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

List of Figures

List of Tables

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

```
# A tibble: 20 x 9
   `Variable Name` `No. missing values`   mean `Std deviation`      min
 * <chr>                           <int>  <dbl>           <dbl>    <dbl>
 1 X1                                  2  9.88           0.764  6.84
 2 X2                                  0 10.2            1.04   6.54
 3 X3                                  1  8.86           0.871  6.42
 4 X4                                  2  8.94           1.28   3.87
 5 X5                                  0 13.9            0.942 10.5
 6 X6                                  0  8.15           1.03   4.81
 7 X7                                  1  0.426          0.278  0.000171
 8 X8                                  3  0.234          0.197  0.000399
 9 X9                                  0  0.717          0.247  0.00608
10 X10                                 0  0.378          0.155  0.000368
11 X11                                 1  9.18           1.09   6.03
12 X12                                 0 11.9            0.977  8.05
13 X13                                 1  8.23           0.806  4.92
14 X14                                 2  7.85           1.24   3.57
15 X15                                 1 10.7
16 X16                                 1  7.81
17 X17                                 2  0.504
18 X18                                 3  0.682
19 X19                                 0  0.544
20 X20                                 0  0.589
# i 4 more variables: `25th %ile` <dbl>, median <dbl>,
#   max <dbl>
```

```
\begin{table*}[ht]
\centering
\begingroup\fontsize{9}{11}\selectfont

\begin{longtable}[t]{cccccccccccccccc}
\caption{\label{tab:feature-data-descriptions}Feature Descriptions}\\
\toprule
skim\_type & skim\_variable & n\_missing & complete\_rate & character.min & character.max & chara
\midrule
\cellcolor{gray!10}{character} & \cellcolor{gray!10}{label} & \cellcolor{gray!10}{0} & \cellcolor
numeric & X1 & 2 & 0.999 & NA & NA & NA & NA & NA & 9.876 & 0.764 & 6.840 & 9.356 & 9.872 & 10.39
\cellcolor{gray!10}{numeric} & \cellcolor{gray!10}{X2} & \cellcolor{gray!10}{0} & \cellcolor{gray
numeric & X3 & 1 & 1.000 & NA & NA & NA & NA & NA & 8.861 & 0.871 & 6.424 & 8.243 & 8.847 & 9.445
\cellcolor{gray!10}{numeric} & \cellcolor{gray!10}{X4} & \cellcolor{gray!10}{2} & \cellcolor{gray
\addlinespace
numeric & X5 & 0 & 1.000 & NA & NA & NA & NA & NA & 13.853 & 0.942 & 10.527 & 13.236 & 13.858 & 1
\cellcolor{gray!10}{numeric} & \cellcolor{gray!10}{X6} & \cellcolor{gray!10}{0} & \cellcolor{gray
numeric & X7 & 1 & 1.000 & NA & NA & NA & NA & NA & 0.426 & 0.278 & 0.000 & 0.185 & 0.375 & 0.678
\cellcolor{gray!10}{numeric} & \cellcolor{gray!10}{X8} & \cellcolor{gray!10}{3} & \cellcolor{gray
numeric & X9 & 0 & 1.000 & NA & NA & NA & NA & NA & 0.717 & 0.247 & 0.006 & 0.532 & 0.751 & 0.888
\addlinespace
\cellcolor{gray!10}{numeric} & \cellcolor{gray!10}{X10} & \cellcolor{gray!10}{0} & \cellcolor{gra
numeric & X11 & 1 & 1.000 & NA & NA & NA & NA & NA & 9.175 & 1.087 & 6.031 & 8.412 & 9.177 & 9.86
\cellcolor{gray!10}{numeric} & \cellcolor{gray!10}{X12} & \cellcolor{gray!10}{0} & \cellcolor{gra
numeric & X13 & 1 & 1.000 & NA & NA & NA & NA & NA & 8.228 & 0.806 & 4.919 & 7.719 & 8.244 & 8.74
\cellcolor{gray!10}{numeric} & \cellcolor{gray!10}{X14} & \cellcolor{gray!10}{2} & \cellcolor{gra
\addlinespace
numeric & X15 & 1 & 1.000 & NA & NA & NA & NA & NA
\cellcolor{gray!10}{numeric} & \cellcolor{gray!10}-
numeric & X17 & 2 & 0.999 & NA & NA & NA & NA & NA
\cellcolor{gray!10}{numeric} & \cellcolor{gray!10}-
numeric & X19 & 0 & 1.000 & NA & NA & NA & NA & NA
\addlinespace
\cellcolor{gray!10}{numeric} & \cellcolor{gray!10}-
\bottomrule
\end{longtable}
\endgroup{}
\end{table*}
```

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.

