Advanced Statistics: Application of supervised and unsupervised methods to biological data

Daniel Hill

2025-04-27

Abstract: Biological data with twenty features and four categorical class labels is explored and analysed using advanced statistical techniques in this report. Both supervised and unsupervised methods were implemented and evaluated, and the broad selection of models includes logistic regression, support vector machine, random forest, agglomerative hierarchical clustering, and gaussian mixture modelling. Comparing the results achieved using a selection of models with different underlying principles gives insight into the nature of the data. For example, the success of model-based clustering compared to hierarchical clustering and tree-based learning suggests the lack of hierarchy among the categorical labels, and the success of factor analysis as a dimensionality reduction technique suggests the presence of underlying biological mechanisms leading to several of the features arising together. Models achieving over 90% accuracy were produced, but all models performed notably worse at separating one of the categories that overlapped the other three.

Contents

1 Introduction **4**

2 Methods **4**

2.1 Data Description 4

2.2 Exploratory Data Analysis Approach 4

2.3 Supervised Learning Methods 9

2.3.1 logistic regression - including class weighting and L2 regularization, and feature selection. 9

2.3.2 Random forest, including feature selection and tuning of mtry. 10

2.3.3 SVM, with feature selection and tuning of kernel selection, gamma and cost parameters. 10

2.4 Unsupervised Learning Methods 10

2.4.1 Agglomerative hierarchical clustering, including tuning of linking metric. . . 10

2.4.2 Gaussian mixture model based clustering, including selecting a model, regularization using shrinkage parameter, and dimensionality reduction using factor analysis. 11

3 Results **11**

3.1 Exploratory Data Analysis Findings 11

3.2 Supervised Learning Results 12

3.2.1 Logistic Regression 12

3.2.2 Random Forest 16

3.2.3 SVM 22

3.3 Unsupervised Learning Results 22

3.3.1 Agglomerative Hierarchical Clustering 22

3.3.2 Gaussian Mixed Model Clustering 22

4 Discussion **22**

5 Conclusion **22**

6 References **22**

List of Figures

1 **Feature and Class Correlation Matrix:** *highlighting relationships between variables and relationships with catagorical labels* 5

2 **Distributions within feature columns:** *histograms showing scale and skewness of data* 6

3 **Distributions within feature columns:** *Boxplots by class label by feature* . . . 7

4 **PCA plotting of class labels:** *scatterplot showing clustering tendency of catagorical classes* 8

5 **Simple Logistic Regression ROC Plot:** *ROC plot for Class D vs Not Class D* 13
6 **Regularized Logistic Regression** 16
7 **Optimal mtry for Random Forest** 21

List of Tables

1 Feature Descriptions 4
2 Hopkins Statistic Scores 9
3 Simple Logistic Regression Confusion Matrix 12
4 Simple Logistic Regression Overall Statistics 12
5 Simple Logistic Regression Statistics by Class 12
6 Simple Logistic Regression with selected features Confusion Matrix 13
7 Simple Logistic Regression with selected features Overall Statistics 13
8 Simple Logistic Regression with selected features Statistics by Class 14
9 Simple Logistic Regression with selected features Confusion Matrix 14
10 Simple Logistic Regression with selected features Overall Statistics 14
11 Simple Logistic Regression with selected features Statistics by Class 14
12 Weighted Logistic Regression Confusion Matrix 15
13 Weighted Logistic Regression Overall Statistics 15
14 Weighted Logistic Regression Statistics by Class 15
15 Logistic Regression Performance with Different Regularization Parameters 15
16 Confusion Matrix for Best Regularized Logistic Regression Model (Decay = 0.001) 16
17 Basic Random Forest Confusion Matrix 17
18 Basic Random Forest Overall Statistics 17
19 Basic Random Forest Statistics by Class 17
20 Basic Random Forest Feature Importance 17
21 Random Forest (5 least important features removed) Confusion Matrix 18
22 Random Forest (5 least important features removed) Overall Statistics 18
23 Random Forest (5 least important features removed) Statistics by Class 18
24 Random Forest (5 least important features removed) Feature Importance 19
25 Random Forest (5 least important features removed) Confusion Matrix 19
26 Random Forest (5 least important features removed) Overall Statistics 19
27 Random Forest (5 least important features removed) Statistics by Class 19
28 Random Forest (5 least important features removed) Feature Importance 20
29 Tuned Random Forest Confusion Matrix 21
30 Tuned Random Forest Overall Statistics 21
31 Tuned Random Forest Statistics by Class 21

1 Introduction

This report takes a moderately sized biological dataset containing 3000 observations, where each observation having 20 numeric varables and one catagorical label. These data are explored thoroughly before being used to train and test an array of statistical modelling techniques.

This is an interesting project because the origin and meaning of the variables in the data are completely unknown - an unusual scenario in the data science field, where usually it is the domain knowledge and problem context that inform the selection and implementation of statistical methods. Here, with this relationship reversed, algorithms have been chosen so that the evaluation of their performance can attempt to uncoverthe underlying biological significance of the variables.

Achieving meaningful results in this task shows the importance of supervised and unsupervised learning to this field, where classification algorithms can build valuable models that have high impact on society such as disease diagnosis models, and unsupervised learning techniques can create breakthoughs in identifying clusters of data that lead to new discoveries and classifications.

2 Methods

2.1 Data Description

Each of the 3000 observations has 20 numeric features and a label placing it in one of four catagorical classes. Though exploritory data analysis, two groups of correlated features were identified. Outliers were identified and removed using z score method. One feature transformed using logarithm to create a more normal distribution to improve the performance of models. All features were scaled and centered. After the preprocessing of the data, 2776 usable observations remained.

Bootstrap sampling was used to create a larger dataset so that the performance of the models could be compared between the original and bootstrapped data.

