Predicting Amphetamine Use through Classification and Machine Learning

Annika White

Carlos Alfonso Ballesteros Carranza

Jose Alonso Serna Jacquez

Constantinos Anastasiou

Data Mining and Machine Learning (STATS5099)

22 March 2024

Introduction

The focus of this study is to construct classification models to predict how recently an individual has used legal and illegal drugs. The data set analyzed in this project contains survey information on 1885 individuals and how recently they reported having consumed amphetamines specifically. There is no distinction made in the data set as to whether those who identified as ever having used amphetamines were doing so legally or illegally. The independent variables within the data set include numerical values for personality measurements for neuroticism, extraversion, openness to experience, agreeableness, conscientiousness, impulsivity, and sensation seeing as well as demographic information detailing their education level, age range, gender, country of residence, and ethnicity. These demographic variables were initially reported as categorical variables and later quantified, but in order to aid the model building process, country of origin and ethnicity were transformed into sets of dummy variables.

In addition to self-reporting when the last time they last used amphetamines, the respondents were asked to report their use of a fake drug (Semeron) to identify those that claimed to use drugs more often than they actually do. These individuals were excluded from the data set going forward. Due to differing sample sizes across the original seven classes, several were combined to result in three final classes: never used amphetamines, used amphetamines in the last year, and last used amphetamines over a year ago. The data was then separated into training and testing sets in two different ways: first using an 80-20 split, and then with oversampling, which samples with replacement from smaller classes to further address the problem of vastly different class sample sizes. Five different classification methods were then employed to determine which is best able to predict how recently an individual has used amphetamines given what we know. These methods include neural networks, random forests, linear discriminant analysis, support vector machines, and k-nearest neighbors, each of which were then analyzed and compared to assess the best predictor.

Exploratory Data Analysis

	Nscore	*Escore*	*Oscore*	*Ascore*	*Cscore*	*Impulsive*	*SS*
Min	*-3.46*	*-3.27*	*-3.27*	*-3.16*	*-3.46*	*-2.56*	*-2.08*
1st Qu	*-0.678*	*-0.695*	*-0.717*	*-0.606*	*-0.652*	*-0.711*	*-0.526*
Median	*0.0426*	*0.00332*	*-0.0192*	*-0.0173*	*-0.00665*	*-0.217*	*0.0799*
Mean	*-0.00514*	*0.00283*	*-0.00435*	*0.00626*	*0.00201*	*0.00974*	*0.0021*
3rd Qu	*0.63*	*0.638*	*0.583*	*0.761*	*0.758*	*0.53*	*0.765*
Max	*3.27*	*3.27*	*2.9*	*3.46*	*3.01*	*2.9*	*1.92*

Table 1: Summary Statistics of our Numerical Variables (3s.f)

The summary statistics on *Table 1* give valuable insights into the training set's numerical variables. Firstly, the mean and median are all very close to each other indicating potential symmetric distributions. Now by looking at the 1st and 3rd quartiles they all have quite similar IQRs (Interquartile range), hence the personality measurements mentioned above exhibit a similar degree of variability. Finally, by examining the minimum and maximum values they all seem to be on the same level with no indication of extreme points (they all mostly lie between -3.5 and 3.46).