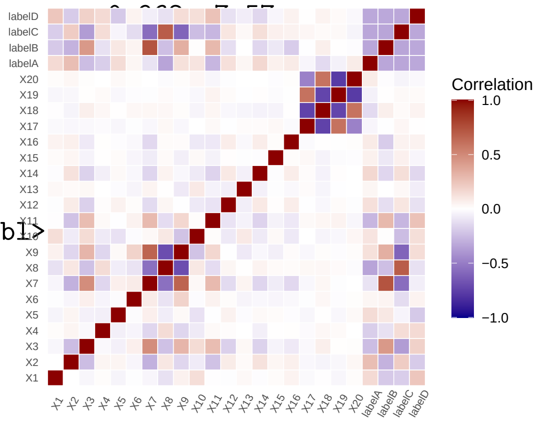


Figure 1: **Feature and Class Correlation Matrix:** *highlighting relationships between variables and relationships with catagorical labels*

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 justifiable 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.

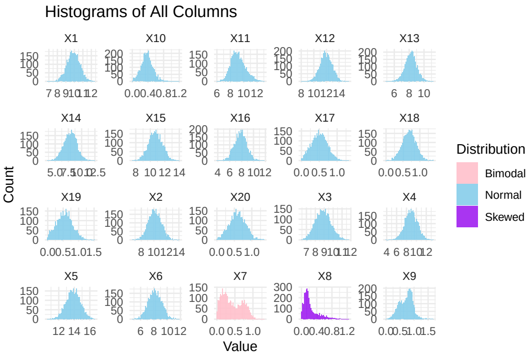


Figure 2: **Distributions within feature columns:** *histograms showing scale and skewness of data*

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.

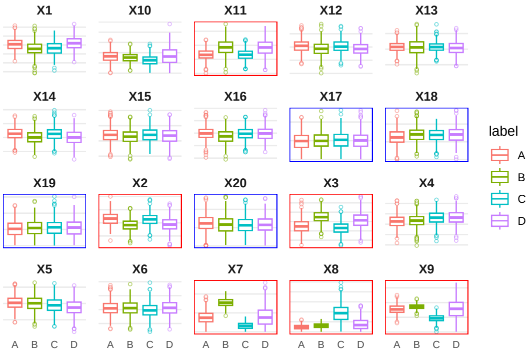


Figure 3: **Distributions within feature columns:** *Boxplots by class label by feature*

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.

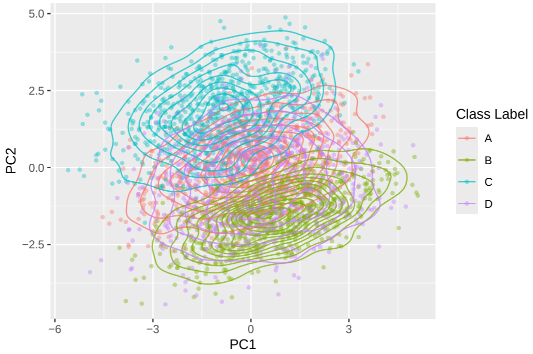


Figure 4: **PCA plotting of class labels:** *scatterplot showing clustering tendency of catagorical classes*

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.

```
\begin{table*}[ht]
\centering
\end{table*}
```

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 nonlinear 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 scatterplot 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

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

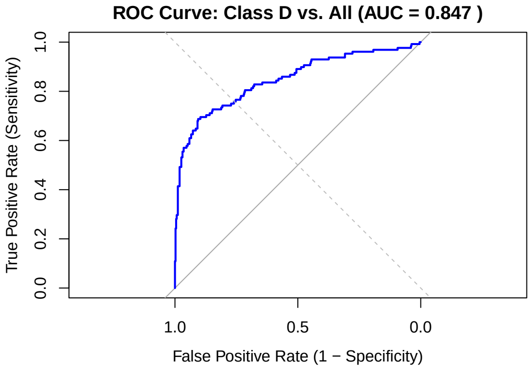


Figure 5: **Simple Logistic Regression ROC Plot:** *ROC plot for Class D vs Not Class D*

