,829	2,056	47	507	19	11,458
Female	1,491	628	4	28	8	2,159
Total	10,320	2,648	51	535	27	**13,617**

The charts and tables given above for the data before it was cleaned remained largely unchanged after receiving the cleaned data. Due to the relatively few corrections made in comparison to the large dataset, very little adjustments were made visually. However, some small changes among the subgroups with a small sample size became apparent upon examination of the box plots in Figures 2.4 and 2.5.

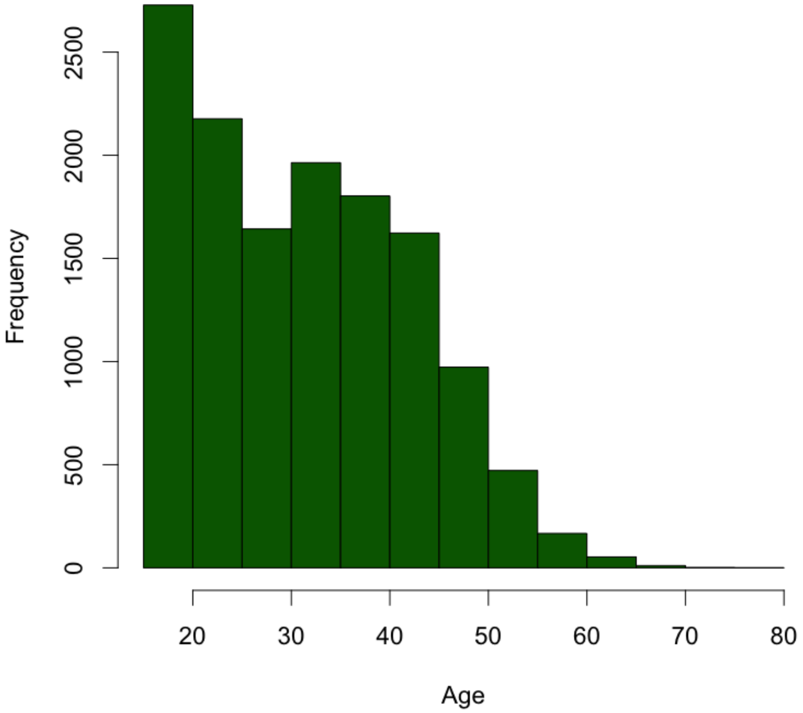

FIGURE 2.3. Distribution of Arrests by Age

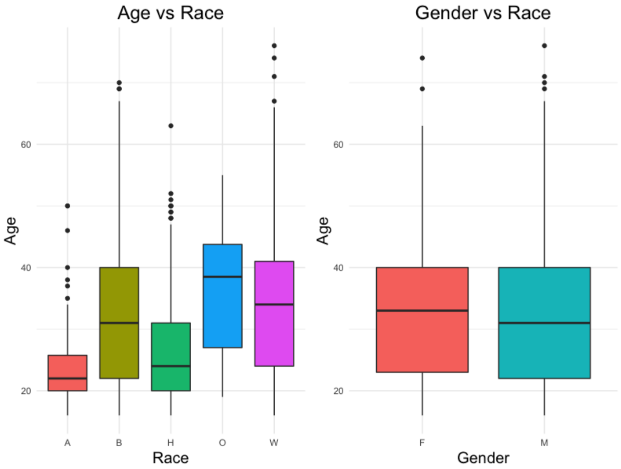

FIGURE 2.4. Before Cleaning

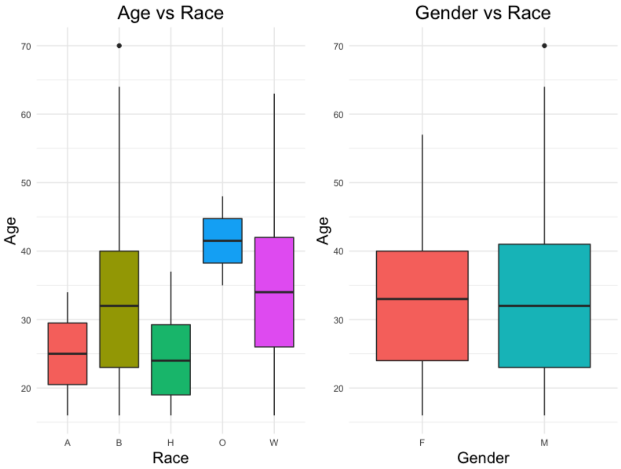

FIGURE 2.5. After Cleaning

The distribution by gender shows almost no difference, but we can see some changes when examining the plots for age by race. The interquartile range for the Asian subgroup increased, and the box plot for the Other subgroup shifted significantly as well.

With the new, clean data we began to examine other factors, like the number of images in the data set by gender and by decade of life.

TABLE 2.4. Number of Images by Gender

	1	2	3	4	5+	Total
Male	372	2,350	3,606	1,975	3,155	11,458
Female	85	478	712	352	532	2,159
Total	457	2,828	4,318	2,327	3,687	**13,617**

In Table 2.4, we see that the vast majority of arrested persons in our dataset are arrested multiple times. Only 457 of the 13,617 unique individuals were not arrested a second time. The most common number of images present in the dataset was 3, followed by 5+, indicating 5 or more images. The trend in image number likely continues to decline after 3, but this category sums the number of individuals with 5 images or more, making it larger than the category for only 4 images.

Table 2.5 summarizes the results when analyzing the number of images by decade of life.

TABLE 2.5. Number of Images by Gender and Decade of Life

	<20	20-29	30-39	40-49	50+	Total
Male	1,966	3,387	3,048	2,297	760	11,458
Female	294	605	687	473	100	2,159
Total	2,260	3,992	3,735	2,832	860	**13,617**

We were asked to consider the first arrest only, so this doesn't give us information about the total number of times an individual was arrested, rather the total number of unique individuals who were first arrested while in the specified age group.

We can see that the age group with greatest number of arrests for males is 20-29, while the age group with greatest number of arrests is 30-39 for females. This may seem to contradict our earlier histogram showing the most arrests in the below 20 age group, but we must take into account that the below 20 age group only includes persons from 16-20, which is a much smaller period than the rest of the groups. If we were to divide the data into age groups of 5 year periods, the below 20 age group would be the largest.

3. Regression Models to Predict Age

As we prepared to fit regression models to predict age from BIF features, we merged our cleaned data with the BIF data. There are over 2,500 Bio-Inspired Features that could be included, but to simplify the process and the computation, we only included the first 20 features for the models.

3.1. Polynomial Regression Diagnostics

The diagnostic plots in Figure 3.1 illustrate the results of the first linear model regressing on age. The Residuals vs Fitted plot is not straight, indicating possible nonlinear relationship not explained by the model. The Normal Q-Q plot isn't perfectly straight, but neither is it horrendously off, indicating residuals are likely normally distributed. The Scale-Location plot gives no strong evidence of heteroskedasticity. On the leverage plot, we can't make out any dashed lines indicating Cook's distance, which doesn't give us any information about potential influential points.

