Behavioral and non-behavioral factors leading to depression diagnosis in Behavioral Risk Factor Surveillance System 2017

# Introduction

This analysis will make use of the comprehensive survey conducted by CDC in analyzing the role of behavioral (i.e., alcohol, drug use) and non-behavioral factors (i.e., alcohol, drug use) to predict which patients may be at risk for depression. In this attempt, a supervised learning models were built using patient information from the survey who have been measured on their depression diagnosis to determine the most consistent variables leading to depression.

The main research question attempts to answer which behavioral and non-behavioral factors contribute and as a result, predict a positive depression diagnosis. This research will mainly use various feature selection methods such as chi square to then apply a random forest and logistic regression model to predict the most important predictors of a depression diagnosis. With knowledge of the most prevalent predictors, this information can be used by healthcare professionals and policy analysts for use in mental illness prevention and treatment. Note, the target variable is ADDEPEV2 and signifies a positive (=1) or negative depression diagnosis(=2) in the test subject.

Preliminary Notes: \*Due to the size of the data of 400K+ rows, 100K rows were sampled and all consequent analysis were performed on the sample. \*\*Some variables which begin with “X\_” as in “X\_ MISFRT1” may appear without the “X\_” in the codebook provided by CDC.

# Literature Review

#### 1 Burcu, Malecki and, Engelman(2018) - Using recursive feature elimination in random forest to account for correlated variables in high dimensional data

In this paper, the authors concluded that using a variant of recursive feature elimination paired with random forest which is also used in this analysis was ineffective when it comes to highly dimensional data. In addition, an interesting finding was that even though previous literature had suggested reducing the mtry value to .1\*number of features in dataset for highly dimensional data, a better model was obtained using the default mtry value of square root of p(features). The random forest model in this analysis hopes to see if using 10% of all features produce similar results.

#### 2 Jiarpakdee, Tantithamthavorn, and Treude(2018) - AutoSpearman: Automatically Mitigating Correlated Software Metrics for Interpreting Defect Models

The packages and functions used in this analysis were derived from this article. Chi squared, and recursive feature elimination used in this analysis were derived from this article

#### 3 Khalilia , Chakraborty and Popescu (2011) - Predicting disease risks from highly imbalanced data using random forest

In this paper, the authors used a very high dimensional and very large dataset similar to the one used in this analysis. Also, similar to the dataset used in this analysis, Khalilia et al., also identify a high class imbalance. Despite the high class imbalance, and large number of features, the authors concluded that random forests provide the most accurate group of predictors.

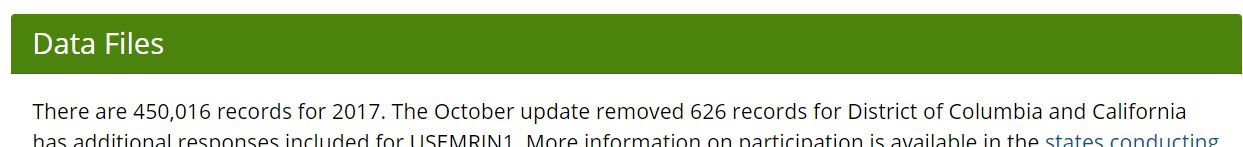
#### 4Wray et. al (2019). Using Smartphone Survey Data and Machine Learning to Identify Situation and Contextual Risk Factors for HIV Risk Behavior Among Men Who Have Sex with Men Who are Not on PrEP. Society for Prevention Research. In this paper, the authors attempt to identify most risky sexual behaviors or predictors of HIV. Similar to this analysis, the paper also uses survey data to find the most important predictors using random forests and logistic regression models. Upon performing random forest and logistic regression, the significant measures found by ROC and AUC were those belonging to random forest models. ROC and AUC will be utilized in this analysis as evaluation measures as well.