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

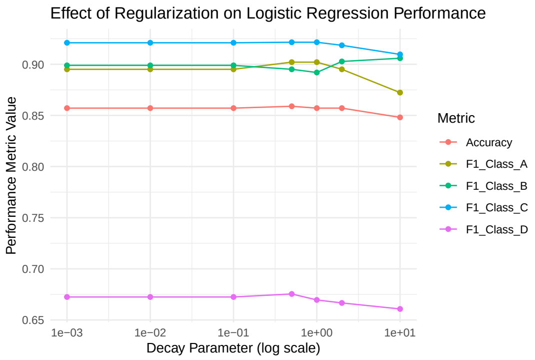


Figure 6: **Regularized Logistic Regression**

```
\begin{table*}[ht]
\centering
\end{table*}
```

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

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}

\begin{table*}[ht]
\centering
\end{table*}

\begin{table*}[ht]
\centering
\end{table*}

\begin{table*}[ht]
\centering
\end{table*}

\begin{table*}[ht]
\centering
\end{table*}

\begin{table*}[ht]
\centering
\end{table*}
```

```
-0.04366812 0.01
-0.004366812 0.01
```

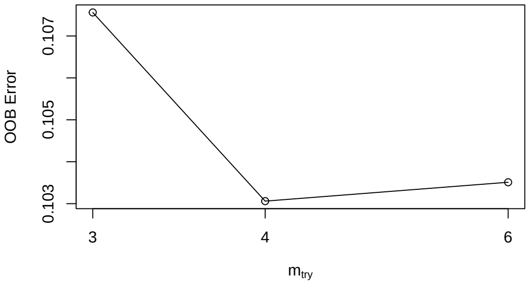


Figure 7: **Optimal mtry for Random Forest**

```
\begin{table*}[ht]
\centering
\end{table*}

\begin{table*}[ht]
\centering
\end{table*}

\begin{table*}[ht]
\centering
\end{table*}
```

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

```
\begin{table*}[ht]
\centering
\end{table*}
```

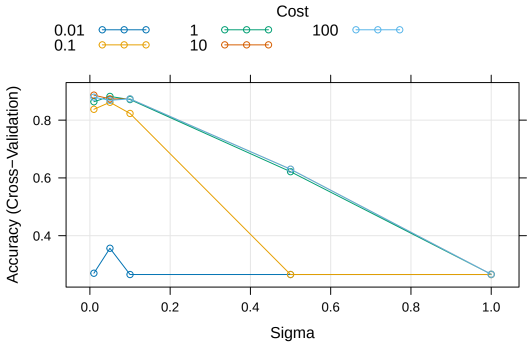


Figure 8: **Hyperparameter Affect on SVM Performance**