2.2 Exploratory Data Analysis Approach

find distributions within each feature, look for correlations between features, scatter between plots that features that have high correlation to the catagorical label or another feature, use PCA to visualize all data together, calculate hopkins statistic to determine the clustering tendency of the data

Table 1: Feature Descriptions

Variable Name	No. missing values	mean	Std deviation	min	25th %ile	median	75th %ile	max
X1	2	9.876	0.764	6.840	9.356	9.872	10.399	12.355
X2	0	10.151	1.040	6.538	9.445	10.138	10.855	14.021
X3	1	8.861	0.871	6.424	8.243	8.847	9.445	12.216
X4	2	8.939	1.275	3.875	8.088	8.926	9.805	13.351
X5	0	13.853	0.942	10.527	13.236	13.858	14.492	16.557

X6	0	8.151	1.026	4.815	7.447	8.134	8.856	11.871
X7	1	0.426	0.278	0.000	0.185	0.375	0.678	1.301
X8	3	0.234	0.197	0.000	0.102	0.170	0.300	1.230
X9	0	0.717	0.247	0.006	0.532	0.751	0.888	1.679
X10	0	0.378	0.155	0.000	0.281	0.372	0.469	1.199
X11	1	9.175	1.087	6.031	8.412	9.077	9.860	13.027
X12	0	11.930	0.977	8.046	11.290	11.913	12.594	15.478
X13	1	8.228	0.806	4.919	7.719	8.244	8.746	11.226
X14	2	7.846	1.238	3.574	7.022	7.884	8.689	12.413
X15	1	10.701	0.962	7.572	10.054	10.701	11.344	14.037
X16	1	7.814	1.052	3.801	7.132	7.830	8.521	11.668
X17	2	0.504	0.221	0.001	0.348	0.502	0.653	1.315
X18	3	0.682	0.204	0.004	0.544	0.686	0.819	1.390
X19	0	0.544	0.254	0.000	0.363	0.545	0.710	1.518
X20	0	0.589	0.231	0.012	0.434	0.587	0.746	1.353

This description of the predictive features shows the range of scales, ranging by and order of magnitude. Several models such as support vector machine analysis are affected by the scale and centering of the data it learns from, so it this was identified as an important preprocessing step that had to be performed.

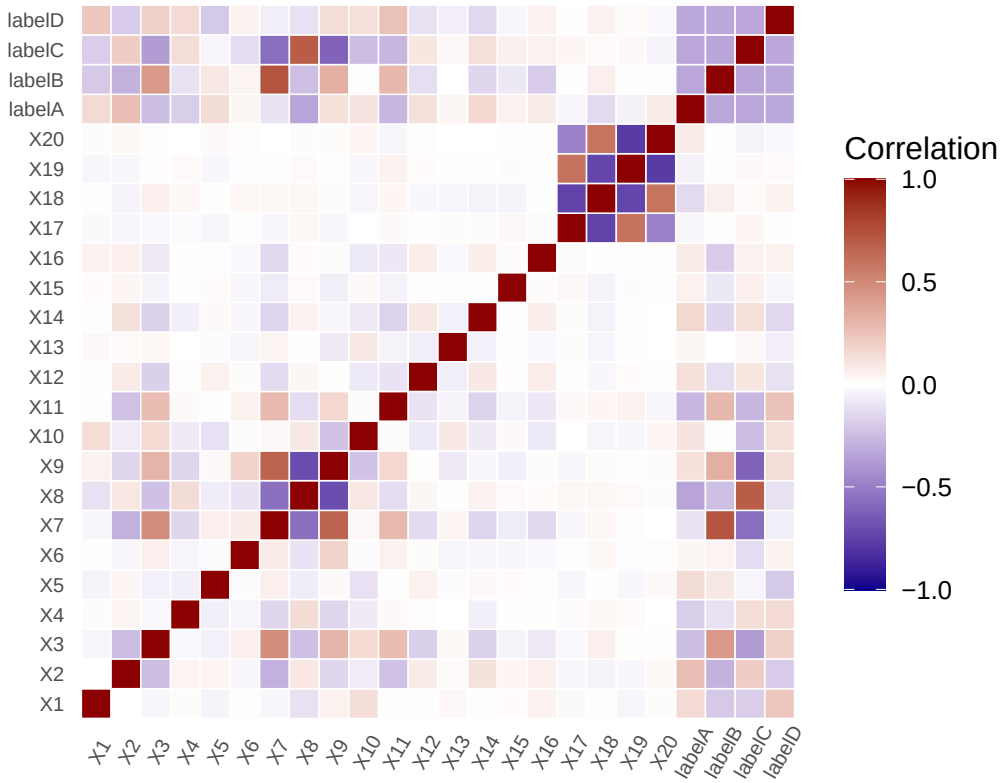

Figure 1: **Feature and Class Correlation Matrix:** *highlighting relationships between variables and relationships with catagorical labels*

Within the data there are two groups of features that correlate together - X7, X8, and X9, and X17 to X20. Noticing groups of correlated features is important since some models such as logistic regression and SVM will struggle with multicolinearity. This will lead us to attempt feature selection or dimensionality reduction with these models, or choose alternative algorithms that are more robust in these cases such as tree-based algorithms.

The difference between these two group is that while X7, X8, and X9 are three features with some of the strongest correlations with the label values, all of X17 to X20 are features without significant correlations. This would lead us to believe that X17 - X20 have low predictive power in classification that aim to predict the class label and so removing them entirely would be a justifyable approach.

Among the other columns, we see that there are definitly some columns with stronger correlations than others.

To produce the correlation matrix, the four catagorical labels were one hot encoded to create four binary columns. This allows us to see that several features have strong predictive power for one or more label but not all. For example, X8 has high correlations with classes A, B, and C, but none very low correlation with class D. This contrasts with a feature like X11 which has equal magnitude of correlation across all four labels.