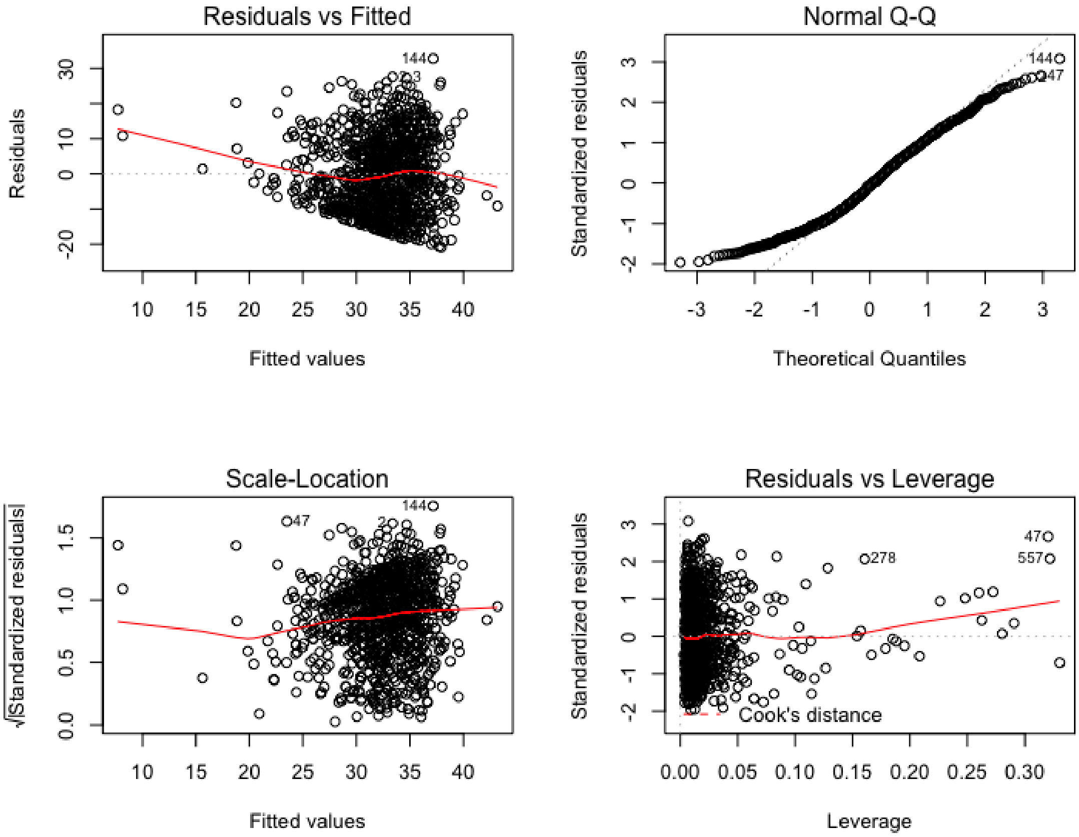

FIGURE 3.1. Linear Regression Diagnostic Plots

We performed polynomial regression with increasingly higher order to compare the results. The adjusted R-squared values increased as the order of the polynomials increased, indicating that a greater proportion of the variance of age was explained by the models as the order increased. There was also a general trend of more variables being considered statistically significant. The R-squared values remained very small up to quintic regression, but this is unsurprising considering that we are feeding our model 20 features out of 2,500 possible features, many of which were not found to be significant.

TABLE 3.1. Results of Polynomial Regression

	Linear	Quadratic	Cubic	Quartic	Quintic
Adjusted R-squared	0.08404	0.1046	0.1139	0.1184	0.1198
Significant Features	2	5	5	3	9

The diagnostic plots for the Quintic Regression shown in Figure 3.2 may lend some evidence to the increased Adjusted R-squared values when comparing Linear to Quintic regression. We can see a slight straightening in the line on the Residuals and Normal Q-Q plots, indicating a slightly better fit and closer to normally distributed residuals. The improvement is very small, however. Going further, I would be interested to see how the results compare when a majority of the significant features are being employed by a model.

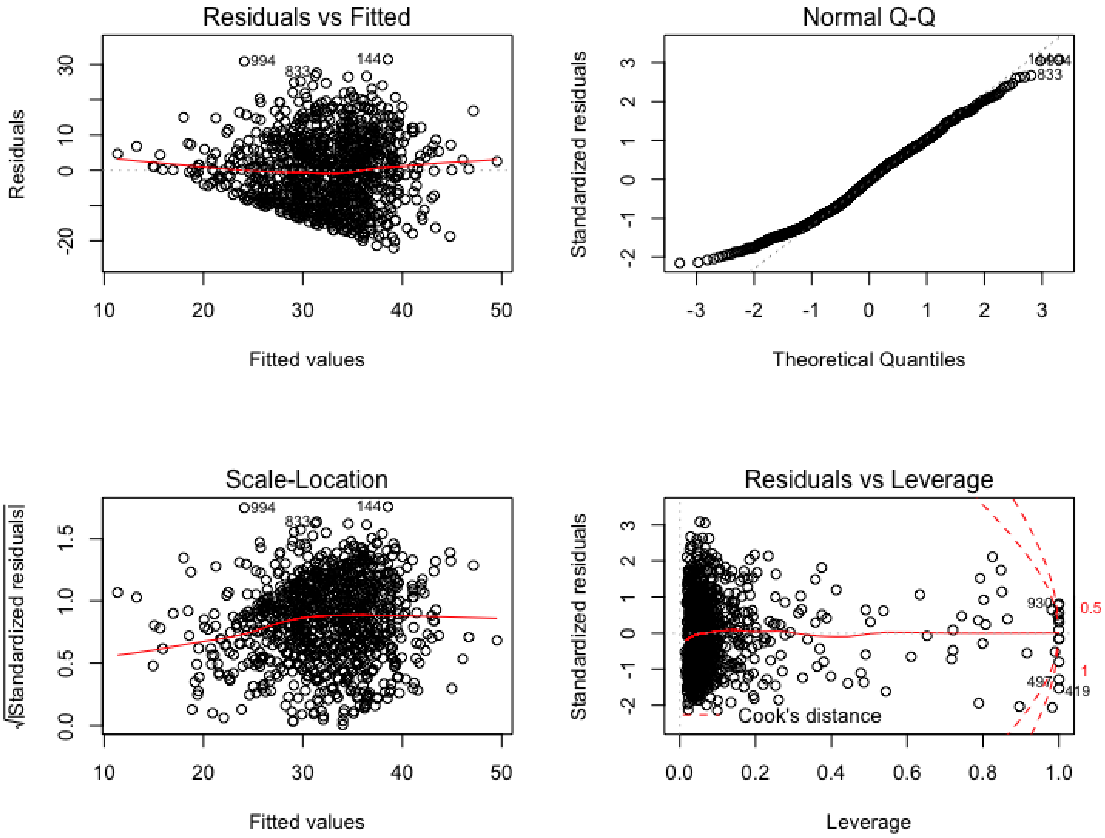

FIGURE 3.2. Quintic Regression Diagnostic Plots

3.2. Mean Squared Error: Comparing All Nine Models

Observing the results of the different degree polynomial regressions, we see the change in adjusted R-squared values beyond cubic regression is negligible. We will only consider the mean squared error (MSE) results for linear, quadratic, and cubic regression here.

As a group, the decision was made to train and test on the entire subset of the data, instead of withholding a portion of the data to test on. Thus, the results are not indicative of how the models would perform on new data. In fact, we observe some clear evidence of overfitting in the results.

As each model was trained and age predictions were made, MSE was calculated by:

$$MSE = \frac{1}{1000}\sum_{n=1}^{1000}(\hat{x_n} - x_n)^2 \tag{3.1}$$

where $\hat{x_n}$ is the predicted age of the nth face, and x_n is the actual age of the nth face.

Figure 3.3 illustrates the resulting MSE values for each model.

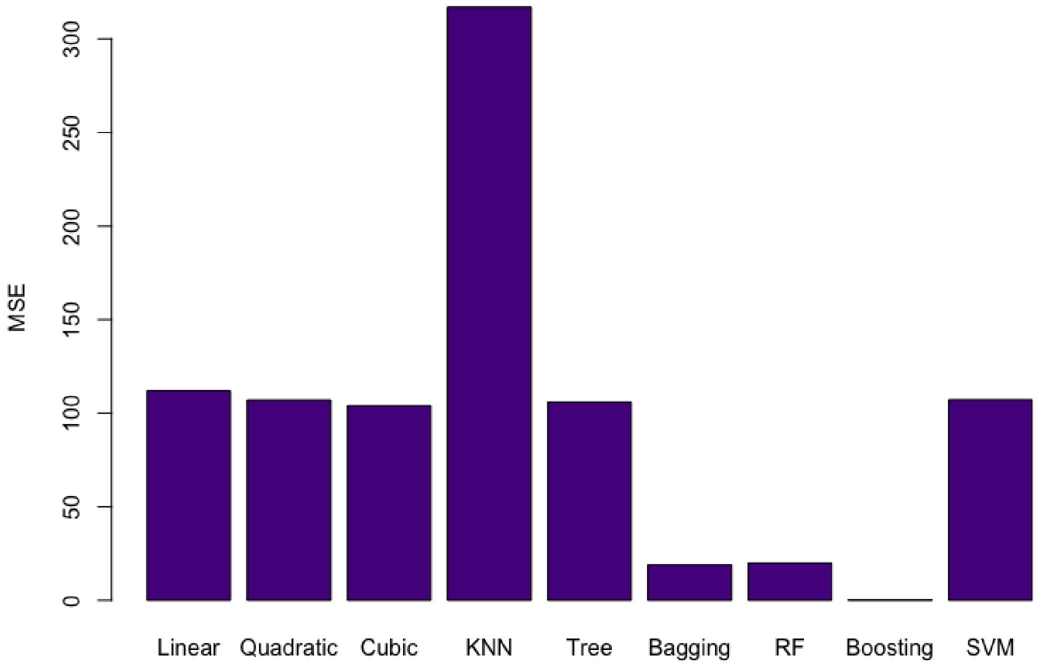

FIGURE 3.3. Mean Squared Error

It is apparent that those models that utilize bootstrapping (Bagging, Random Forest, and Boosting) performed the best. Boosting far outperformed any other model, attaining a MSE value close to 0. This is likely due to severe overfitting as a result of training and testing on the same data. These results were attained by adjusting tuning parameters to 5000 trees, shrinkage of 0.1, and interaction depth of 4. After some tuning, the results of the SVM model changed very little. The lowest MSE was attained with a radial kernel SVM, whose results are presented here. It should be noted that by training and testing on the same data, we allow the possibility of tuning many of these models to aggressively overfit the data. The decision tree, for example, could be extended to account for every data point separately, resulting in 0 MSE. The model whose performance is presented used the default values given by R.

Table 3.2 presents the same results as the bar graph numerically.