```
    sigma      C
301 0.052 0.1975
```

```
\begin{table*}[ht]
\centering
\end{table*}

\begin{table*}[ht]
\centering
\end{table*}

\begin{table*}[ht]
\centering
\end{table*}
```

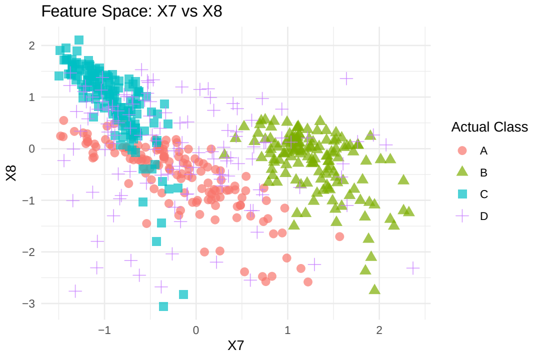


Figure 9: **Misclassifications - X7 by X8**

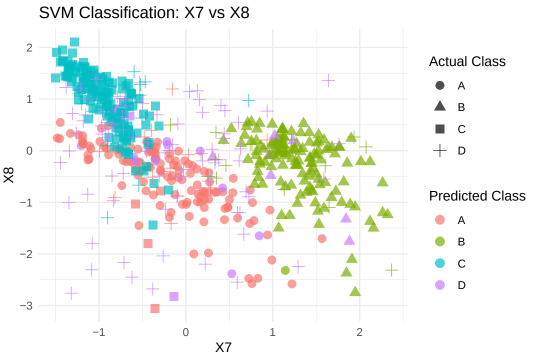


Figure 10: **Misclassifications - X7 by X8**

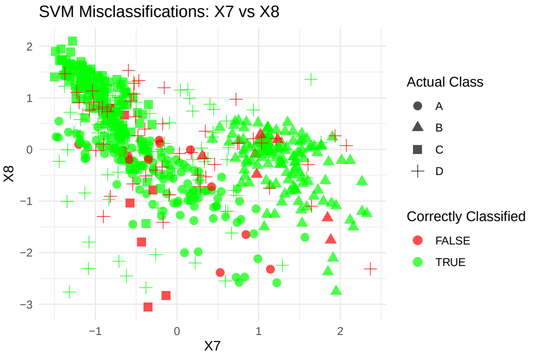


Figure 11: **Misclassifications - X7 by X8**

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

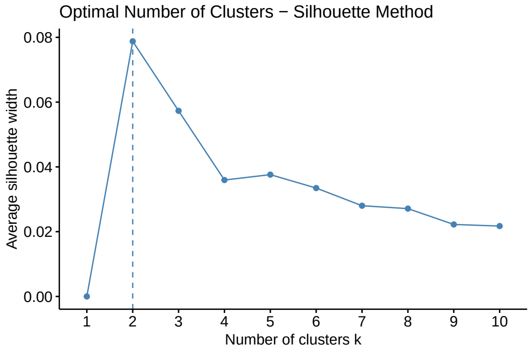


Figure 12: **Silhouette Plot**

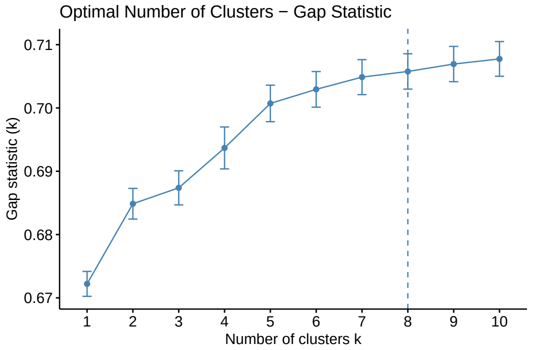


Figure 13: **Gap Statistic Plot**

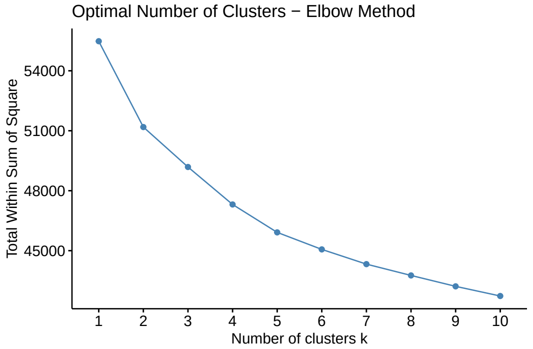


Figure 14: **Elbow Graph**

ahc was good not great.

3.3.2 Gaussian Mixed Model Clustering

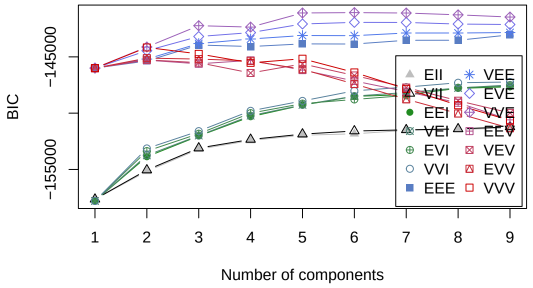
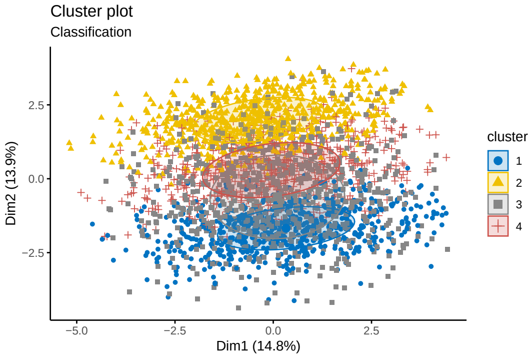


Figure 15: **Baysian Information Criteria Plot for Choosing Optimal Model**

VVE was found to be the best performing model. Promisingly, the plot peaked at four clusters. Since it is known that there actually are four categorical classes in the original data, this shows that the model is learning from the distributions of all four labels. Since we saw such overlap in classes and many misclassifications in class D previously, it would have been unsuprising to see three as the optimum number of classes,suggesting that data was better described by three categories than four.

In heirarchical clustering, we didn't see clear confirmation of four groups from the graphs using either the elbow method, gap method, or silloette method.

The three letters in the model name describe the shape, orientation, and orientation of the clusters that the model predicts. In this case, VVE indicates that the model is predicting clusters that are elipsoids of equal orientation, but varying volume. This makes sense based on the PCA scatterplots where the data is shown as three roughly equally size elipses one above the other, with a fourth, larger elipses overlaying them, with the major axis of all four being roughly horizontal (along the first principle component)



This looks really similar to our original PCA plot, with class A B and C seperated and class D overlapping all three.

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

Gaussian mixed model clustering performed very well. It immediately had overall accuracy comparible with random forest, and performed notebly better than other models at seperating the fourth class successfully, with an F1 score of 0.8.

Attempts were made to further tune the model through regularization:

