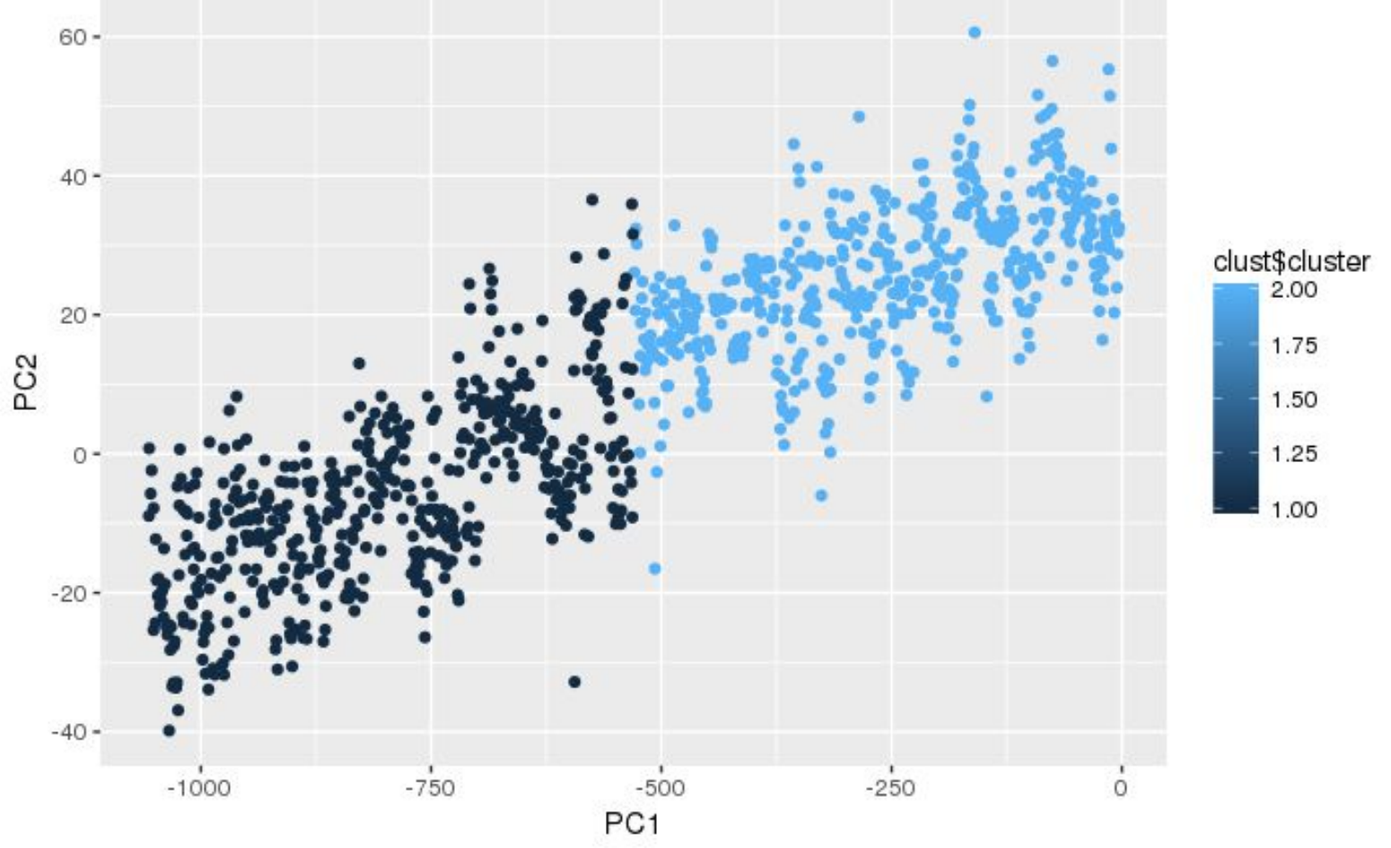
Weesnaw

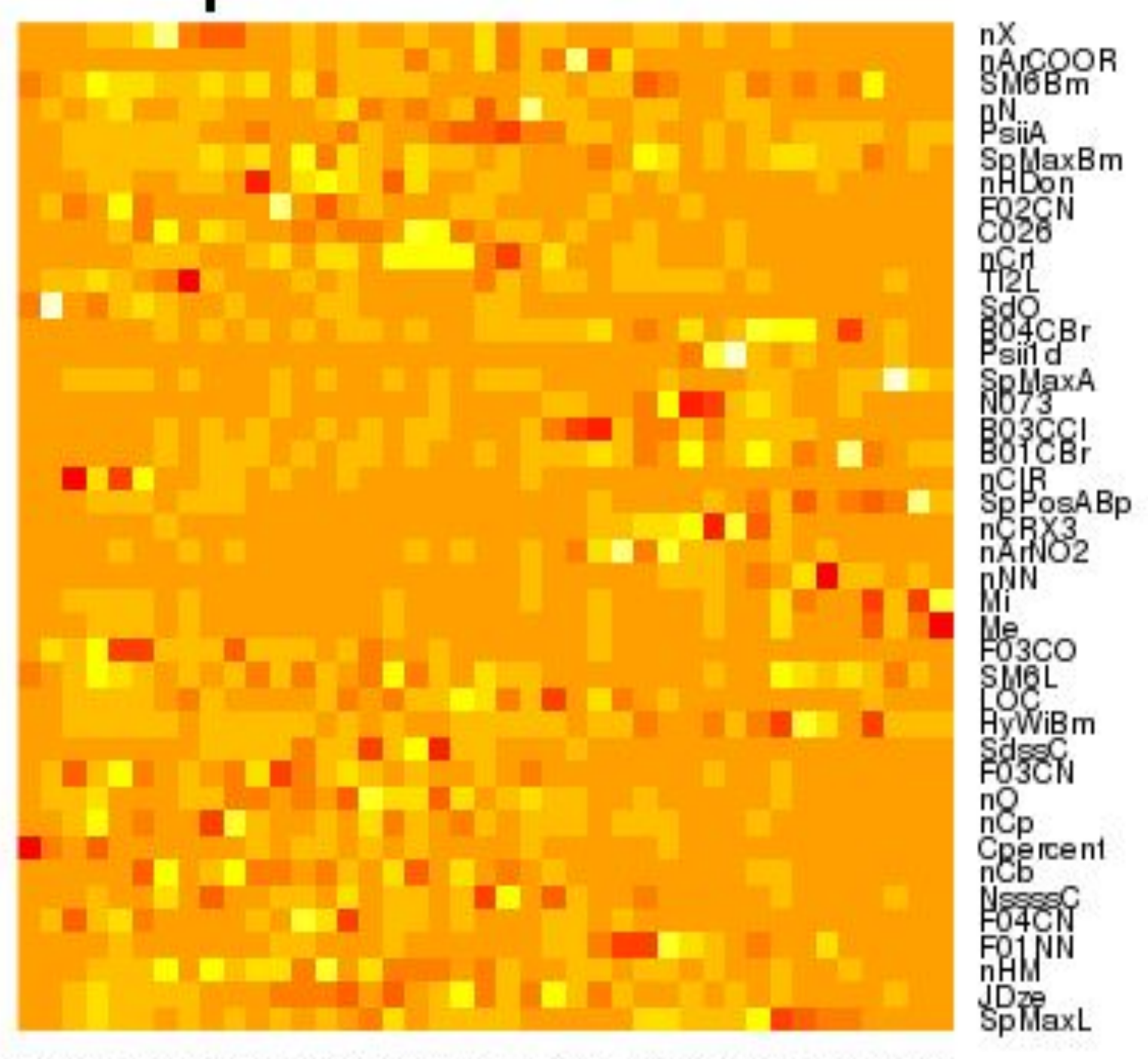
Data Analysis

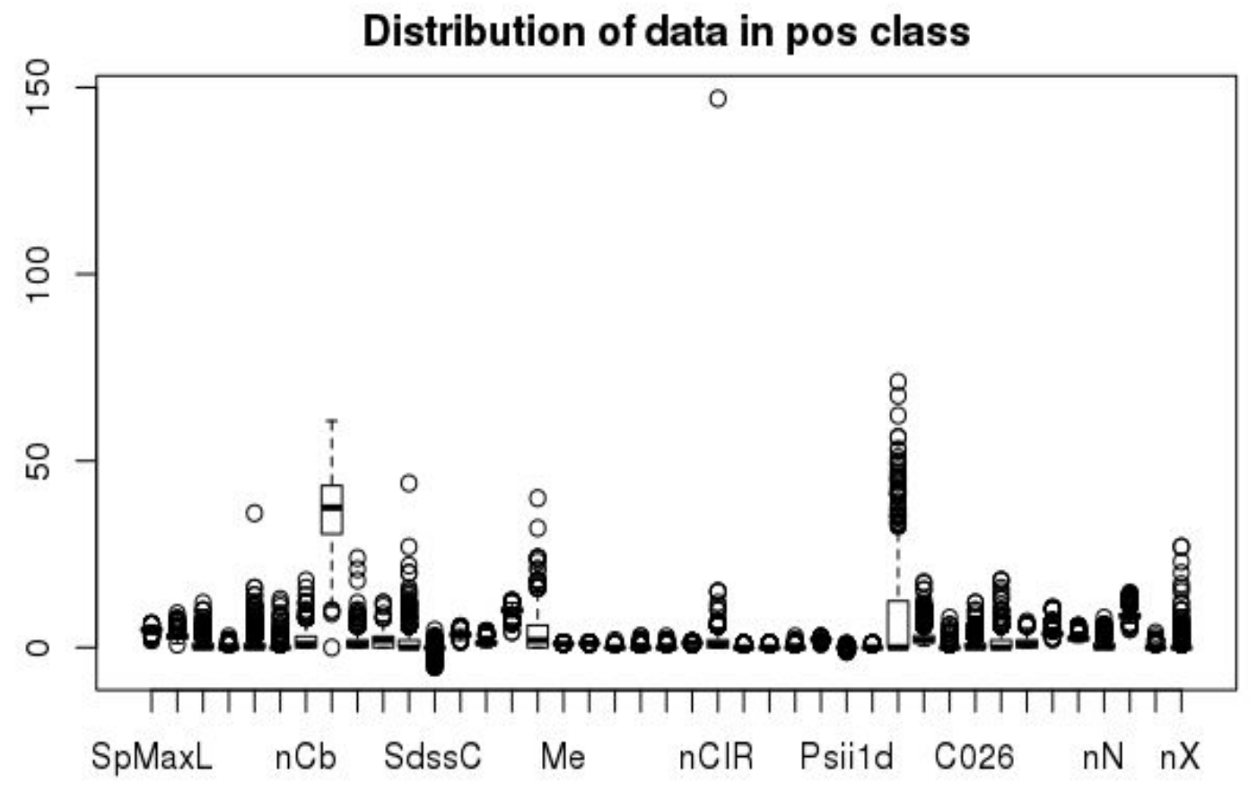


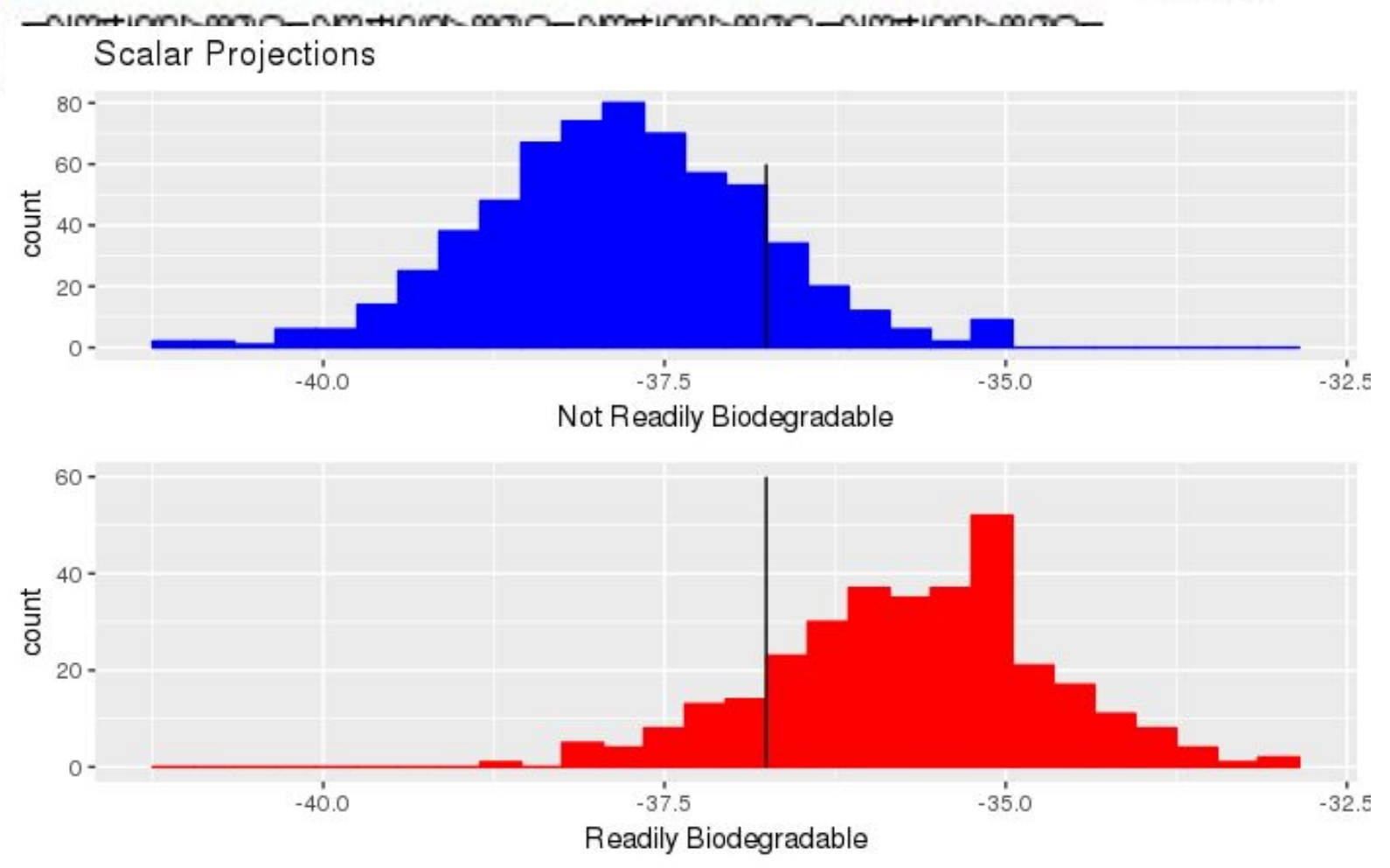
Introduction

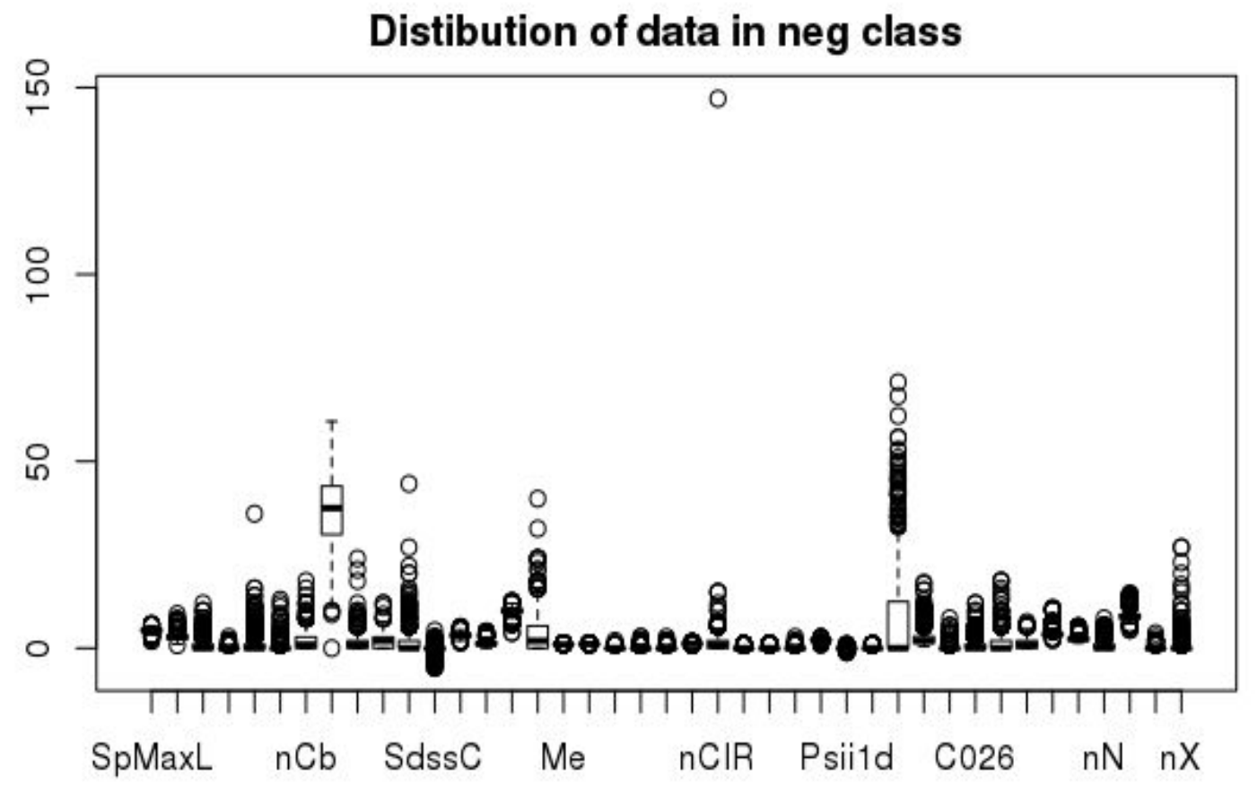
Our firm "Weesnaw" was hired by Chems-R-Us to develop a computational model that analyzes and generates the biodegradability of certain molecules. We were given a data set of with molecule data and 1,055 chemicals of biodegradable values. With the data from chemsrus.csv provided to us, we then created a model to predict the biodegradation of chemicals by using molecular data. The data signified where the last column determines