TABLE 3.2. Mean Squared Error

	Linear	Quadratic	Cubic	KNN	Tree	Bag	RF	Boost	SVM
MSE	112.1	107.4	104.0	317.2	106.0	19.4	20.5	0.097	107.1

We observe that K-Nearest Neighbors with $K = 3$ performed by far the worst, with a mean squared error almost three times that of the next highest model MSE.

To illustrate the danger of overfitting, Figure 3.4 shows the default decision tree whose MSE is given above, while Figure 3.5 shows an aggressively overfitted decision tree that predicts exact ages for each of the 1000 data points.

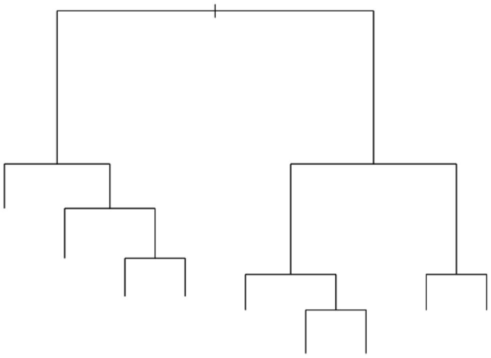

FIGURE 3.4. Default Tree

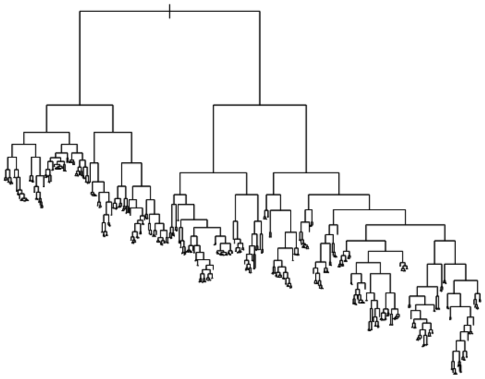

FIGURE 3.5. Overfitted Tree

The complexity will not transfer well to new data, giving poor results. For this reason, other methods of obtaining MSE, such as cross-validation, should be used.

4. Classification Models to Predict Gender

Having finished comparing model performance for age prediction, we moved on to training model for gender classification. We used 9 different models and compared them using 5-Fold Cross-Validation (5FCV) and Leave-One-Out Cross-Validation (LOOCV). An explanation of these methods is not given in this paper, but an understanding of them is recommended before proceeding. I elected to use the toy data set, composed of 1000 of the 55,134 images, for ease of computation. Furthermore, the algorithms for the models in R will not converge for a large amount of features, so the decision was made to reduce the 2,500 features to only the first 100 features for functionality in R.

To measure model performance, we examine accuracy, standard error, sensitivity, specificity, and run time for each model. After some tuning was performed, the SVM model was built with a linear kernel and cost set to 0.0001, KNN used $K = 7$, and Boosting interaction depth was adjusted to 4. All other parameters remained default.

4.1. 5-Fold Cross-Validation

We first employed 5-Fold Cross-Validation to compare the models' ability to classify images into genders, male or female. The data was randomized and 200 images were allocated to each of the 5 folds. The five accuracies for each model are visualized in a box plot in Figure 4.1 to show the magnitude and spread of the results.

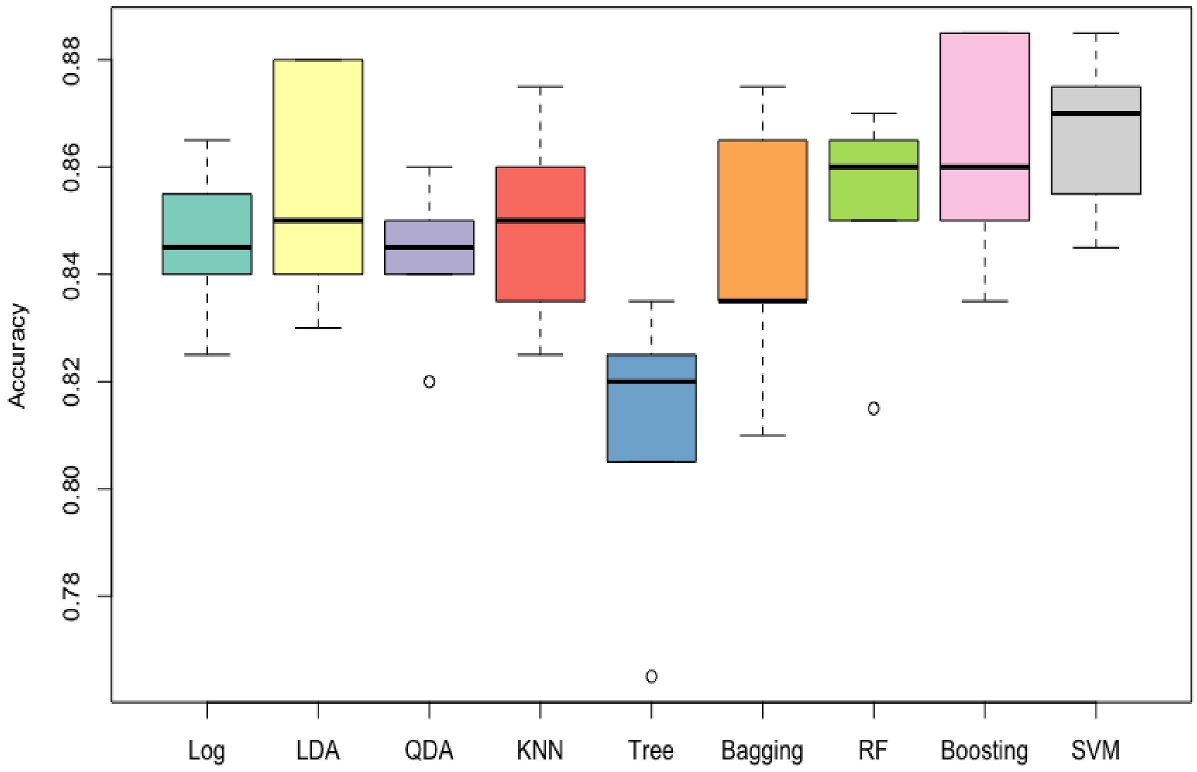

FIGURE 4.1. 5-Fold Cross-Validation Accuracies

The average accuracies of most of the models are comparable, with Boosting and SVM slightly outperforming the other models. It is clear from the plot that the Decision Tree model had a much lower accuracy then the other models.

The SVM and Boosting models had the highest mean accuracy and also had high sensitivity. SVM had relatively low standard error, but rather low specificity has well. Boosting had rather high standard error comparatively, but higher specificity as well. While their performances were comparable, the SVM model took less than half the time that the Boosting model required to run. Figure 4.2 illustrates these results. Once again, we see the comparatively poor performance of the Decision Tree model. It has the lowest accuracy and sensitivity, and highest standard error of any of the models.

Across every model, we find the trend of very high sensitivity and quite low specificity, likely due to the male-dominated dataset. Efforts to tune the models for rare event situations may improve results. It's interesting to note that the Quadratic Discriminant Analysis model resulted in a specificity of 1 and sensitivity of 0. The model achieved this by classifying every image as male.

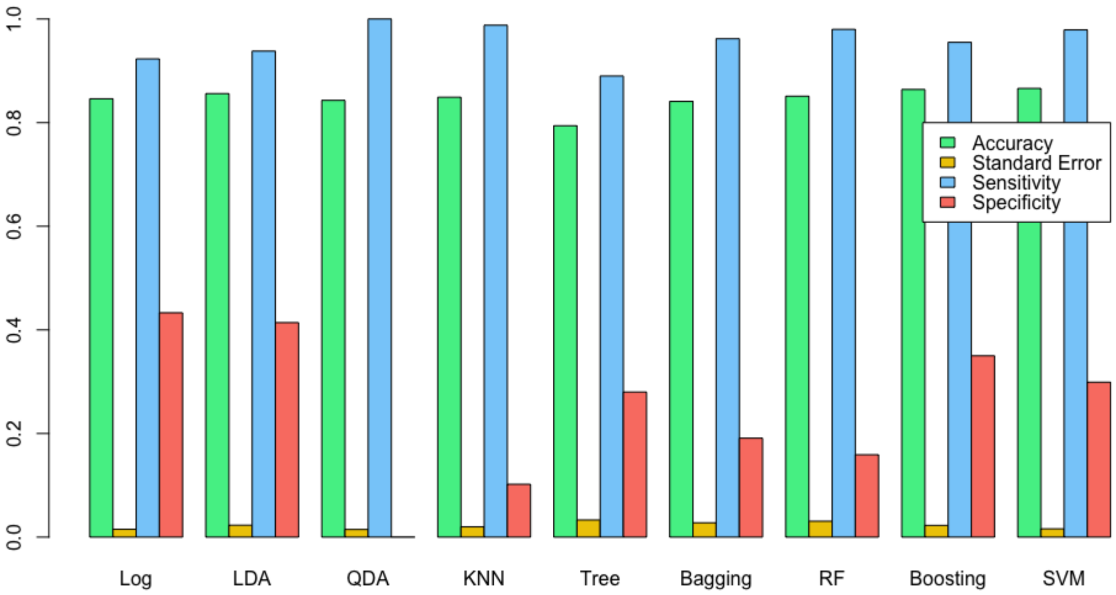

FIGURE 4.2. 5-Fold Cross-Validation Results

The same information is presented numerically in Table 4.1. The Decision Tree model is the only model with an accuracy that fell below 80%. Its standard error is highest at 0.033, closely followed by the Random Forest model at 0.031. The first five models are computationally more simple than the last four, as is evident in their run times. The first five models took less than a second to run cross-validation with 5 folds, while the other four took over ten seconds. The LDA model at 0.856 accuracy and 0.5 seconds of run time and the KNN model at 0.849 accuracy and 0.2 seconds of run time are the most cost-effective from a temporal perspective. This will have a larger impact when working with a large dataset.