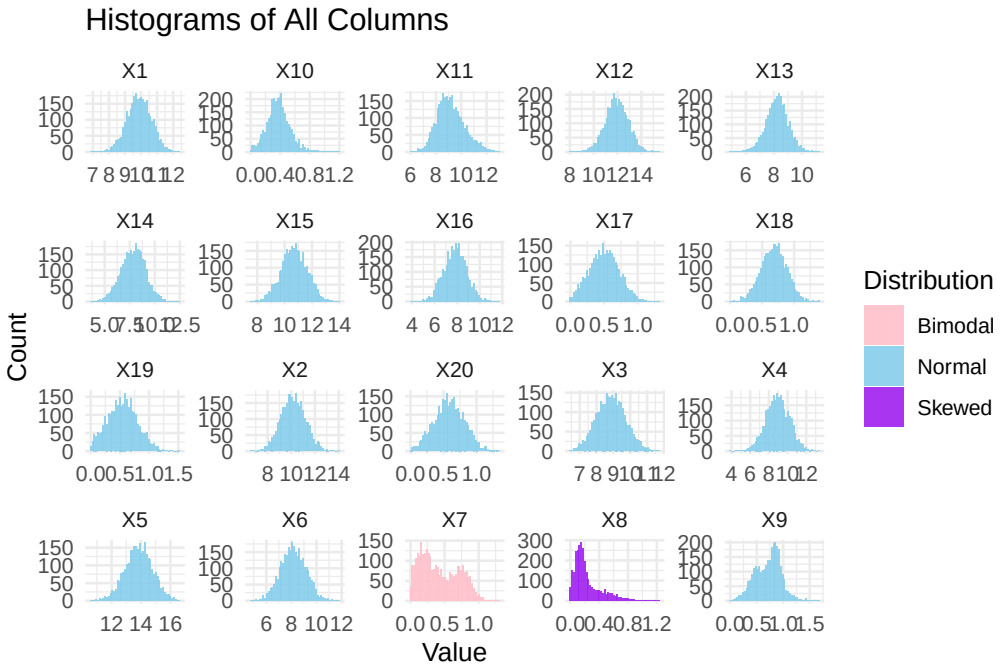

Figure 2: **Distributions within feature columns:** *histograms showing scale and skewness of data*

The majority of the twenty predictive variables appeared to follow a normal distribution. Noteable exceptions were X8, which is heavily right skewed, and X7 which has a bimodal appearance. With both of these features showing strong correlations with the labels leading to a high probability

that they have strong predictive power, they should not be removed. X8 will be transformed, and the natural logarithm of X8 will be used in all modelling.

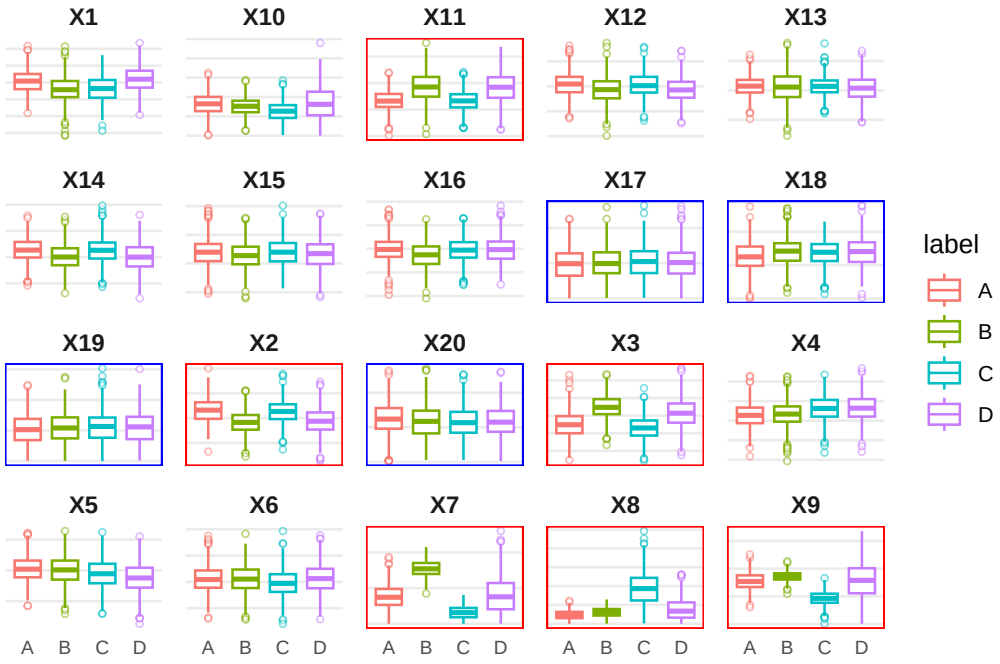

Figure 3: **Distributions within feature columns:** *Boxplots by class label by feature*

Across the entire dataset, the distribution of labels is uniform, with roughly on quarter of the observations falling into each category.

When the distribution of each feature is observed by class label, we see that some features have significantly different characteristics for each class. In the figure, highlighted with a red boarder are features where we can see notable differences in the key descriptive statistics such as medians and interquartile ranges between different classes. Highlighted in blue boarders are the features where the boxplots look almost identical from one feature to another. This gives us more insight into which features will be important for building effective models.

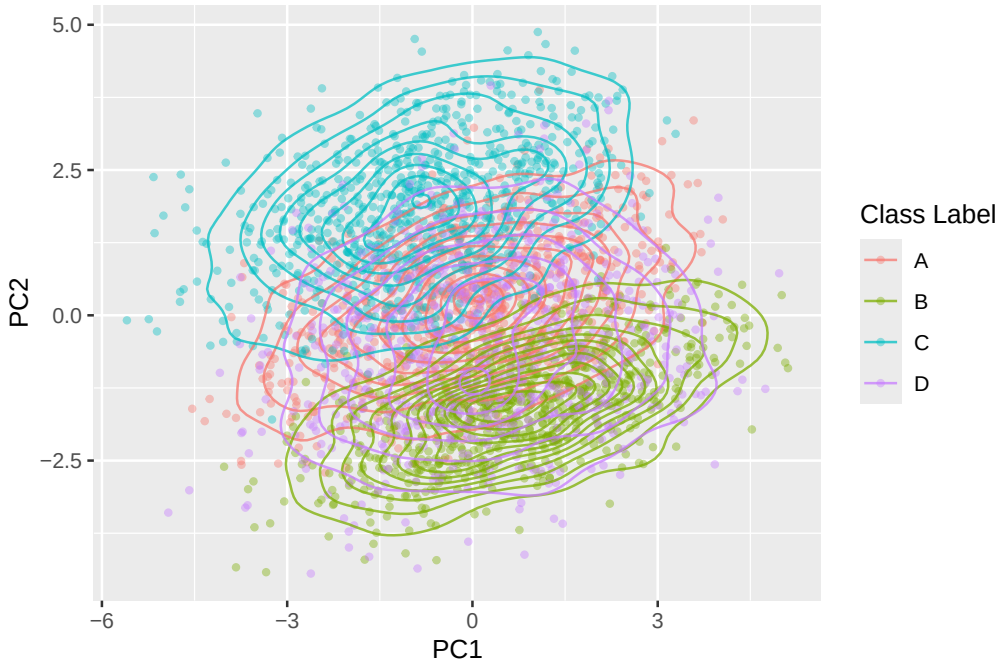

Figure 4: **PCA plotting of class labels:** *scatterplot showing clustering tendency of catagorical classes*

It is important that we learn about the structure of the underlying data in order to understand our chances of sucess with unsupervised methods where we do not have the value of the class label to train the model. Using Principle Componant Analysis, we can capture most of the information in the system in one scatter plot. This shows us that there is definitly some underlying structure to the data that will lead to the formation of clusters, although it is fairly loose in this plot. Annotating the points with the target label shows us that classes A, B, and C form elipsoidal groups that are roughly equal in size and orientation, and offset along the second principle component. The fourth class (D) forms a larger elipsoid that overlaps significantly with the other three.