whether or not the last column is biodegradable (either 1 or -1). Our job was to find the best way to classify these points.

Heatmap of mean of each class









3.0 Data Description

- This 'chemsdata.csv' dataset is consisted of 1,055 entries consisting of an id captioning the molecule as well as 41 attributes that the specific molecule has, and finally checking if it is biodegradable or not. After checking the classes, there are 356 biodegradable points and 699 non-biodegradable points
- The 'chemstest.csv' file that was given to us is a csv file filled with data of molecules explained by the 41 attributes but the biodegradability has not been provided, the class column is empty.

Presentation by: Ramin Chowdhury, Sebastian Castillo-Sanchez, Tenzin Tashi, Madison Chamberlain

4.0 LDA vs. Mean Method

- The results from running LDA and Mean Method show that LDA is more accurate
- LDA correctly classified 86% of the positive training data and 86.58% of the negative training data

	Training		Testing	
	Pos	Neg	Pos	Neg
LDA	86.06	86.581	75.758	84.932
Mean method	69.659	63.898	63.636	69.863

5.0 Model Improvements

Some ideas we tried were:

- Iterate over data points numerous times and update the normal vector when we found a misclassified point
- Another idea we had was to perform a pca and observe which components are well separated. When calculating the threshold, we find a method to give more weight to the well separated components since our classification methods prefer well separated data

6.0 Reccomended Model

Our previous tests indicate that LDA is superior to the mean method. Further prediction analysis claims LDA correctly classifies 86.4% of data while the mean method only classifies 82% of data correctly.

7.0 Additional Analysis

- Used ggplot function to examine different components to determine the best way to separate data
- It is shown that the data is not well separated in every component, which will explain why a number of points could not be classified correctly with our linear model
- Used scree plot function to show total variance in the data represented by the principal components

8.0 Conclusion

So in conclusion, the results from our analysis shows that the Fisher LDA Method is shown to be more accurate than the Mean Method. While we could not achieve better accuracy through other methods, we can still expect our model to work correctly for most points. This is expected since we are using a linear model on a data set that we saw is not well separated in all of its components. Our final consultation recommendation to Chems-R-Us is that they use our Fisher LDA model to predict the biodegradation of chemicals.