TABLE 4.1. Comparison of Classification Methods Using 5-Fold Cross-Validation

	Log	LDA	QDA	KNN	Tree	Bag	RF	Boost	SVM
Accuracy	0.846	0.856	0.843	0.849	0.794	0.841	0.851	0.864	0.866
Standard Error	0.015	0.023	0.015	0.020	0.033	0.028	0.031	0.023	0.016
Sensitivity	0.923	0.938	1.000	0.988	0.890	0.962	0.980	0.955	0.972
Specificity	0.433	0.414	0.000	0.102	0.280	0.191	0.159	0.350	0.299
Time (s)	0.50	0.50	0.33	0.20	0.37	24.2	15.0	25.8	10.5

4.2. Leave-One-Out Cross-Validation

The results of each model when employing LOOCV instead of 5FCV showed little change in accuracy, but drastic change in standard error and run time across every model. The corresponding bar graph for LOOCV is given in Figure 4.3.

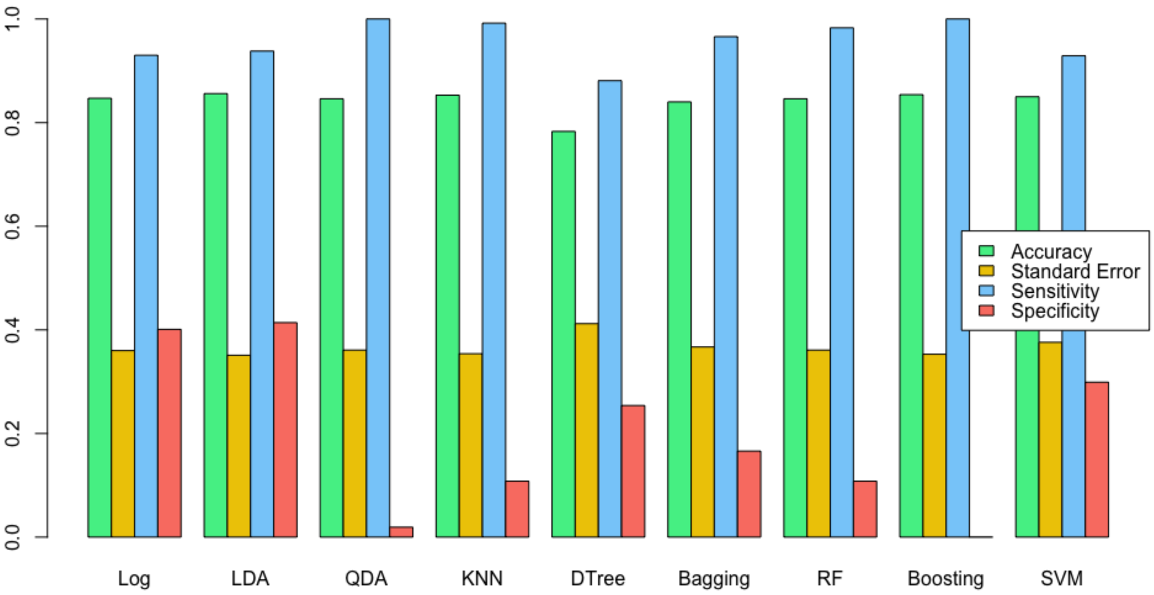

FIGURE 4.3. Leave-One-Out Cross-Validation Results

We see the same trends in accuracy and sensitivity that we saw in the 5-Folds Cross-Validation. Clearly, the standard error for every model is much higher, with Decision Tree still showing the highest error.

By studying Table 4.2, we see the most drastic change occurred in run time. Pay careful attention to observe the units in this table are in minutes, while Table 4.1 had units of seconds.

TABLE 4.2. Comparison of Classification Methods Using LOO-CV

	Log	LDA	QDA	KNN	Tree	Bag	RF	Boost	SVM
Accuracy	0.847	0.856	0.846	0.853	0.783	0.840	0.846	0.854	0.830
Standard Error	0.360	0.351	0.361	0.354	0.412	0.367	0.361	0.353	0.376
Sensitivity	0.930	0.938	1.000	0.992	0.881	0.966	0.983	1.000	0.929
Specificity	0.401	0.414	0.191	0.108	0.254	0.166	0.108	0.000	0.299
Time (min)	0.93	0.81	0.83	0.08	1.12	111.4	70.6	109.0	34.3

Once again, we see a huge difference in run time between the first five models and the last four. The models that employ bootstrapping require over an hour to complete. While the most accurate model was LDA for this test, the KNN model was the most cost-effective with a run time of only 0.08 minutes.

4.3. Comparison

While there may be specific occasions that warrant use of Leave-One-Out Cross-Validation as opposed to K-Folds Cross-Validation, in this situation we see a massive increase in computational rigor with little to no performance improvement. In fact, the accuracy of the SVM model dropped from 0.866 to 0.830 when we switched to LOOCV. With such a higher cost and low potential for improvement, I recommend employing 5 or 10-fold cross-validation over LOOCV.

When considering which model to use, if cost is not of consequence, the SVM and Boosting models performed with the highest accuracy, but the SVM model was considerably faster. For cost efficiency, the LDA and KNN models performed with impressive accuracy in a much smaller amount of time. All these factors should be considered when making a decision.

5. Latent Class Analysis

Latent Class Analysis (LCA) is a method that allows us to characterize categorical ”latent,” or unobserved variables by analyzing relationships between many observed categorical variables (McCutcheon, 1987). In preparation for this project, we received new data containing levels of exposure to nine chemicals, labeled $m1, m2, ..., m9$, for each individual in the first 500 rows of our toy data set, as well as whether or not each person was convicted.

5.1. LCA Methodology

We began by comparing our subset to the rest of the data to determine if it was a representative sample. Table 5.1 displays the demographic results of the proportion of the data belonging to each subgroup.

TABLE 5.1. Subset for LCA: Proportions

	Black	**W**hite	**A**sian	**H**ispanic	**O**ther	**Total**
Male	0.668	0.164	0.002	0.002	0.000	0.836
Female	0.122	0.040	0.000	0.002	0.000	0.164
Total	0.790	0.204	0.002	0.004	0.000	**1.000**

By comparing our results to Table 5.2, we can see that the proportions for each subgroup split by gender and race are pretty close to one another, with the exception of Hispanic Males, who are not well-represented in our subgroup. However, the other subgroups all match well, showing that our sample subset is fairly representative and can be used to train our model. We found similar results when we compared the age distribution of our subgroup to that of the total population.

TABLE 5.2. Full Data: Proportions

	Black	**W**hite	**A**sian	**H**ispanic	**O**ther	**Total**
Male	0.648	0.151	0.003	0.037	0.001	0.841
Female	0.109	0.046	0.000	0.002	0.001	0.159
Total	0.758	0.197	0.004	0.039	0.002	**1.000**

We employed an unsupervised LCA approach to establish classes to which each individual was assigned based on their levels of chemical exposure. We suspected some of these unknown chemicals may have strong correlations that may give us greater insights into the classes to be formed. Figure 5.1 displays a heat map of the correlations of the chemical averages.