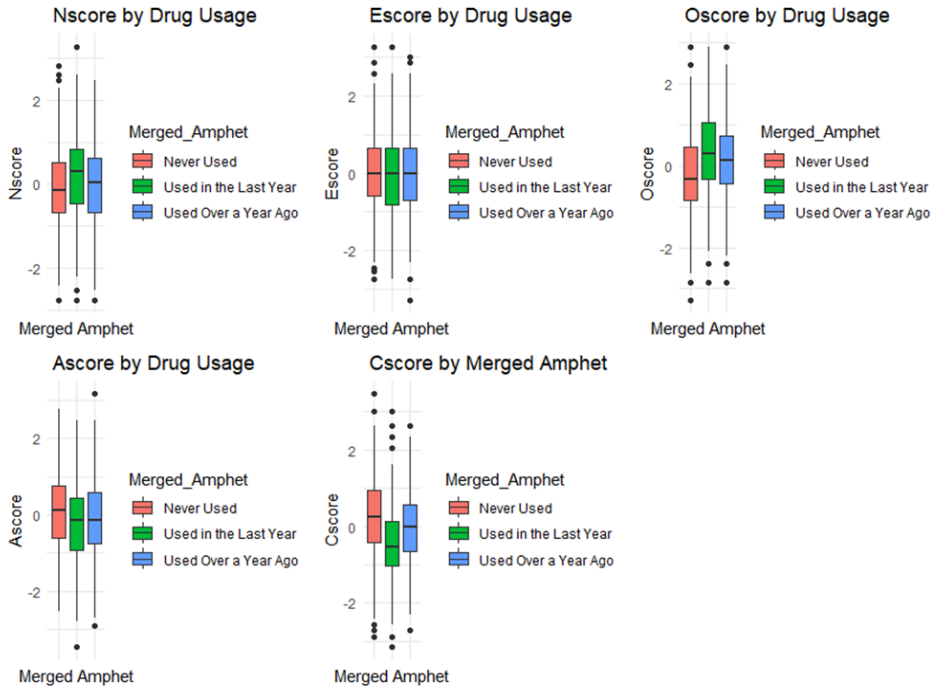

Figure 1: Boxplots of personality measurements against Drug Usage

Firstly, from *Figure 1* it is obvious that most outliers can be found in the Conscientiousness (Cscore) boxplots while Agreeableness (Ascore) have the fewest. Additionally, the boxplot representing the Used in the Last Year class has a higher median than the other classes in the Neuroticism (Nscore) and Openness to experience (Oscore) plots, while in the Ascore and Cscore it has a lower median than them. Extraversion (Escore) seems to be the most symmetrical out of them with all of the classes having the same median and almost the same quartiles (there are some minor differences on the 3rd quartile value). For these reasons, Escore, Ascore, and Cscore were excluded from the models.

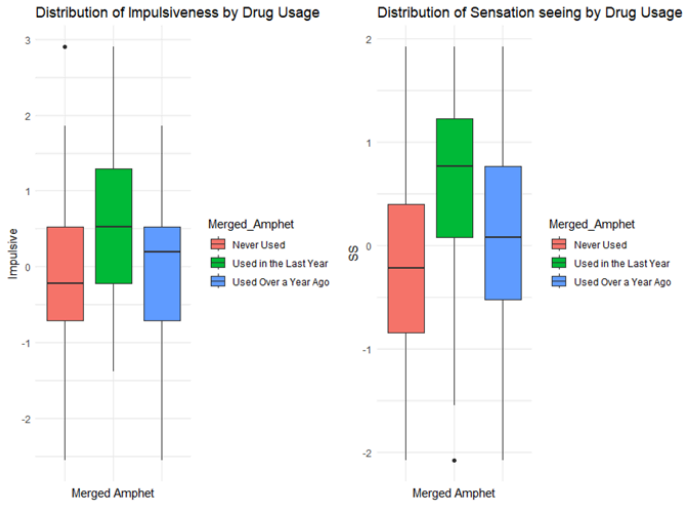

Figure 2: Boxplots of Impulsiveness and Sensation Seeing against Drug Usage

Based on *Figure 2* the used in the last year boxplot once again has a higher median than the other ones in both the Impulsiveness and the Sensation Seeing graphs, while the "used over a year ago" boxplot has a higher median than the "never used" boxplot in both cases as well. Their IQRs seem to be fairly similar so their spread is expected to be similar as well. It is also good to note that no huge number of outliers is observed on either of these plots.