The seperation of the first three classes suggests that there is information in the features that allow the statistical models to decern between them, but the overlap of class D means that this class may have more misclassifications. It may be challenging to find an approach that is effective for this label.

The fact that the three classes that are distinct are displaced along the second principle componant and not the first principle component means that it is the dimensions with less variation that distinguish them. It would be easier to seperate the groups with predictive models if they were offset along the first principle conmponent.

To learn more about the underlying structure of the dataset, the clustering tendency can be examined by calculating the Hopkins statistic, where a score close to 1 indicates strong clustering tenency, a score of 0.5 indicates random distribution of occurences, and 0 a uniform distribution.

Table 2: Hopkins Statistic Scores

Columns used	Hopkins Statistic Score
All Features + Label	0.9999829
All Features with Label Removed	0.9999211
All Features Binary Class D vs Rest	0.9999700
Features X2,X3,X7,X8,X9,X11	0.9977374
Features X17,X18,X19,X20	0.9982615
Features X1,X4,X6,X10,X12,X13,X14,X15,X16	0.9961911

We can see there is a very high clustering tendency for various treatments of the data. The statistic remains high when the label is removed from the feature set. This is promising for persuing unsupervised approaches, since it confirms that the 20 numeric features contain the information that are structuring the data into clusters, rather than the label itself providing a significant amount of this structure. If we saw the score drop when the label was removed, this would suggest that the label was necessary for dividing the data into classes and the features themselves did not contain sufficient predictive information to do so.

Similarly, we see that the score remains high for three different groups of features. These three groups are the groups we see with different coloured boarders in the boxplots. This means that even the features with very little correlation with the labels provide structure to the data. This could suggest that there are other ways that the data could be structured when using unsupervised models.

Overall, we see from exploritory data analysis that this data contains high amounts of information usable for predictive modelling, and a high clustering tendency which is promising for unsupervised clustering. Multicolinarity has been observed among several features that may prevent models from performing well.

2.3 Supervised Learning Methods

2.3.1 logistic regression - including class weighting and L2 regularization, and feature selection.

A simple model, logistic regression is quick to implement and will reveal more about the data, in particular how different features contribute to predictions.

There are variations such as weighting and regularization which will be implemented - the success of lack of success of these techniques will reveal characteristics about our data that will be valuable to inform the selection and implementation of other models

2.3.2 Random forest, including feature selection and tuning of mtry.

Random Forest was chosen due to its resilience. One of the more robust option, it is a good choice for handling the data without extra processing. Several characteristics of the data have been identified that may cause othe models such as logistic regression and SVM to struggle: - Correlated of features -Features that appeared to have low linear correlations with the label values (from heatmap in eda) but still contributed to the predictive ability of the model. This suggests there might be some non-linear relationships between features and the target variable - Features that aren't perfectly normally distributed, such as the bimodal peak in X7

Random Forest is a robust algorithm with few underlying assumptions that will handle these considerations well. Random Forest resists overfitting because of the sampling approach, it handles non linearity well, and it is naturally suited to multinomial classifications problems, like the one we have with four possible values for label. I also think that tree based models may peform well at distinguishing class A from class D, which was the biggest challenge that held back our logistic regression modelling. This is because it can prioritize at an early node in the tree a feature such as X11, which is one of the few that had high importance for descerning between class A and D, and then refine the selection in further nodes.

2.3.3 SVM, with feature selection and tuning of kernel selection, gamma and cost parameters.

SVM is a powerful and popular algorith. SVM has options for different kernals that can be tuned, and this is a promising approach to solving the challenges of seperating class A from class D that is evedent from the data analysis and the results of logistic regression. It might be the case that A and D aren't linearly seperable, but a non-linear kernal will have success.

2.4 Unsupervised Learning Methods

2.4.1 Agglomerative hierarchical clustering, including tuning of linking metric.

Agglomerative heirarchical clustering was chosen because it is interesting to explore a model where the number of clusters is not specified and let the natural structure of the data reveal itself.

Biological data is often naturally heirarchical, for example animals can be classified by deviding them into first kingdoms, then families of species, and finally species and sub-species.

Although we don't know the exact meaning of each feature in our data, we know that it is biological in origin, perhaps gene expressions or environmental factors. This means that there might be a heirarchy of classes in our data.

Using this method without specifying that there are four values for label might reveal that some of the labels have a strong tendency to form sub classes, or that there is little structure in the data to justify asserting there are four classes. These would both be interesting finds.

It is also a model that handles the feature correlation well, which allows us to keep in all the collumns that correlate like X7, X8, and X9.

2.4.2 Gaussian mixture model based clustering, including selecting a model, regularization using shrinkage parameter, and dimensionality reduction using factor analysis.

So far while working with this data set, we have struggled to seperate class D, and by plotting the results of some of the methods in specific dimensions, we have been able to show that class D significantly overlaps the other classes. We have also seen from the two dimensional PCA scatter-plot that this is general overlap between all the classes in the the first two principle componants of the data.

Lots of clustering methods struggle with seperating overlapping clusters, so for the final method I wanted to choose one that might perform better with this challenge in mind. I have chosen to try a gaussian mixture model - a model-based clustering technique that assumes that all the data is distributed according to the combination of different normal distributions. I think it might have a good chance of performing well on our dataset because it is probabalistic, calculating the probability that a data point is in each cluster. This can help it perform better than other methods like K-means when there aren't clear boundaries between the clusters such as we see with class D.