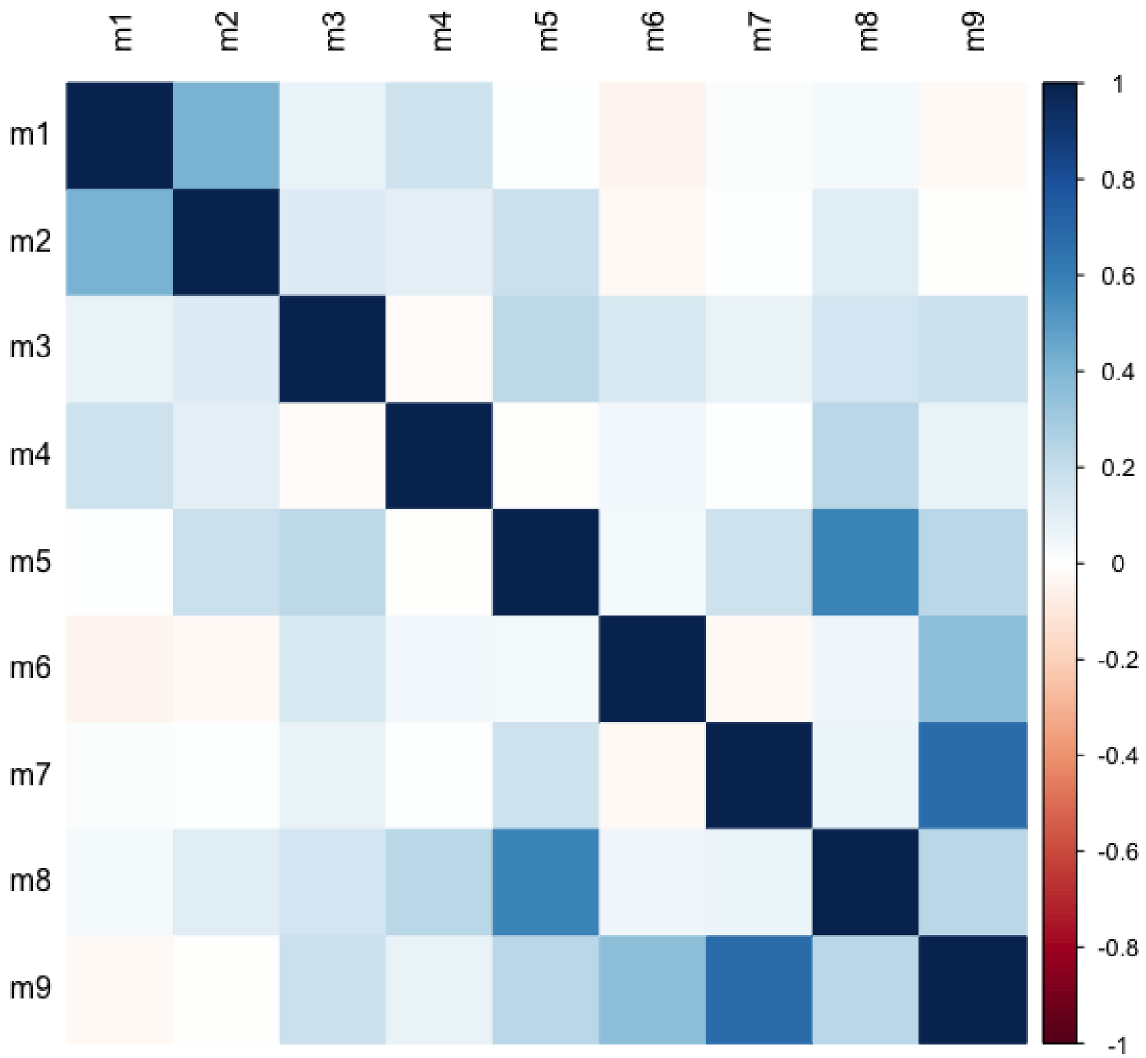

FIGURE 5.1. Heat Map of Chemical Correlations

Most correlations are small. The most notable are between $m7$ and $m9$ at 0.673, $m5$ and $m8$ at 0.582, and $m1$ and $m2$ at 0.414. These correlations will prove significant in the results of the distribution of chemical exposure in the latent classes.

Before training the model, we dichotomized the levels of chemical exposure by assigning 1 if the level of any chemical was below the median level for that chemical, and 2 if it was above the median.

To decide on the number of latent classes to use, we tried values from 2 to 10 and compared different measures of goodness-of-fit for the each model. Different measures gave different results. Using G^2 or χ^2, indications suggest 10 classes would be optimum. However, interpretability would be very difficult with 10 different classes. BIC suggested 3 classes would be the best, while AIC implied 6 or 7 classes would produce the best goodness of fit. I elected to meet in te middle with 4 classes. This should improve interpretability and maintain decent goodness-of-fit for the model. The results of these tests are presented in Table 5.3.

TABLE 5.3. Goodness-of-Fit Measures

Classes	Log Likelihood	AIC	BIC	G^2	χ^2
2	-3056.222	6150.444	6230.521	691.836	653.139
3	-2993.575	6045.151	**6167.375**	566.543	525.154
4	-2973.026	6024.051	6188.421	525.443	495.887
5	-2959.818	6017.636	6224.152	499.028	487.474
6	-2944.984	**6007.967**	6256.629	469.359	447.177
7	-2934.505	**6007.009**	6297.817	448.401	415.744
8	-2928.569	6015.139	6348.093	436.531	413.093
9	-2918.572	6015.144	6390.244	416.536	385.562
10	-2909.000	6016.000	6433.246	**397.392**	**362.941**

We trained the LCA model with 4 classes and calculated the posterior probability to determine how confident we are in the assignments for each subject. Results were promising, with the 1st Quartile of the probabilities at 0.81, the Median at 0.97, and the 3rd Quartile at 1.00. A histogram of the resulting posterior probabilities can be found in Figure 5.2.

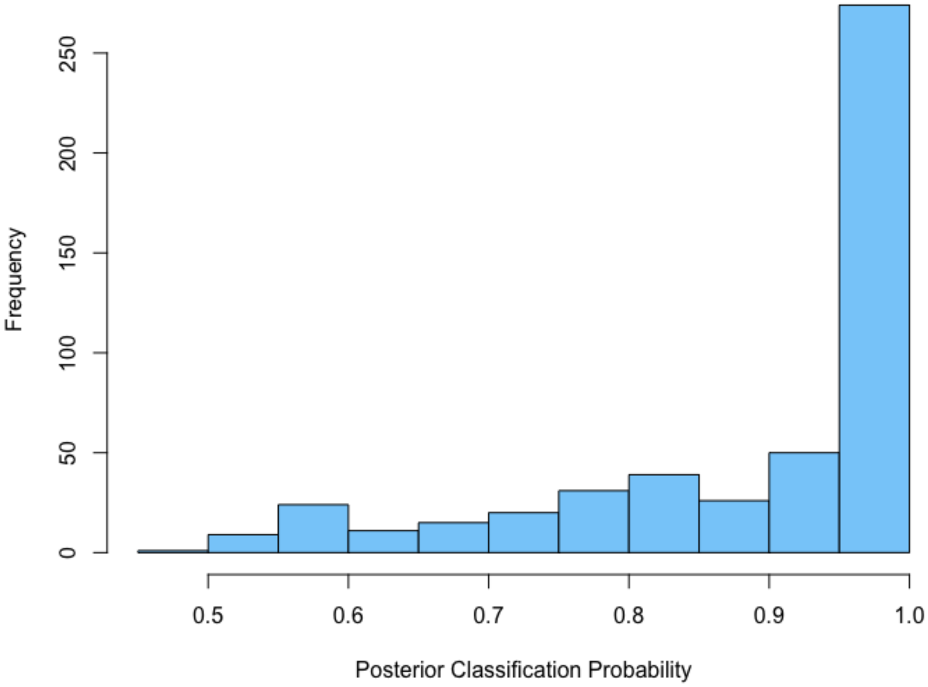

FIGURE 5.2. Posterior Probabilities

5.2. LCA Results

The resulting classes with their chemical exposure levels are presented in Figure 5.6. An attempt was made to label these classes by their defining characteristics. In other figures, class 1 corresponds to ”High m8, m9”, class 2 corresponds to ”Low m5, m8”, class 3 corresponds to ”High Exposure”, and class 4 corresponds to ”Low Exposure.” It is of interest to note that the defining characteristics of one class are low exposure to m5 and m8, which we previously found to have a high correlation.

Next, we assessed the relationship between these classes and gender, race, and age. Of the 500 individuals in our subset, 100 were assigned to class 1, 106 to class 2, 57 to class 3, and 237 to class 4.