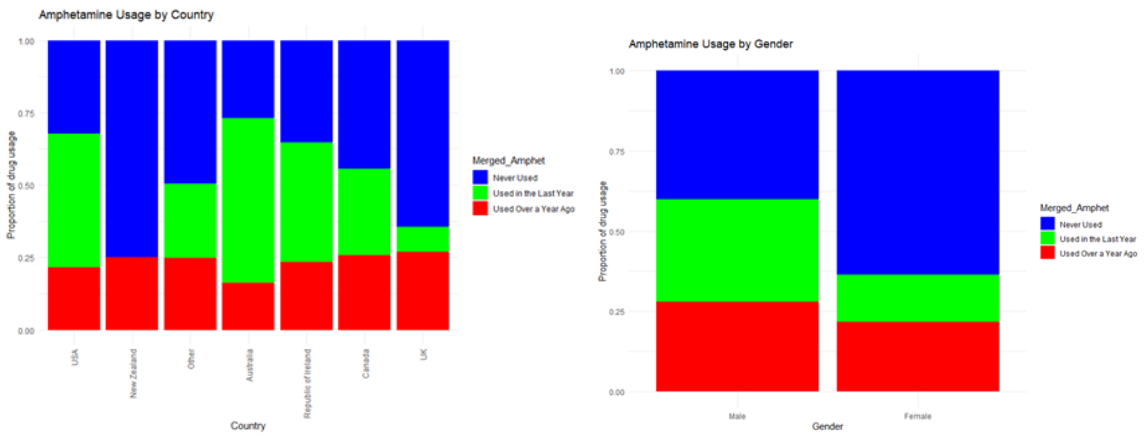

Figure 3: Bar charts of Drug usage against country and gender

It can be deduced from *Figure 3* that the country that has the most people who have never used amphetamines in the sample is New Zealand while Australia has the lowest proportion of them and has the largest proportion of people who used in the last year out of all the countries. Australia also has the lowest proportion of people that used over a year ago but the difference with the proportions of the other countries is very low. Furthermore, women seem to have the largest proportion of never used, and the lowest proportion of used in the last year and used over a year ago, as opposed to men.

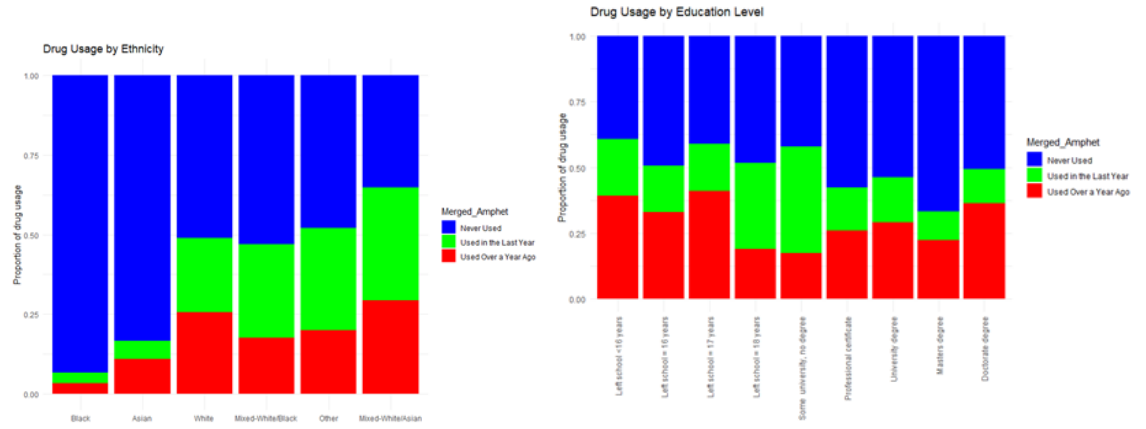

Figure 4: Bar charts of Drug usage against Education level and ethnicity

From *Figure 4* it can be observed that most people who have never used amphetamines are Black, followed by Asians, while Mixed-White/Asians have the lowest proportion of

people who have never used them. From the education level bar chart, one would expect that as you increase the education level of each individual you would get a larger number of people who have never used any. Instead, this pattern is present in people who used drugs in the last year. As the education level increases the proportion of people using amphetamines in the last year decreases.