Another advantage it has is that it has some flexibility in the geometry of the clusters it produces, unlike k-means which tends to produce spherical clusters. This is important because we have seen that in some dimensions our classes produce fairly elipsoidal clusters. It also operates on very different fundemental principles to our other unsupervised method - agglomerative heirarchical clustering - so it will be good to compare the two. If model-based clustering performs much better then it could suggest that the classes of our data are not heirarchical in nature.

3 Results

3.1 Exploratory Data Analysis Findings

most features are normally distributed except for X8 which is highly skewed. Features originally had different scales. X7, X8, X9 columns are correlated and correlate highly with the labels. X17, X18, X19 and X20 are highly correlated together and have very low correlation with the labels.

The PCA showed the four labels had some clustered structure, but also some significant overlap. The hopkins statistic showed there was a moderate clustering tendency, but that the label column when included made the clustering tendency extremely high. This is an initial suggestion that supervised learning would be more effective than unsupervised learning

3.2 Supervised Learning Results

3.2.1 Logistic Regression

Table 3: Simple Logistic Regression Confusion Matrix

	Reference			
Prediction	A	B	C	D
A	120	1	2	23
B	0	130	0	16
C	0	0	144	16
D	20	7	1	73

Table 4: Simple Logistic Regression Overall Statistics

	Statistic	Value
Accuracy	Accuracy	0.8445
Kappa	Kappa	0.7921
AccuracyLower	AccuracyLower	0.8115
AccuracyUpper	AccuracyUpper	0.8737
AccuracyNull	AccuracyNull	0.2658
AccuracyPValue	AccuracyPValue	0.0000
McnemarPValue	McnemarPValue	NaN

Table 5: Simple Logistic Regression Statistics by Class

Statistic	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.8571	0.9420	0.9796	0.5703
Specificity	0.9370	0.9614	0.9606	0.9341
Pos Pred Value	0.8219	0.8904	0.9000	0.7228
Neg Pred Value	0.9509	0.9803	0.9924	0.8783
Precision	0.8219	0.8904	0.9000	0.7228
Recall	0.8571	0.9420	0.9796	0.5703
F1	0.8392	0.9155	0.9381	0.6376
Prevalence	0.2532	0.2495	0.2658	0.2315
Detection Rate	0.2170	0.2351	0.2604	0.1320
Detection Prevalence	0.2640	0.2640	0.2893	0.1826
Balanced Accuracy	0.8971	0.9517	0.9701	0.7522

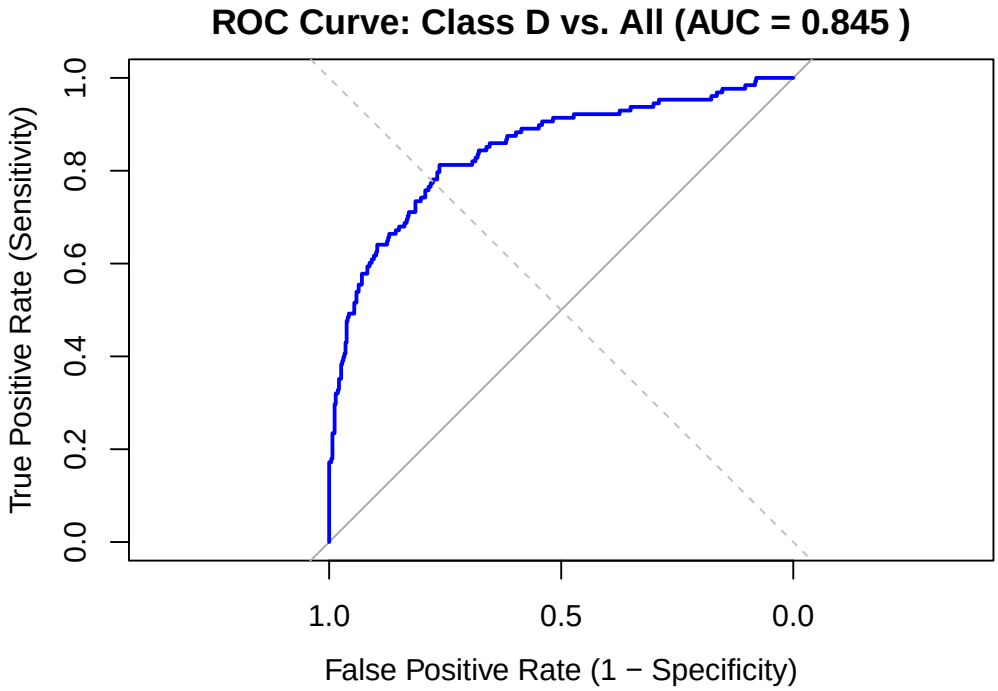

Figure 5: **Simple Logistic Regression ROC Plot:** *ROC plot for Class D vs Not Class D*

Table 6: Simple Logistic Regression with selected features Confusion Matrix

	Reference			
Prediction	A	B	C	D
A	117	1	1	26
B	0	130	0	15
C	0	0	145	15
D	23	7	1	72

Table 7: Simple Logistic Regression with selected features Overall Statistics

	Statistic	Value
Accuracy	Accuracy	0.8391
Kappa	Kappa	0.7849
AccuracyLower	AccuracyLower	0.8057
AccuracyUpper	AccuracyUpper	0.8687
AccuracyNull	AccuracyNull	0.2658
AccuracyPValue	AccuracyPValue	0.0000
McnemarPValue	McnemarPValue	NaN

Table 8: Simple Logistic Regression with selected features Statistics by Class

Statistic	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.8357	0.9420	0.9864	0.5625
Specificity	0.9322	0.9639	0.9631	0.9271
Pos Pred Value	0.8069	0.8966	0.9062	0.6990
Neg Pred Value	0.9436	0.9804	0.9949	0.8756
Precision	0.8069	0.8966	0.9062	0.6990
Recall	0.8357	0.9420	0.9864	0.5625
F1	0.8211	0.9187	0.9446	0.6234
Prevalence	0.2532	0.2495	0.2658	0.2315
Detection Rate	0.2116	0.2351	0.2622	0.1302
Detection Prevalence	0.2622	0.2622	0.2893	0.1863
Balanced Accuracy	0.8840	0.9529	0.9747	0.7448