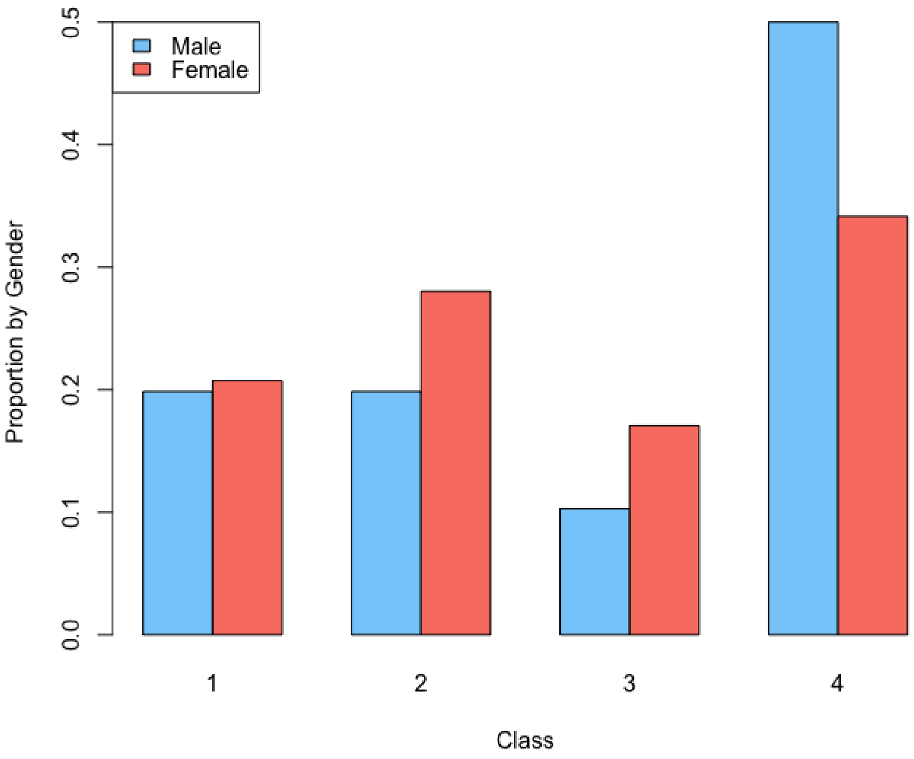

FIGURE 5.3. Proportion of Gender in Classes

In Figure 5.3 we can see that the class 4 favors males slightly more than it favors females, but the proportions of gender in the other classes are fairly evenly spread. We did not find any significant relationship between the distribution of class and gender.

Results were similarly inconclusive in our search for a relationship between class and race. In Figure 5.4, the proportions of Hispanics and Asians seem significant at first, but we must keep in mind that there were only two Hispanic people in the data set and only 1 Asian person. The only races for which we have a significant sample size, Black and White, carry very similar proportions in each of the classes, as illustrated by Figure 5.4.

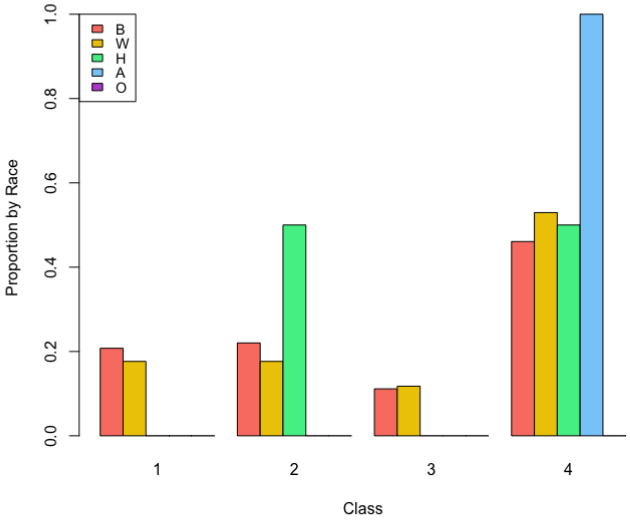

FIGURE 5.4. Proportion of Race in Classes

We on to compare the distribution of age in each class. The histograms that resulted were very similar to the histogram for the over all data set. A smaller proportion of individuals over the age of 50 was assigned to Class 3 than the other classes, while Class 2 had a heavier concentration of individuals between the ages of 30 and 40 than the other classes. However, the overall pattern remains the same for each class.

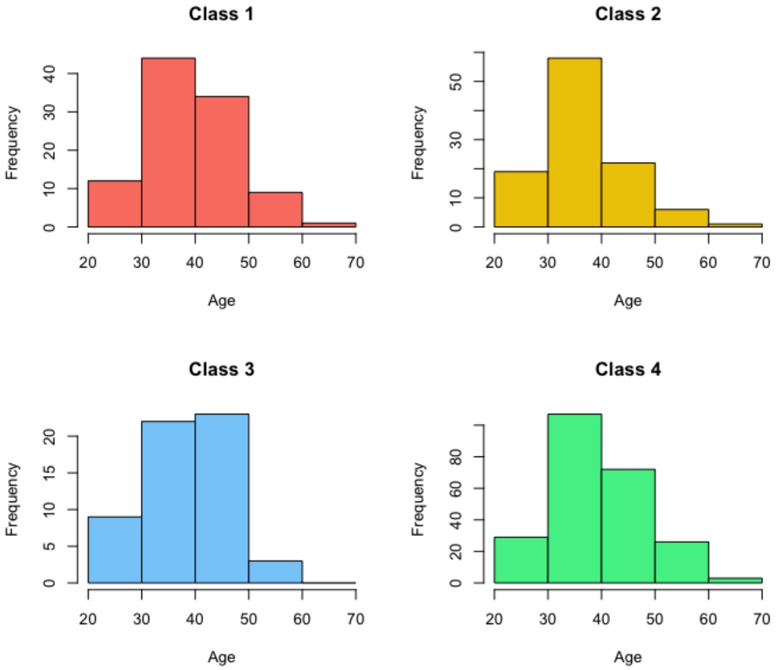

FIGURE 5.5. Histograms of Age by Class

Note the mean percentages of individuals with high exposure across all biomarkers (chemicals) under the title of each of the classes. The titles were chosen to describe what appears to be the defining characteristic of each class.

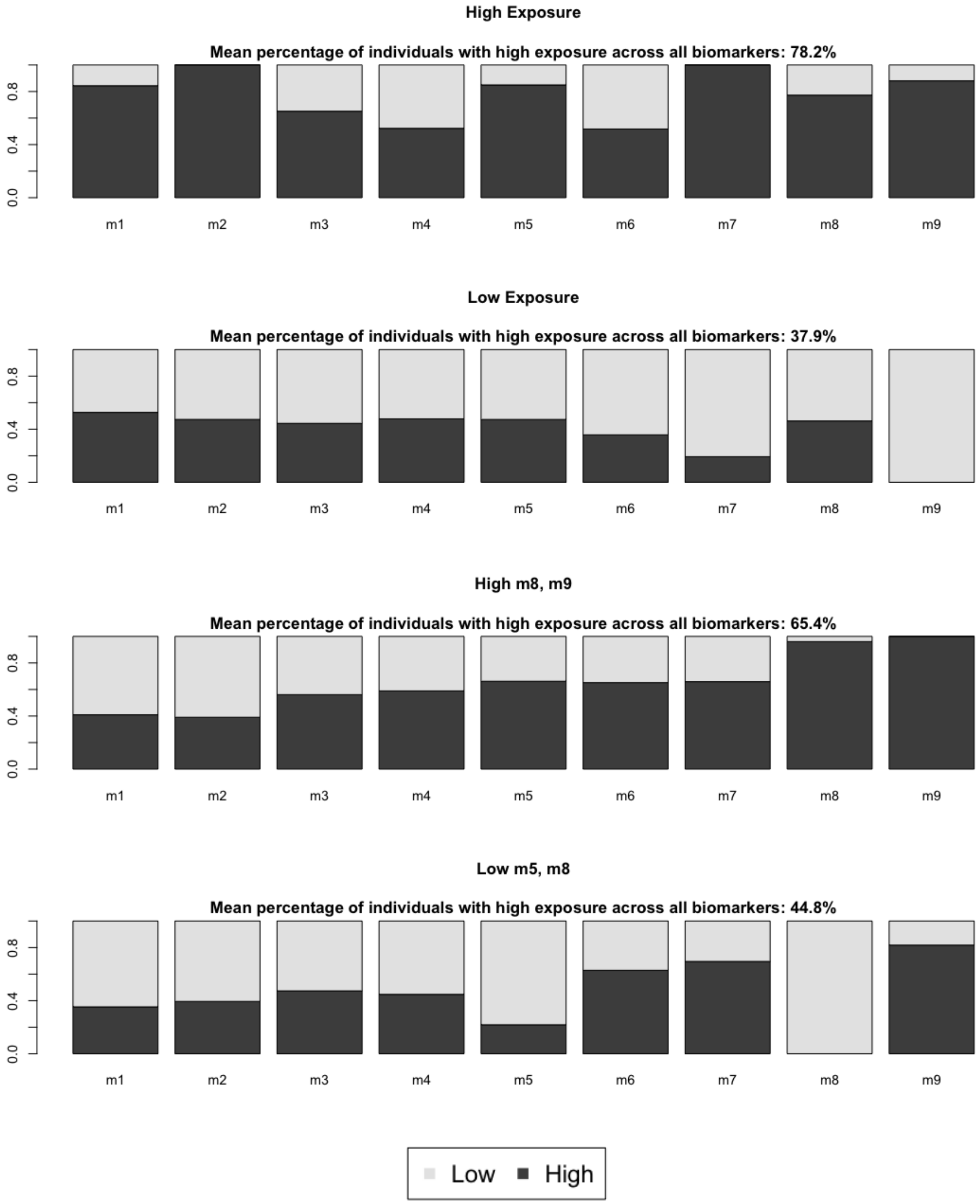

FIGURE 5.6. Latent Classes

We further attempted to determine if a relationship existed between the latent classes and age or between the classes and whether the individual was convicted.