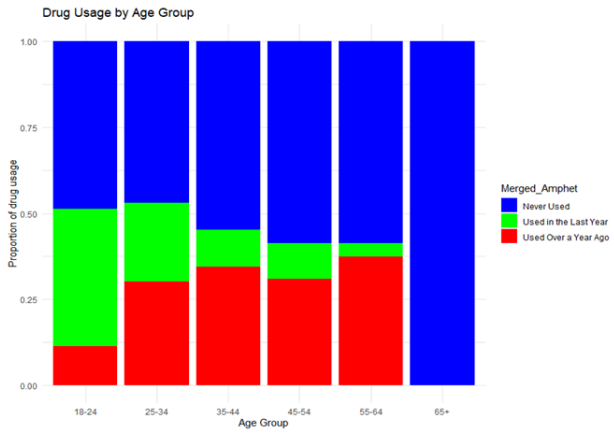

Figure 5*: Bar chart of Drug usage against different age groups*

From *Figure 5* people aged 65+ have never used amphetamines, whilst most people who used drugs in the last year are between 18 and 24 years of age, which is expected, as younger people usually tend to be the ones using them, especially in university and college. Based on the plot, in the dataset, as the age group increases, the proportion of people who have used amphetamines in the last year decreases, and the proportion of people who used them over a year ago increases as one would expect.

Methods

Support Vector Machines (SVM)

Support Vector Machines (SVMs) stand out as a formidable tool for classification tasks within the sector of machine learning. They are widely favored for their adaptability and precision. Notably, they often deliver outstanding results, especially when confronted with scenarios where the number of variables surpasses the number of observations within the

training set. This superiority stems from the characteristic that the quantity of parameters in SVMs is independent of the number of variables but instead relies on the number of training instances. The core concept of SVMs revolves around the endeavor to establish a hyperplane capable of classifying the data points into segregated regions.

Firstly, what was done to correctly implement this method was to create possible parameters to get the best possible model we can afterwards. The parameters of the SVM function are pivotal in shaping the behavior and performance of the Support Vector Machine model. These parameters, including the kernel type, cost (C), gamma (for radial kernel), degree (for polynomial kernel), enable customization of the SVM model to suit the specific characteristics of the dataset and desired outcome. The choice of kernel function, such as linear, polynomial, radial, determines the transformation of the input space to a higher-dimensional feature space for potential linear separation. The cost parameter regulates the trade-off between maximizing the margin and minimizing classification errors, while gamma controls the influence of individual training samples on the decision boundary, affecting the smoothness and complexity of the boundary. Similarly, the degree parameter in polynomial kernels adjusts the nonlinearity of the decision boundary.

Therefore, the polynomial model and radial model were tuned and their summary outputs were compared in order to get the best possible model parameters. After comparing them, the radial model seemed to be the better fit with the following parameters leading to the lowest error: cost =10, and gamma =1 . So, the final radial SVM was fitted with the parameters mentioned above and C-classification and predicted the classes of the test dataset. Then, a confusion matrix was constructed to compare these predictions against the actual classes in the test dataset. From the confusion matrix, the overall accuracy of 0.5718085. Thus, approximately 57% of the instances in the test dataset were correctly classified by the SVM model.

Linear Discriminant Analysis (LDA)

Moreover, Linear Discriminant Analysis (LDA) was applied as a linear method for classification in our dataset. This technique bases itself on dimension reduction which determines a subspace that can be compared to the original data sample (Xanthopoulus et al., 2013, p.27). In turn, this process creates separate classes using boundary decisions. This method is generally used when the response variable (or classes) has a linear relationship with the explanatory variables and can be commonly applied for pattern recognition, category detection, and behavior analysis. LDA's importance derives from its versatile nature that allows its application in many different fields and analyses.

This method was selected since it provided a powerful and robust method for predicting the different classes. Not only does it reduce the dimensionality, but it also extracts the most discriminative (linear) features between classes, offering insights into the structure of the data and helping identify important variables for classification. LDA is computationally efficient compared to more complex ML algorithms.