```
\begin{table*}[ht]
\centering
\end{table*}
```

The best performing model did not use regularization.

The motivation for using regularization is to help the model perform better when using features that correlate together. Since regularization in fact hindered performance, another approach is to use factor analysis to reduce the dimensionality of the data. Factor analysis is suitable because it is suited to handling the groups of correlated features notable in the EDA results, and because the data is biological in origin. With biological data, there is often an underlying cause, like a gene expression, that can have many measurable implecations, like disease symptoms or physical characteristics. Because we know these mechanisms may exist in the source of our data, factor analysis is a good choice to reduce dimensions while preserving as much information as possible.

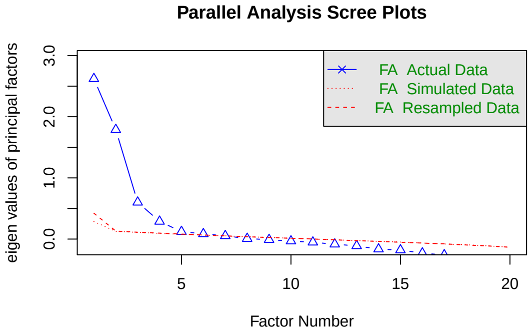


Figure 16: **Scree Plot:** *for Choosing Number of Factors for Dimensionality Reduction*

The parallel analysis results suggest that the optimum number of factors is 6. Using the maximum likelyhood method finds the number of factors where the value of the eigenvalue is above what would be expected by random chance.

The underlying values from the analysis show that the first two factors contribute the most, with a sharp drop after that. So a dimensionality reduction to two factors could be a reasonable option. We can also see that the seventh and eigth eigenvalues are not much smaller than the sixth, so swapping some of the smaller factors could also be a justifyable experiment.

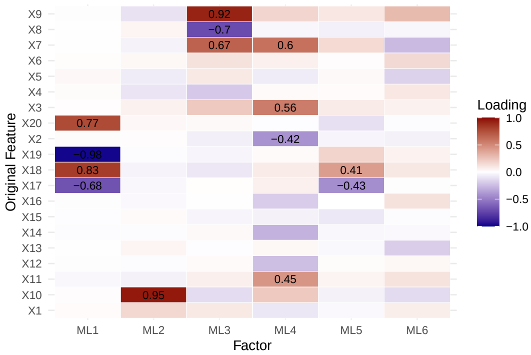


Figure 17: **Factor Analysis Loadings:** *loadings values by feature*

The loadings of features to factors shows the highly correlated features in the correlated groups are all heavily loaded to the same factor, which will succesfully decrease the multicolinearily present in the dataset.

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