We used linear regression to predict age by class and achieved a Mean Squared Error of 65.016. The model predicts one of four ages depending on what class the individual is in. The ages are 38.636, 38.968, 39.300, and 39.632, which all fall close to the mean age of the sample group. Clearly this model isn't ideal, but it performed as expected considering the limitation of 4 classes.

We then attempted to use a logistic regression model to predict whether or not each person was convicted. The model did not show an excellent performance with an accuracy of 0.764. This isn't abysmal, but upon further inspection, we observed the model was simply predicting that everyone was convicted, which was true 76.4% of the time. Perhaps another model geared toward rare-event detection would perform better in this situation, or perhaps the latent classes don't provide a foundation for determining conviction status.

6. Principal Component Analysis

In sections 3 and 4 we trained models on a small subset of our dataset and tested their performance in gender and age prediction. In this section, we use PCA to enable those same models to use all the data available to use for training, without requiring extremely high computation cost.

Principal Component Analysis is a dimension reduction technique that relies on extracting ”principal components” from a data set and using a number of those principal components for statistical learning instead of the original features of the dataset. To decide how many principal components to use for our models, we examined what proportion of the variance in the data was explained by the principal components. There is a trade-off between variance explained and computational cost. We elected to use 108 principal components, as it was low enough to use with our models without too much trouble and it was the smallest number of principal components that accounted for at least 80% of the variance.

6.1. Gender Classification

Once again, we employed 5-Fold Cross-Validation and Leave-One-Out Cross-Validation to measure model performance. Calculating principal components added a fair amount of computational complexity, as it had to be done for each fold before training the models.

Table 6.1 presents our results numerically, while Figures 6.1 and 6.2 display graphical results.

TABLE 6.1. Comparison of Classification Methods Using PCA

	Log	LDA	QDA	KNN	Tree	Bag	RF	Boost	SVM
5FCV Accuracy	0.882	0.922	0.843	0.870	0.782	0.846	0.843	0.873	0.881
5FCV Standard Error	0.021	0.014	0.012	0.040	0.033	0.012	0.012	0.007	0.017
LOOCV Accuracy	0.897	0.924	0.844	0.868	0.782	0.851	0.843	0.876	0.893
LOOCV Standard Error	0.391	0.331	0.365	0.411	0.456	0.386	0.365	0.371	0.290
5FCV Time (min)	0.010	0.004	0.010	0.002	0.008	0.472	0.268	0.165	0.144
LOOCV Time (min)	1.374	0.851	0.757	0.023	1.794	129.6	73.03	40.63	157.0

We note that accuracies using both 5-Fold and LOO Cross-Validation increased using PCA compared to other methods. In both situations, the LDA model performed the best, with the highest accuracy, low standard error, and low computation time. once again, we see our Decision Tree model produced the lowest accuracy, while SVM and Boosting models performed very well. The QDA, Bagging, and Random Forest models didn't improve with PCA as the other models did.

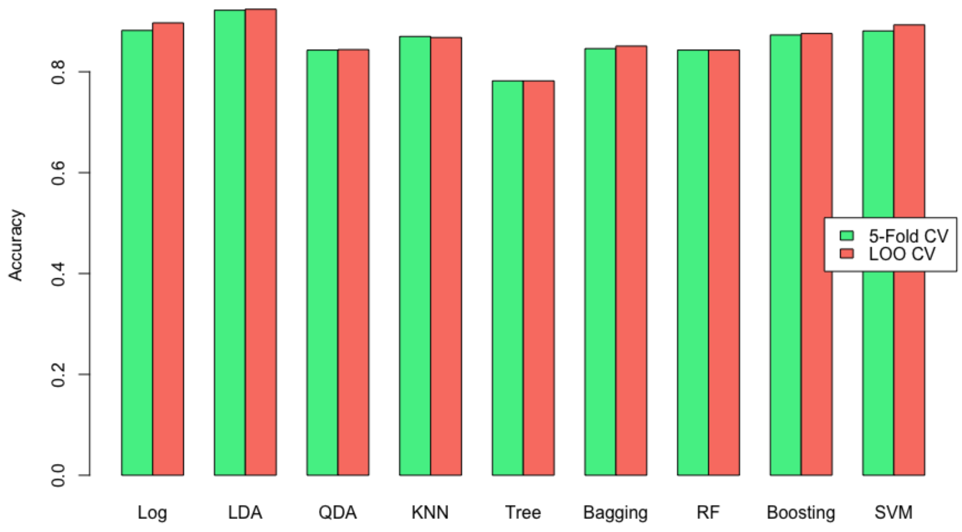

FIGURE 6.1. Accuracy with PCA

We can see from Figure 6.1 that there is hardly any improvement at all in accuracy from the 5-Folds to Leave-One-Out models. Considering the greatly increased temporal complexity and standard error, as shown in Figure 6.2, it is advisable to use 5-Folds Cross-Validation for these models in this situation.

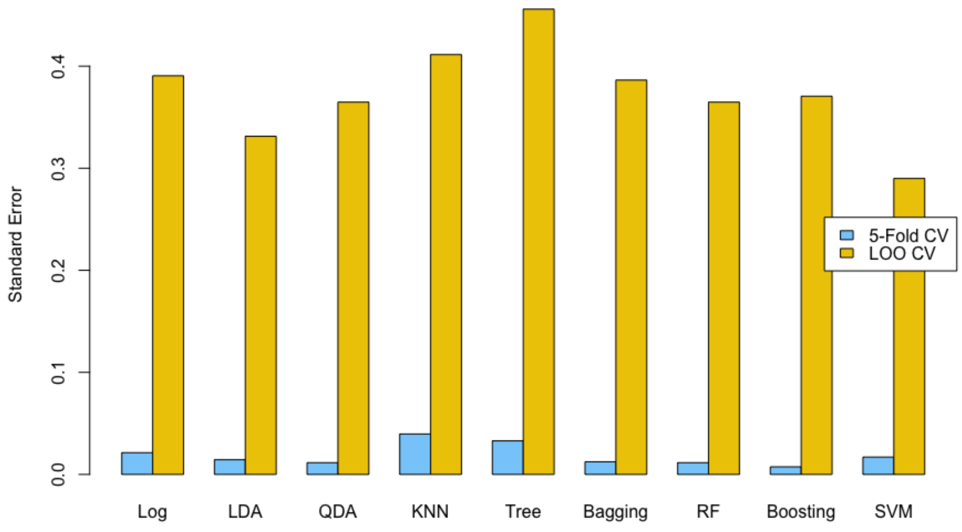

FIGURE 6.2. Standard Error with PCA

After seeing the effects of PCA on our model success, we moved to try KPCA, or Kernel Principal Component Analysis. KPCA projects the data onto a higher-dimensional space in an attempt to make it linearly separable so PCA can work well. We made use of the Gaussian Radial Basis Function Kernel to perform PCA and compare the results. After some tuning, our kernel parameter sigma was set to 1e-10. Results for accuracy with this method are shown in Figure 6.3.

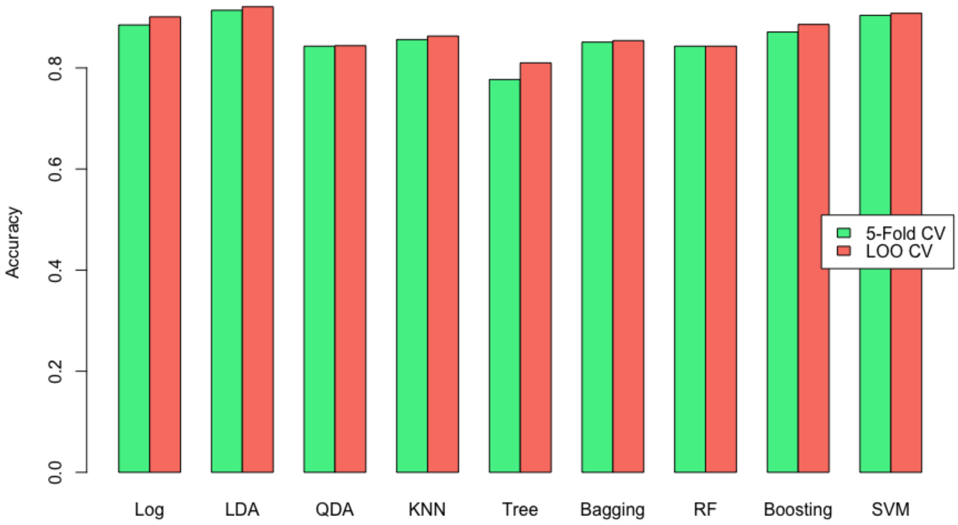

FIGURE 6.3. Accuracy with KPCA

We can see a pattern similar to that of the PCA results. The Tree model remains the worst and the LDA model the best in terms of accuracy. Table 6.2 shows the exact numbers. Some models improved in accuracy from the PCA method, while some performed worse. There were no dramatic changes, however. Accuracy, standard error, and even time elapsed were comparable between KPCA and PCA.