Table 9: Simple Logistic Regression with selected features Confusion Matrix

	Reference			
Prediction	A	B	C	D
A	104	5	12	21
B	3	123	0	17
C	14	0	134	17
D	19	10	1	73

Table 10: Simple Logistic Regression with selected features Overall Statistics

	Statistic	Value
Accuracy	Accuracy	0.7848
Kappa	Kappa	0.7123
AccuracyLower	AccuracyLower	0.7482
AccuracyUpper	AccuracyUpper	0.8184
AccuracyNull	AccuracyNull	0.2658
AccuracyPValue	AccuracyPValue	0.0000
McnemarPValue	McnemarPValue	NaN

Table 11: Simple Logistic Regression with selected features Statistics by Class

Statistic	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.7429	0.8913	0.9116	0.5703
Specificity	0.9080	0.9518	0.9236	0.9294
Pos Pred Value	0.7324	0.8601	0.8121	0.7087
Neg Pred Value	0.9124	0.9634	0.9665	0.8778
Precision	0.7324	0.8601	0.8121	0.7087
Recall	0.7429	0.8913	0.9116	0.5703
F1	0.7376	0.8754	0.8590	0.6320

Prevalence	0.2532	0.2495	0.2658	0.2315
Detection Rate	0.1881	0.2224	0.2423	0.1320
Detection Prevalence	0.2568	0.2586	0.2984	0.1863
Balanced Accuracy	0.8254	0.9216	0.9176	0.7499

Table 12: Weighted Logistic Regression Confusion Matrix

	Reference			
Prediction	A	B	C	D
A	118	1	2	23
B	0	130	0	16
C	0	0	144	16
D	22	7	1	73

Table 13: Weighted Logistic Regression Overall Statistics

	Statistic	Value
Accuracy	Accuracy	0.8445
Kappa	Kappa	0.7921
AccuracyLower	AccuracyLower	0.8115
AccuracyUpper	AccuracyUpper	0.8737
AccuracyNull	AccuracyNull	0.2658
AccuracyPValue	AccuracyPValue	0.0000
McnemarPValue	McnemarPValue	NaN

Table 14: Weighted Logistic Regression Statistics by Class

Statistic	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.8429	0.9420	0.9796	0.5703
Specificity	0.9370	0.9614	0.9606	0.9294
Pos Pred Value	0.8194	0.8904	0.9000	0.7087
Neg Pred Value	0.9462	0.9803	0.9924	0.8778
Precision	0.8194	0.8904	0.9000	0.7087
Recall	0.8429	0.9420	0.9796	0.5703
F1	0.8310	0.9155	0.9381	0.6320
Prevalence	0.2532	0.2495	0.2658	0.2315
Detection Rate	0.2134	0.2351	0.2604	0.1320
Detection Prevalence	0.2604	0.2640	0.2893	0.1863
Balanced Accuracy	0.8900	0.9517	0.9701	0.7499

Table 15: Logistic Regression Performance with Different Regularization Parameters

Decay	Accuracy	Kappa	F1 Score (A)	F1 Score (B)	F1 Score (C)	F1 Score (D)

0.001	0.8445	0.7921	0.8392	0.9155	0.9381	0.6376
0.01	0.8445	0.7921	0.8392	0.9155	0.9381	0.6376
0.1	0.8445	0.7921	0.8392	0.9155	0.9381	0.6376
0.5	0.8427	0.7897	0.8351	0.9123	0.9381	0.6376
1	0.8391	0.7849	0.8322	0.9053	0.9381	0.6316
2	0.8391	0.7849	0.8293	0.9053	0.9412	0.6316
10	0.8336	0.7776	0.8315	0.9010	0.9320	0.6133

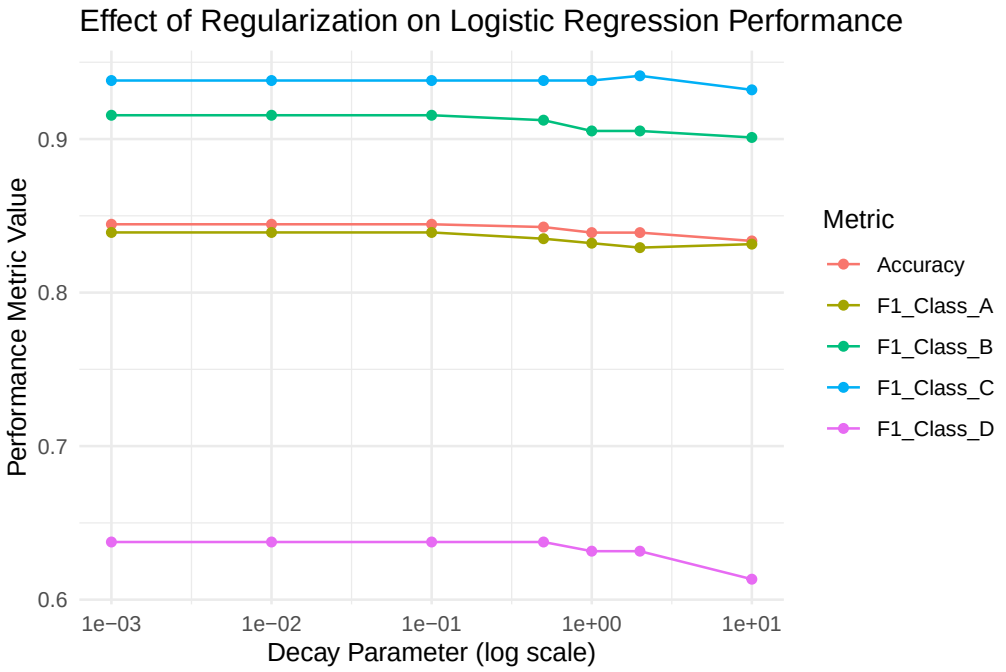

Figure 6: **Regularized Logistic Regression**

Table 16: Confusion Matrix for Best Regularized Logistic Regression Model (Decay = 0.001)

	Reference			
Prediction	A	B	C	D
A	120	1	2	23
B	0	130	0	16
C	0	0	144	16
D	20	7	1	73

logistic regression was not great. weighting did nothing, as expected. regularization didn't really do anything. feature selection did not improve the model. We saw that the model particularly underperformed at classifying class D correctly.