```
\begin{table*}[ht]
\centering
\end{table*}
```

Operating on the factors rather than the original features, overall accuracy is slightly higher and the balanced accuracy of cluster 3 (class D) is slightly lower. This is a trade off that requires more knowledge of the intended use of the model to make a choice between the two.

However, the fact that factor analysis leads to increased accuracy overall while decreasing the dimensionality of the data so far is an interesting finding.

3.3.2.1 Model Evaluation

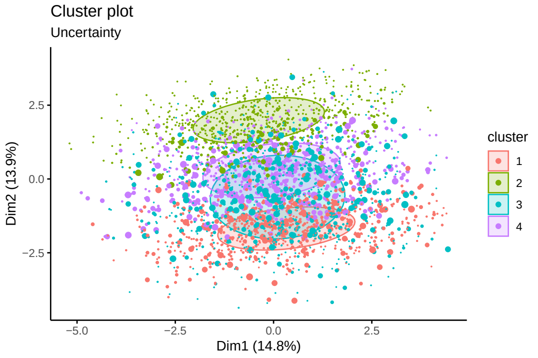


Figure 18: **Cluster Uncertainty Plot:** *Clusters shown with uncertain points by size*

```
\begin{table*}[ht]
\centering
\end{table*}
```

4 Discussion

The relative performance of different learning models can give some insight to the underlying structure of the data. If random forest and agglomerative hierarchical clustering performed significantly better than the other models, this would be evidence that the data was heirarchical in nature. If logistic regression performed as well as any of the other models, it would indicate that the data was extremely structured and easy to model, with strong linear relationships between the predictive features and the labels. The results found that random forest performed best followed by the radial kernel SVM. Both of these algorithms are better at modelling with non-linear relationships. This suggests that the the relationship between the features and the label exhibits some non-linearity.

There are also lessons learned from the challenge of seperating class D which arrose in every model that was implemented. Class D was harder for every model to correctly predict. In the context of biological data there are a several explainations of why this would happen, including that D is a super-class in a hierarchy where the other three are sub classes. This appears unlikely due to the underperformance of agglomerative hierarchical clustering. Another explaination is that D could be a transition state between other classes. This is not supported by the visualization of the data where class D seems to form a differently shaped sigmoid to the other classes, and overlap all three. Its still possible that D represents an immature state that will later develop into one of the other three. The uncertainy of the predictions of this class by an array of models with such different underlying principles suggests that the difficulty is not a limitation of any algorithm, and it is likely that the features of the data do not have the predictive power necessary for efficient seperation of class D. Better performing models could be developed if data was collected with additional variables.

At the outset of the project, it was reasonable to assume that the supervised learning would outperform the unsupervised learning models. The gaussian mixture model achieving the highest performance metrics speaks to the high amounts of noise present within each cluster and the dataset as a whole, as performing well on these kinds of data is a hallmark of the gmm algorithm.

Factor analysis had some success as the gmm model was able to retain its high accuracy and increase the F1 score of class D with a considerable reduction in the dimensionality of the data, suggesting the existance of underlying biological mechanisms or environmental factors that lead to the observed features arrising. Although well performing statistical models were trained using the features and the factor analysis approach, better results mights be possible using the underlying factors themselves if it is possible to measure them.

The best models had good overall accuracy but caution is advised for implementing any model with this data due to the underperformance in class D - if false positives or false negatives in this class have serious implications, some models become immediately unusable. Examples were this would be the case include when the classes represent risk of side effects to a medication option.

5 Conclusion

The results of this project demonstrate how advanced statistical techniques are relevent to fields of research using biological data, with the potential for powerful models evident. The gaussian mixed model would be the one model that could be taken forward for use classifying new data collected or for generalizing to another data set because of its high overall accuracy but particularly because of its high balanced accuracy in class D compared to other models. The greatest limitation of the models trained in this project is the underperformance of classifying class D, which could be critical in some applications. The collection of data with more descriptive features could allow for the development of more powerful models using the same techniques, or the work in this project could be build upon to create more refined models that perform better for specific use cases. Further research should focus on improving the results of the random forest and gaussian mixture model approaches. An algorithm such as A gradient boosting algorithm such as XGBoost would build upon the successful tree-based approach of random forest. With each tree in a gradient boosting approach able to learn from the one before and the high tunability of the model, a model that performs better at the shortcomings of models in this report is possible. An exciting direction to take forward the success of gmm would be to experiment with semi-supervised learning. A sample of class labels supplied to a model based clustering algorithm to guide the initial assignment of cluster can lead to learning distributions that are more specific to each cluster, with better results.

6 References