TABLE 6.2. Comparison of Classification Methods Using KPCA

	Log	LDA	QDA	KNN	Tree	Bag	RF	Boost	SVM
5FCV Accuracy	0.885	0.914	0.843	0.856	0.777	0.851	0.843	0.871	0.904
5FCV Standard Error	0.018	0.016	0.012	0.018	0.031	0.014	0.012	0.007	0.017
LOOCV Accuracy	0.901	0.921	0.844	0.863	0.810	0.854	0.843	0.886	0.908
LOOCV Standard Error	0.389	0.352	0.253	0.400	0.473	0.317	0.353	0.356	0.376
5FCV Time (min)	0.010	0.004	0.003	0.002	0.009	0.522	0.268	0.168	0.014
LOOCV Time (min)	1.326	0.767	0.665	0.020	1.973	144.7	74.19	40.78	3.412

Figure 6.4 displays visually the standard error results with KPCA. Standard error follows a pattern similar to that of the PCA results.

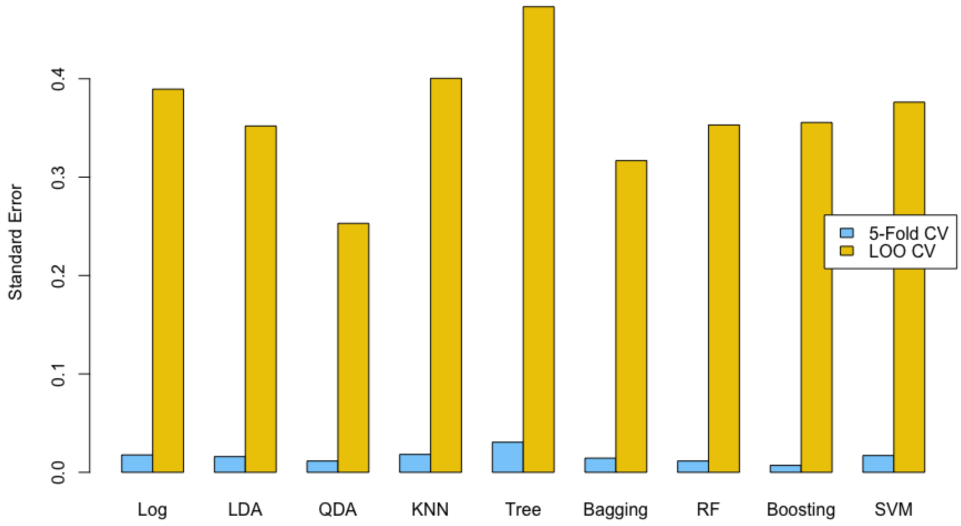

FIGURE 6.4. Standard Error with KPCA

6.2. Age Prediction

We went on to compare the methods of PCA and KPCA on our regression models for age prediction. We continued to use 108 principal components for training and elected to use a validation set approach to testing, instead of testing and training with the same data as we did before. Testing on the training data yielded untrustworthy results, so we hoped to have more acurrate reflections of model performance here.

Tuning yielded the following parameter values for PCA: $k = 3$ for KNN, interaction depth of 1 for boosting, and a cost of 0.0001 on a linear kernel for SVM. For KPCA, we used $k = 3$ for KNN, interaction depth of 2 for boosting, a cost of 10 on a linear kernel for SVM, and a sigma value of 1e-8 for the KPCA model. The Mean Squared Errors attained by each model for both PCA and KPCA are given in Table 6.3.

TABLE 6.3. Mean Squared Error

	Linear	Quadratic	Cubic	KNN	Tree	Bag	RF	Boost	SVM
MSE with PCA	80.70	87.52	92.49	188.4	198.5	162.2	166.4	99.8	80.39
MSE with KPCA	88.97	103.9	136.5	226.1	176.4	150.3	154.5	101.6	88.67

These values are also compared visually in Figure 6.5. For most models, KPCA has a higher mean squared error, but for Decision Tree, Bagging, and Random Forest, the PCA method results in a higher MSE.

Overall, neither of the methods produced very good results for age prediction. Perhaps with a greater number of principal components or more time spent tuning models the results could be improved. The Linear regression and SVM models performed the best with lowest MSE.

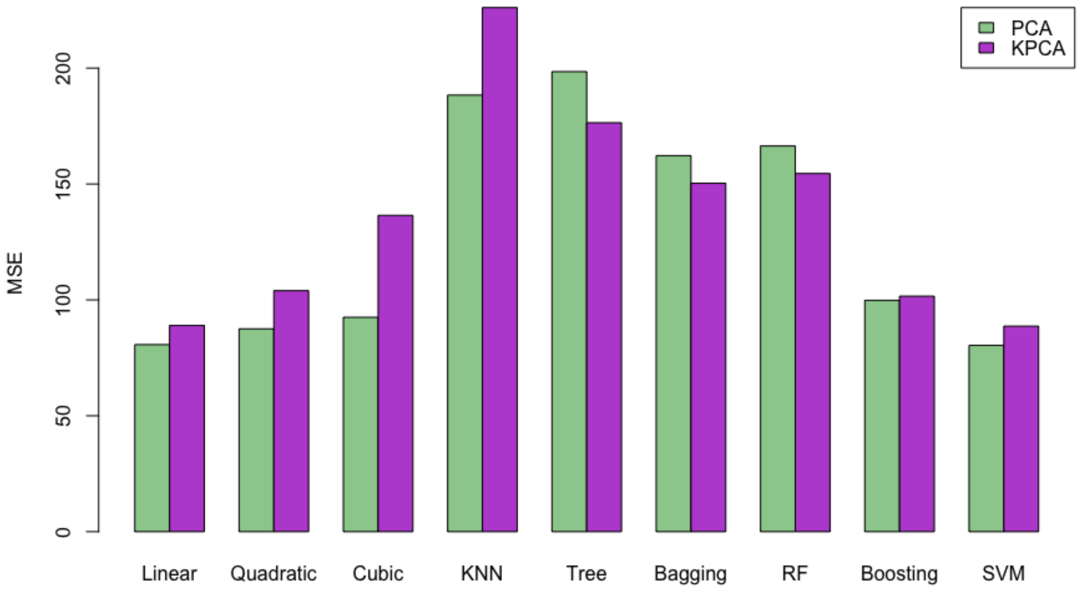

FIGURE 6.5. Mean Squared Error with PCA & KPCA

7. Conclusion

The MORPH-II dataset has a number of inconsistencies. Upon cleaning up these errors, an analysis of the data showed that the the most frequently occurring race in the dataset was by far Black, followed by White. Males far outnumbered females in both total images and total unique individuals. The youngest to be arrested in the dataset was 16, with the oldest at 77, and the average age 32.62. The data indicated that the most common age to be arrested according to this sample was from 16-20 years old. Furthermore, the most common amount of times to be arrested is 3.

Results from our foray into age prediction with regression models pointed to Boosting as the optimal model to use when tuned, but this is a reflection of overfitting. It was decided as a group that we were to train and test on the same data, which allowed for such overfitting to occur. More accurate reflections of model performance should be attained through tests using cross-validation and employing more than the first 20 BIF features.

Each of the nine classification methods performed pretty well with only 100 features given to classify gender, although the decision tree model was clearly less accurate than the rest. The SVM model achieved the highest accuracy at 0.866, but many other models were not far behind. The KNN model was by far the fastest and most cost-efficient model.

As we moved on to using dimension reduction techniques such as PCA and KPCA, we were able to use more than those 20 or 100 BIF features and achieve better results. For age prediction, the SVM model performed the best with MSE of 80.39. For gender classification, the LDA model achieved the highest accuracy with 92.4% using PCA and LOOCV.

The Latent Class Analysis was intriguing, and the classes had well-defined characteristics, but we did not find any relationship between the levels of chemical exposure and any other covariate, such as age, gender, race, or conviction status. Efforts to predict age or conviction status by class were fruitless.

Further tests should be undergone to compare different tuning parameters that were not considered, as well as alternate kernels to be used for KPCA. With more time and resources, accuracy much higher than the 92% seem here could be achieved. Overall, for this project it is recommended to use 5-Fold Cross-Validation over LOOCV due to the massive increase in cost with little potential for improvement and to utilize dimension reduction methods instead of small, randomized subsets of the dataset for training models. With these methods, using BIF features for gender classification and age prediction is optimized.

8. Acknowledgements

Thank you to Dr. Yishi Wang and Dr. Cuixian Chen at UNCW for their mentorship throughout this project. Additional thanks go to the National Science Foundation for funding this program (Grant #1659288).

References

McCutcheon, A. L. (1987). *Latent class analysis*. Number 64. Sage.

Ricanek, K. and Tesafaye, T. (2006). Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345. IEEE.

(D. Johnston) Department of Mathematics and Statistics, The University of North Carolina Wilmington, Wilmington, NC 28403, USA

E-mail address, Corresponding author: `drew.johnston13@gmail.com`