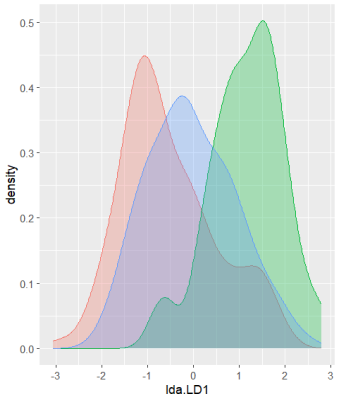

Fig. 6

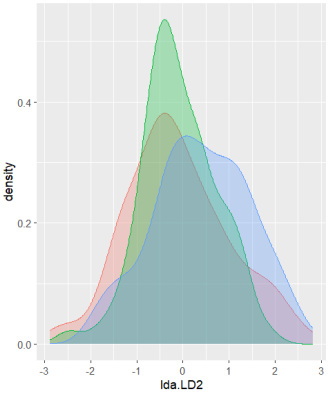

Fig. 7

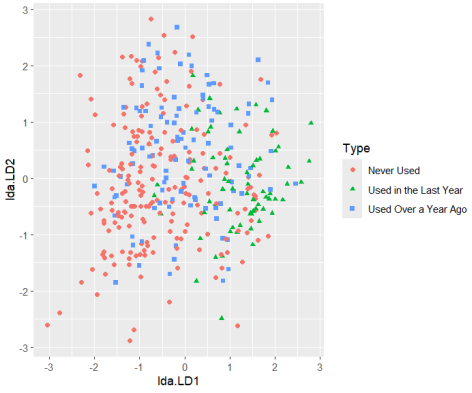

Fig. 8

After running the LDA, two linear discriminants (LD1, LD2) were obtained with a proportion of trace of 0.84 and 0.16 respectively. As seen in Figure 6. LD1 was able to separate the "Used in the Last Year" fairly well from the other two categories. Unlike LD2, where there was no clear boundary decision (Figure 7). When plotting LD1 vs. LD2, some clusters

of classes can be moderately observed, however, the result was not definitive as the observations rested on top of each other in many instances (Figure 8). The model predicted fairly well in the “Never Used” and “Used in the Last Year” categories, but in the “Used Over a Year Ago” class, the model faced issues erroneously predicting it as the other two categories. The results showed a 54.78% accuracy, meaning that even when the data was similar within the classes, the model was able to improve the accuracy from the prior probability.

k-Nearest Neighbors (kNN)

The k-nearest neighbors classification aims to categorize each new observation based on the class that each of its k-nearest neighbors fall into. It is simple in understanding and implementation, and it can easily handle non-linear data like that of the data set used here since it makes no assumptions about the data itself, which is why it was chosen as a classification strategy for this project. However, this method has a lot of ambiguity in how to choose the best model within this classification method, as its performance is sensitive to some aspects such as the value of k chosen, what distance calculation is used, and what variables are kept in the model. To overcome some of these limitations, leave-one-out cross-validation was used to choose the optimal value of k, and extensive exploratory data analysis was conducted to make informed decisions on which variables were kept in the final model. Euclidean distance was used as it is the default distance metric in R. K-nearest neighbors can also be rather computationally costly, but this is a reasonably small data set, so this was not an issue in this case.

As previously mentioned, this method is sensitive to several factors that can be changed while fitting the model. While cross-validation was used to minimize the limitation of choosing the optimal value for k, this data set includes dummy variables and several categorical variables, so a different distance metric that can handle categorical and dummy variables may have yielded better results. Unlike the other classification methods, the k-nearest neighbors model was fit with a simple 80-20 training and testing split of the data.

This is because if the oversampling data split had been used instead, the optimal k would be 1, as samples that were repeated could be predicted with 100% accuracy.