3.2.2 Random Forest

Table 17: Basic Random Forest Confusion Matrix

	Reference			
Prediction	A	B	C	D
A	130	0	1	20
B	0	135	0	15
C	0	0	145	8
D	10	3	1	85

Table 18: Basic Random Forest Overall Statistics

	Statistic	Value
Accuracy	Accuracy	0.8951
Kappa	Kappa	0.8598
AccuracyLower	AccuracyLower	0.8665
AccuracyUpper	AccuracyUpper	0.9194
AccuracyNull	AccuracyNull	0.2658
AccuracyPValue	AccuracyPValue	0.0000
McnemarPValue	McnemarPValue	NaN

Table 19: Basic Random Forest Statistics by Class

Statistic	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.9286	0.9783	0.9864	0.6641
Specificity	0.9492	0.9639	0.9803	0.9671
Pos Pred Value	0.8609	0.9000	0.9477	0.8586
Neg Pred Value	0.9751	0.9926	0.9950	0.9053
Precision	0.8609	0.9000	0.9477	0.8586
Recall	0.9286	0.9783	0.9864	0.6641
F1	0.8935	0.9375	0.9667	0.7489
Prevalence	0.2532	0.2495	0.2658	0.2315
Detection Rate	0.2351	0.2441	0.2622	0.1537
Detection Prevalence	0.2731	0.2712	0.2767	0.1790
Balanced Accuracy	0.9389	0.9711	0.9833	0.8156

Table 20: Basic Random Forest Feature Importance

	A	B	C	D	MeanDecreaseAccuracy	MeanDecreaseGini
X7	50.178	123.195	48.798	8.689	94.272	360.168
X8	70.546	24.694	59.057	-9.548	69.003	249.545
X10	49.286	22.251	58.352	20.140	65.924	137.078
X9	36.026	30.389	57.324	12.945	59.435	243.792
X11	41.100	10.247	40.549	26.969	54.407	109.409
X3	27.326	28.103	32.068	11.605	44.401	109.383
X2	21.802	15.885	12.888	9.189	27.226	64.198

X1	5.962	14.128	7.705	14.315	21.040	52.068
X14	13.434	0.341	7.159	6.457	14.516	39.462
X18	11.272	5.765	5.430	-0.569	12.927	26.816
X13	8.198	-0.523	10.302	5.309	12.894	29.595
X4	13.610	3.251	-1.849	7.126	12.437	39.760
X5	10.726	-0.609	-0.043	7.295	10.755	34.441
X12	7.179	0.349	7.800	5.186	10.384	30.063
X19	3.480	4.531	2.829	-1.073	5.023	20.314
X16	2.354	2.495	-0.401	4.082	4.629	25.166
X17	4.903	2.814	1.571	-1.040	4.548	21.503
X20	6.057	2.140	1.583	-1.709	4.516	22.280
X6	1.262	-0.676	2.216	1.539	2.255	24.902
X15	0.444	3.233	0.437	0.290	2.025	24.404

Table 21: Random Forest (5 least important features removed) Confusion Matrix

	Reference			
Prediction	A	B	C	D
A	129	1	0	20
B	0	134	0	13
C	0	0	145	6
D	11	3	2	89

Table 22: Random Forest (5 least important features removed) Overall Statistics

	Statistic	Value
Accuracy	Accuracy	0.8987
Kappa	Kappa	0.8647
AccuracyLower	AccuracyLower	0.8705
AccuracyUpper	AccuracyUpper	0.9226
AccuracyNull	AccuracyNull	0.2658
AccuracyPValue	AccuracyPValue	0.0000
McnemarPValue	McnemarPValue	NaN

Table 23: Random Forest (5 least important features removed) Statistics by Class

Statistic	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.9214	0.9710	0.9864	0.6953
Specificity	0.9492	0.9687	0.9852	0.9624
Pos Pred Value	0.8600	0.9116	0.9603	0.8476
Neg Pred Value	0.9727	0.9901	0.9950	0.9129
Precision	0.8600	0.9116	0.9603	0.8476
Recall	0.9214	0.9710	0.9864	0.6953
F1	0.8897	0.9404	0.9732	0.7639
Prevalence	0.2532	0.2495	0.2658	0.2315
Detection Rate	0.2333	0.2423	0.2622	0.1609

Detection Prevalence	0.2712	0.2658	0.2731	0.1899
Balanced Accuracy	0.9353	0.9698	0.9858	0.8288

Table 24: Random Forest (5 least important features removed) Feature Importance

	A	B	C	D	MeanDecreaseAccuracy	MeanDecreaseGini
X7	58.778	170.331	58.103	12.361	115.060	393.832
X8	79.792	25.300	76.407	-9.359	87.035	261.263
X10	58.892	25.012	68.278	21.863	81.340	156.837
X9	43.211	33.091	75.760	16.200	74.779	269.860
X11	44.359	10.145	46.135	29.869	60.058	119.295
X3	30.197	28.932	32.616	14.103	48.321	114.500
X2	21.424	16.846	12.326	9.676	28.820	63.799
X1	5.386	14.134	9.653	14.498	21.435	50.625
X14	13.600	4.097	6.321	9.991	18.016	39.849
X4	14.789	4.920	-1.521	11.160	17.127	41.036
X5	12.953	2.051	-2.077	8.866	12.988	35.297
X12	7.641	-0.037	8.708	6.206	11.506	32.122
X13	5.177	-0.472	9.949	6.446	11.252	31.485
X18	6.163	-1.981	2.941	-0.384	3.670	27.229
X16	2.665	0.627	-0.780	3.256	3.275	27.314

Table 25: Random Forest (5 least important features removed) Confusion Matrix

	Reference			
Prediction	A	B	C	D
A	123	3	2	22
B	3	133	0	12
C	4	0	143	12
D	10	2	2	82

Table 26: Random Forest (5 least important features removed) Overall Statistics

	Statistic	Value
Accuracy	Accuracy	0.8698
Kappa	Kappa	0.8259
AccuracyLower	AccuracyLower	0.8389
AccuracyUpper	AccuracyUpper	0.8967
AccuracyNull	AccuracyNull	0.2658
AccuracyPValue	AccuracyPValue	0.0000
McnemarPValue	McnemarPValue	NaN

