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 is a section where i write the introduction.

2 Methods

2.1 Data Description

20 features, a label with four catagorical classes. two groups of correlated features. Outliers removed using z score method. one feature transformed using logarithm. All features scaled and centered. All features numeric.

Bootstrapping was used to create a larger dataset.

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

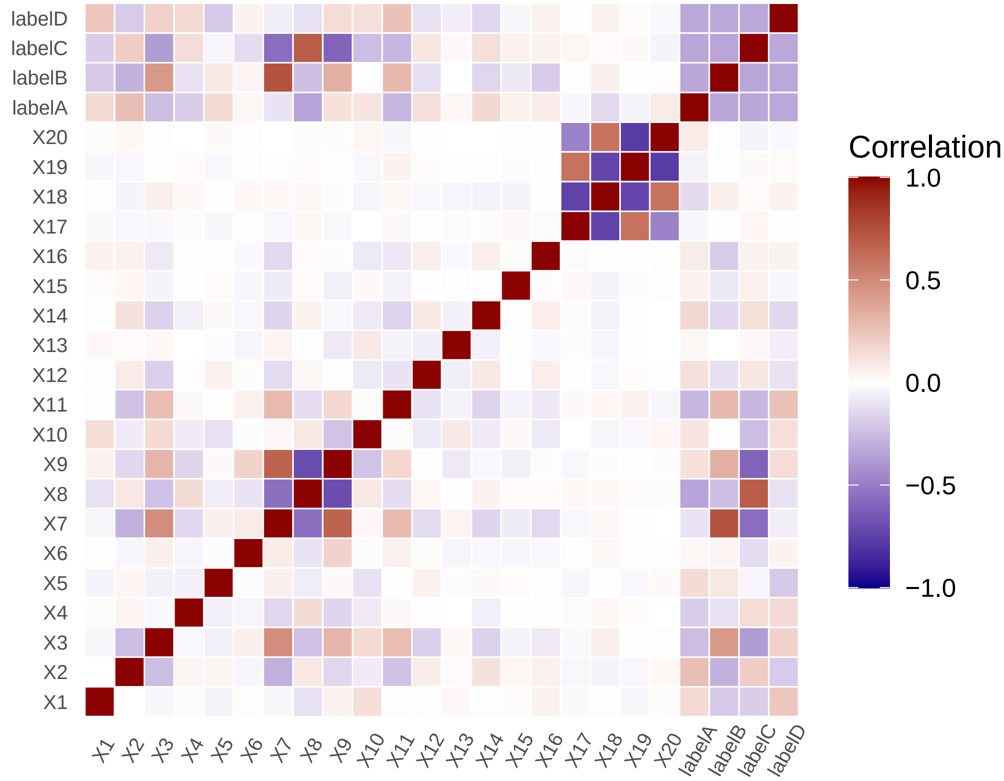

Figure 1: **Feature and Class Correlation Matrix:** *highlighting relationships between variables and relationships with catagorical labels*

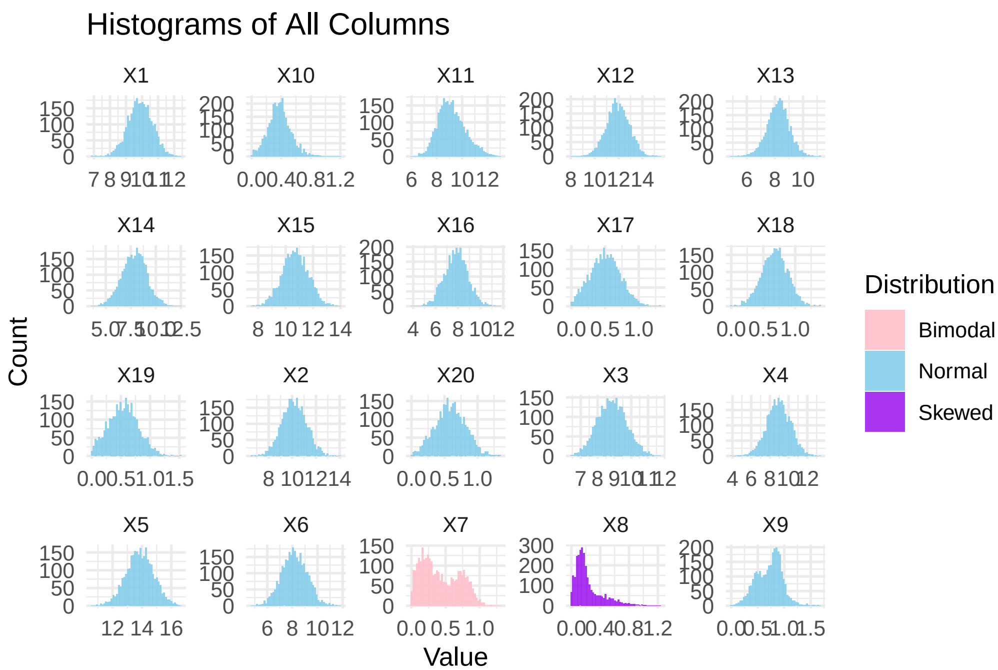

Figure 2: **Distributions within feature columns:** *histograms showing scale and skewness of data*