To find the optimal number of neighbors for prediction, leave-one-out cross-validation was used on the training data set for a range of k-values from 1 to 50. The value of k that produced the highest value for the correct classification rate for the training data was 38. The final model then used the 38-nearest neighbors in the training set to predict the class of the individuals in the testing data set. The final correct classification rate, or the accuracy score, was calculated as the percentage of correctly classified observations in the testing set, which was 59.31%.

Neural Network

Artificial neural networks are robust non-linear models that are composed of layers and nodes. The output of each layer is connected to every node in the following layer with a different weight, then the constant (bias) term is added to each node and it is finally transformed by an activation function. This method is powerful for both regression and classification problems as it is able to identify intricate patterns in the data through the use of nonlinear functions and hidden layers.

A neural network with 2 hidden layers each containing 50 nodes was chosen. It was not possible to fit the model using the full training data set due to computational cost, so the training data set was reduced to 80% of its size. The number of nodes was set at 50 given that trials with fewer nodes did not converge to the minimum error threshold required by the model.

The final testing accuracy was 52.12% which was a significant decrease from the 95.17% accuracy obtained in the training data set. Different iterations of the model were attempted to improve testing results, but no significant improvement was achieved.

Random Forest

The random forest classifier is composed of several classification trees. Each individual tree is fit on a sample with replacement of the original data set (bagging) and a random subset of the original variables (to decorrelate each tree). The result of each tree is recorded and then averaged to decide which class obtained the majority vote.

A random forest classifier with 150 trees was fitted, which yielded a 57.71% testing accuracy. The three most important variables by mean decrease gini were NScore, Oscore and SS. Suggesting a strong relationship between neuroticism, openness to experience, sensation seeing, and drug use.

Results

As seen by the tables in figure 9, none of the models performed overly better than others. The highest accuracy score came from the k-nearest neighbors approach, which had a correct classification rate of nearly 60%. It also had some of the highest scores for other metrics such as F1 score, sensitivity, etc. for the “never used” class. However, this is the only method that was unable to use the oversampled data set, meaning that it may have a higher accuracy simply because it could achieve a high score by classifying most people as never having used amphetamines. This is further confirmed by the comparison of sensitivity and F1 score across classes for this method, as most of the metrics show high performance for the “never used” class and poor performance for the “used in the last year” and “used over a year ago” classes.