Table 27: Random Forest (5 least important features removed) Statistics by Class

Statistic	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.8786	0.9638	0.9728	0.6406
Specificity	0.9346	0.9639	0.9606	0.9671
Pos Pred Value	0.8200	0.8986	0.8994	0.8542
Neg Pred Value	0.9578	0.9877	0.9898	0.8993
Precision	0.8200	0.8986	0.8994	0.8542
Recall	0.8786	0.9638	0.9728	0.6406
F1	0.8483	0.9301	0.9346	0.7321
Prevalence	0.2532	0.2495	0.2658	0.2315
Detection Rate	0.2224	0.2405	0.2586	0.1483
Detection Prevalence	0.2712	0.2676	0.2875	0.1736
Balanced Accuracy	0.9066	0.9638	0.9667	0.8038

Table 28: Random Forest (5 least important features removed) Feature Importance

	A	B	C	D	MeanDecreaseAccuracy	MeanDecreaseGini
X7	46.928	131.294	62.714	11.698	106.991	394.017
X9	45.490	33.844	117.233	20.621	88.116	329.713
X10	48.203	24.909	58.850	21.496	66.109	148.005
X11	42.252	13.808	42.799	27.582	52.903	116.436
X3	27.622	36.977	36.642	15.727	52.104	125.651
X2	21.159	18.514	12.862	11.968	31.020	73.567
X1	7.424	16.590	11.568	14.707	23.943	62.662
X14	13.939	2.007	7.069	9.486	16.032	46.087
X4	12.400	5.259	3.249	10.335	15.806	48.080
X13	6.283	1.648	9.093	6.989	12.536	36.913
X18	9.273	5.833	7.010	3.238	12.423	32.457
X5	10.625	1.883	-1.438	9.410	11.117	39.770
X12	5.089	0.856	8.879	6.496	10.821	38.286
X20	6.161	2.901	5.964	-0.197	7.679	28.957
X6	4.089	-0.262	5.483	3.494	6.711	30.751
X17	3.116	5.812	4.081	-0.494	6.341	27.301
X19	3.203	3.596	3.133	1.176	5.410	26.548
X15	1.684	3.629	3.830	-1.466	3.806	28.961
X16	2.068	1.645	-0.925	3.210	3.320	30.286

```
-0.05263158 0.01
0.03508772 0.01
0.01818182 0.01
-0.02777778 0.01
```

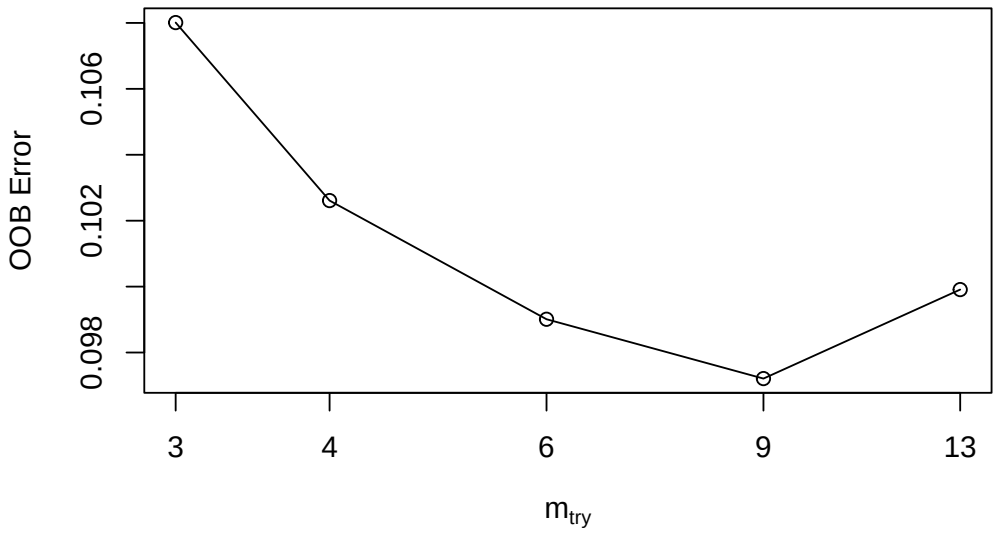

Figure 7: **Optimal mtry for Random Forest**

Table 29: Tuned Random Forest Confusion Matrix

	Reference			
Prediction	A	B	C	D
A	126	2	1	17
B	2	132	0	11
C	0	0	144	4
D	12	4	2	96

Table 30: Tuned Random Forest Overall Statistics

	Statistic	Value
Accuracy	Accuracy	0.8951
Kappa	Kappa	0.8598
AccuracyLower	AccuracyLower	0.8665
AccuracyUpper	AccuracyUpper	0.9194
AccuracyNull	AccuracyNull	0.2658
AccuracyPValue	AccuracyPValue	0.0000
McnemarPValue	McnemarPValue	NaN

Table 31: Tuned Random Forest Statistics by Class

Statistic	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.9000	0.9565	0.9796	0.7500
Specificity	0.9516	0.9687	0.9901	0.9576
Pos Pred Value	0.8630	0.9103	0.9730	0.8421
Neg Pred Value	0.9656	0.9853	0.9926	0.9271

Precision	0.8630	0.9103	0.9730	0.8421
Recall	0.9000	0.9565	0.9796	0.7500
F1	0.8811	0.9329	0.9763	0.7934
Prevalence	0.2532	0.2495	0.2658	0.2315
Detection Rate	0.2278	0.2387	0.2604	0.1736
Detection Prevalence	0.2640	0.2622	0.2676	0.2061
Balanced Accuracy	0.9258	0.9626	0.9849	0.8538

The random forest performed well without any configuration. feature selection was not effective. still struggled at seperating class D. mtry was tuned.

3.2.3 SVM

svm was good but not as good as random forest. lots of tuning. feature selection was uneffective

3.3 Unsupervised Learning Results

3.3.1 Agglomerative Hierarchical Clustering

ahc was good not great.

3.3.2 Gaussian Mixed Model Clustering

Gaussian mixed model clustering performed very well. dimensionality reduction using factor analysis was slightly effective.

4 Discussion

the final model had good overall accuracy but caution is advised when using a ml model with this data due to the poor performance in class D - if false positives or false negatives in this class have serious implications, some models become immediately unusable.

5 Conclusion

6 References