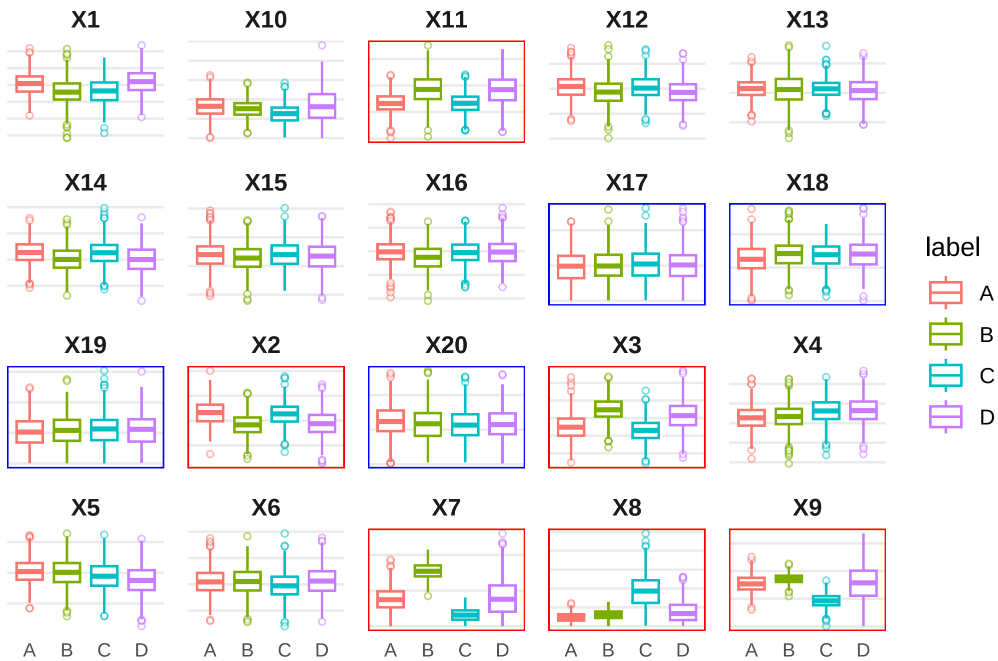

Figure 3: **Distributions within feature columns:** *Boxplots by class label by feature*

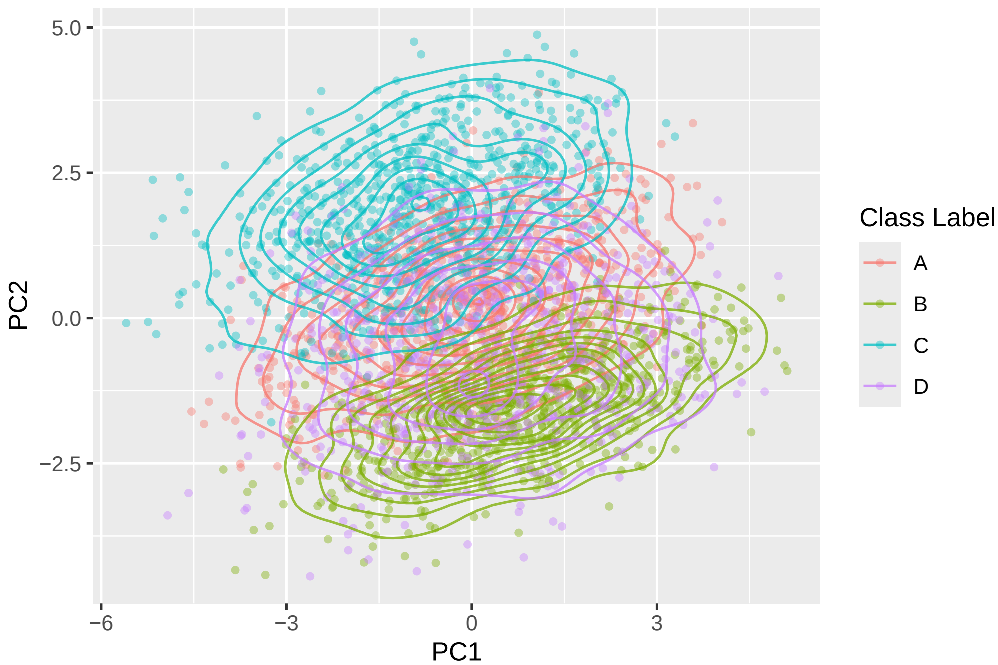

Figure 4: **PCA plotting of class labels:** *scatterplot showing clustering tendency of catagorical classes*

Table 2: Hopkins Statistic Scores

Columns used	Hopkins Statistic Score
All Features + Label	0.9999870
All Features with Label Removed	0.9999754
All Features Binary Class D vs Rest	0.9999785
Features X2,X3,X7,X8,X9,X11	0.9961250
Features X17,X18,X19,X20	0.9987979
Features X1,X4,X6,X10,X12,X13,X14,X15,X16	0.9938103

2.3 Supervised Learning Methods

logistic regression - including class weighting and L2 regularization, and feature selection. Random forest, including feature selection and tuning of mtry. SVM, with feature selection and tuning of kernel selection, gamma and cost parameters.