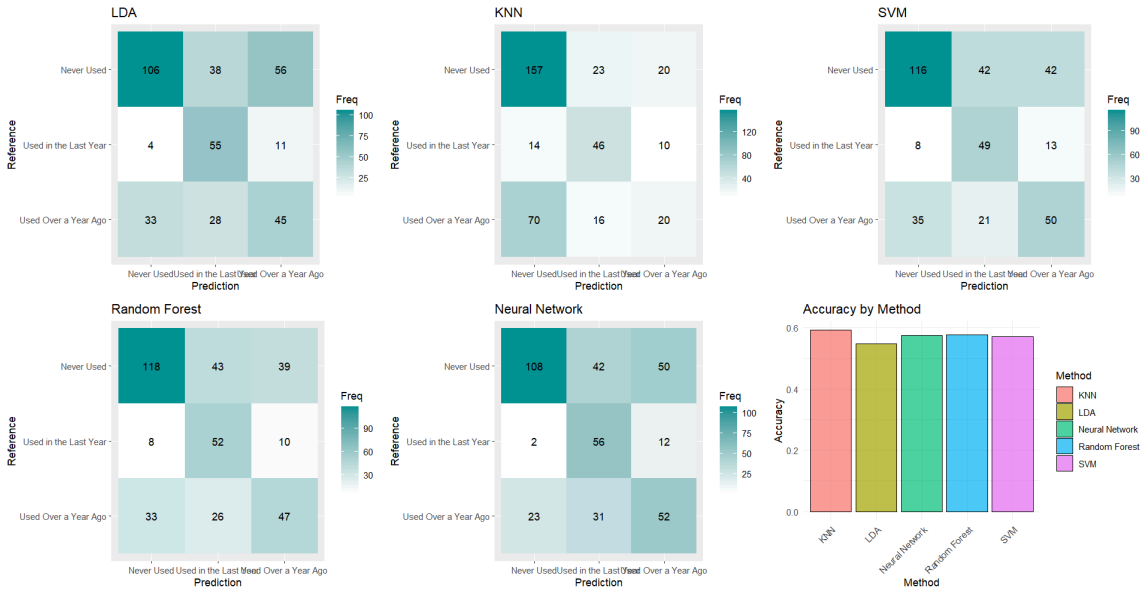

Figure 9: Confusion matrices and accuracy scores

Given that the accuracy may be misleading due to different sampling methods in the training set, it was not used as the primary assessor of performance. It has been concluded that the best model for predicting how recently an individual within our sample has used amphetamines is through the random forest approach. While this method has the next highest accuracy score (after kNN), it also performs decently well across all other performance metrics and all three classes relative to the other methods, as seen in figure 10. More specifically, it has very consistent results for the 6 metrics in figure 10 unlike other methods, such as neural networks or linear discriminant analysis, which have relatively high performance across some classes and relatively low performance across others. Since none of these methods had exceedingly high correct classification rates or predictive performance in general, none of them seem to be particularly fitting for the data in the sample, but it has been determined that the random forest is the most adequate from the information gathered.

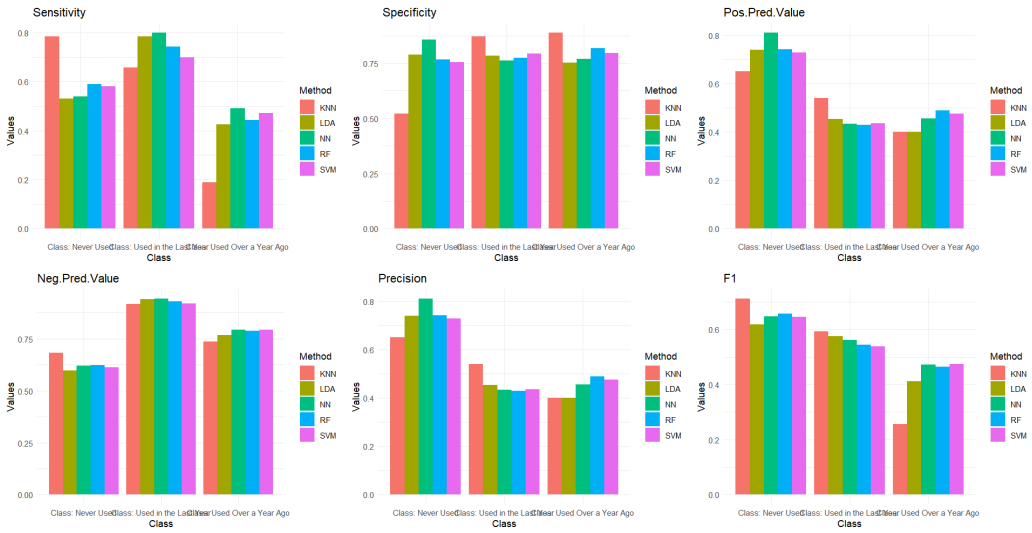

Figure 10: Performance metrics across classification methods and classes

Conclusions

Prediction for this data set proved rather difficult due to a number of limitations previously highlighted, such as differing sample sizes across classes, having quite a few categorical variables, etc. Oversampling, reducing the number of classes, introducing dummy variables, and more were all strategies employed to reduce the impact of these limitations, but they still proved to negatively impact the predictive ability of several of the models fitted to the training data.

Random forest seems to be the best approach to predicting whether individuals should be classified as never having used amphetamines, used amphetamines in the last year, or last used amphetamines over a year ago. When compared to the other methods employed in this project, the random forest model seems to have a relatively high accuracy as well as consistently high performance indicated by the sensitivity, specificity, and F1 scores across the three classes.

References

Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Robust Data Mining. Springer New York.