2.4 Unsupervised Learning Methods

Agglomerative hierarchical clustering, including tuning of linking metric. Gaussian mixture model based clustering, including selecting a model, regularization using shrinkage parameter, and dimensionality reduction using factor analysis.

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

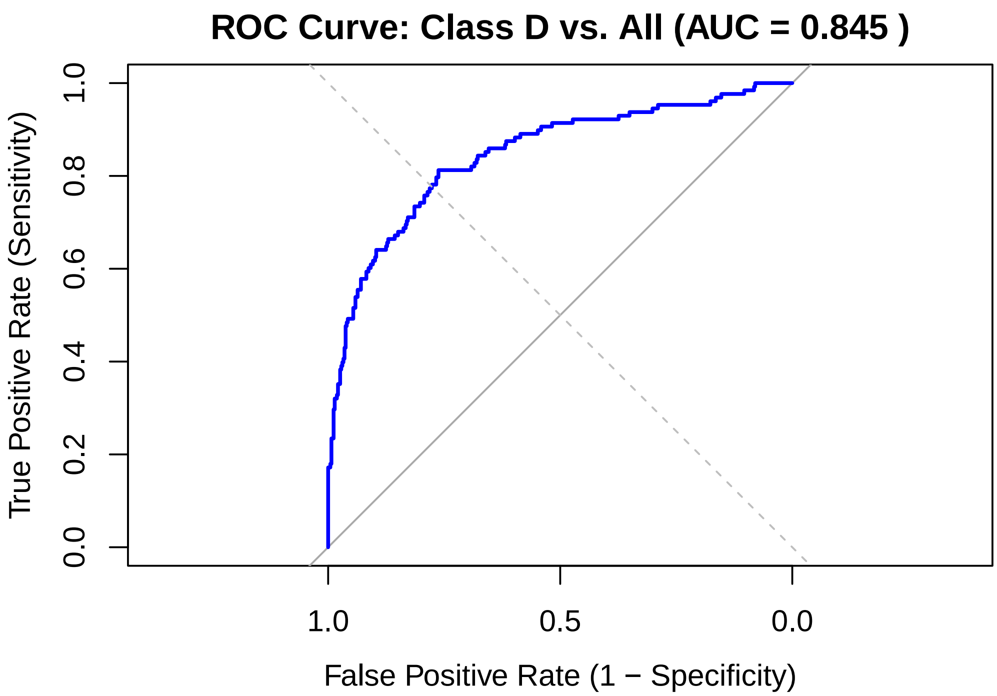

Figure 5: **Simple Logistic Regression ROC Plot:** *ROC plot for Class D vs Not Class D

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

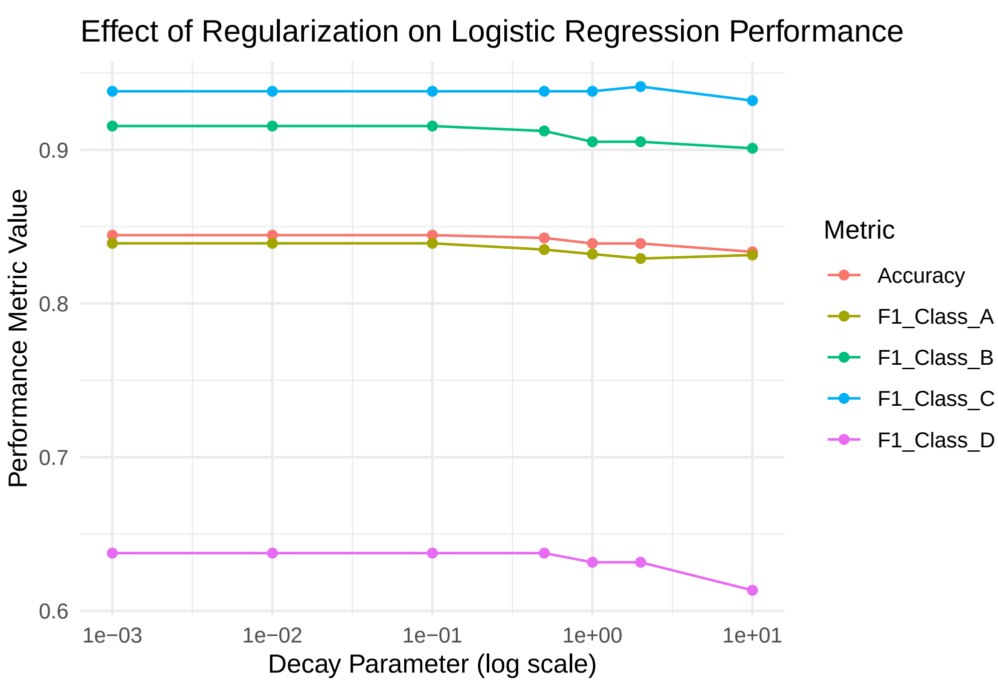

Figure 6: **Regularized Logistic Regression**

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

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