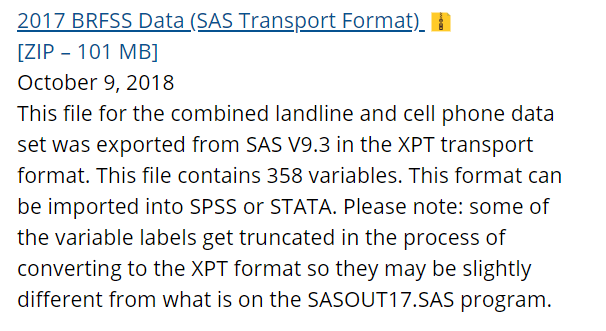
# Dataset

### Source references:

### Codebook: <https://www.cdc.gov/brfss/annual_data/2017/pdf/codebook17_llcp-v2-508.pdf>

#### Datafile found here: <https://www.cdc.gov/brfss/annual_data/annual_2017.html>





# Approach

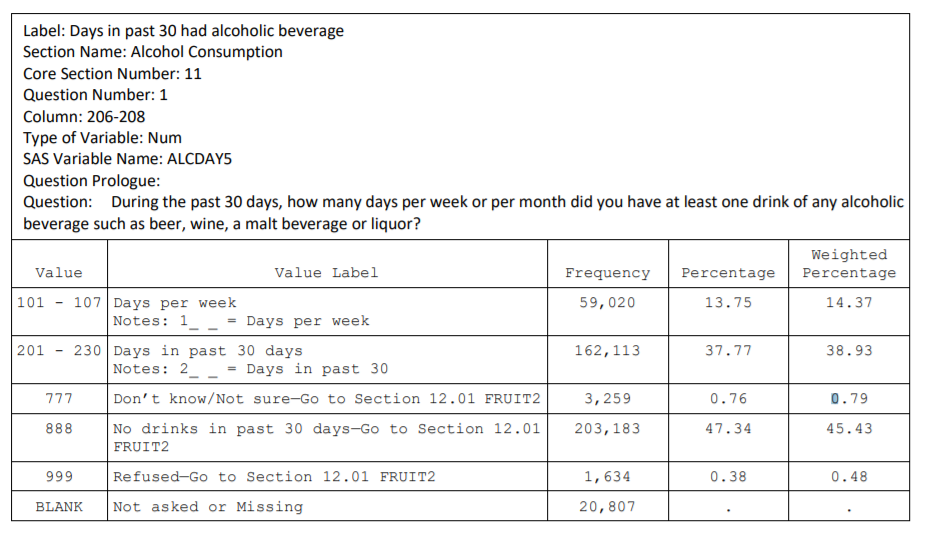
## Step 1: Data preparation

#### *Data import*: The data was provided in the form of a .sas file format and was converted using the read.xport() function. The converted .csv file too large to upload is converted within the dataCleaning.R file.

Data cleaning: Four main types of data cleaning were performed for missing values, re-scaling, re-encoding variables, and factorizing/re-factorizing variables.

All variables which were identified as having 50% or more missing values were removed from analysis. Deleted variable and included variables are identified in the file labeled dataDescription.csv on Github and in submission on the course shell.

Some variables were computed because of the mixed scales in their parent variables. For example, the variable ALCDAY5 also shown below was reported using 2 different scales – days per week (101-107) and days per month (201-230). As a result, a variable called alc30Week is created with an assigned value of 1 if value in ALCDAY5 is between 101-107(inclusive) and assigned value of 2 if value in ALCDAY5 is between 201-230(inclusive). These variables can be found in the dataDescription.csv spreadsheet and the encoding procedure in the file dataCleaning.R.



Next, values encoded as “Don’t know/Not sure” or “Refused” were re-encoded as NAs. Using the example of ALCDAY5, the values 777 and 999 were re-encoded as NAs. Also, 888 and other value which represent a response of “No drinks” or “Never” were re-encoded as zero.

In addition, some variables were removed from analysis as they posed no real value to the research question at hand. For instance, variables such as X\_MISFRT1 was removed as it represents missing values for fruit intake variable and thus, not useful for analysis. Other variables such as IDATE and SEQNO, were removed as they were not considered meaningful predictors. WEIGHT2 and HEIGHT3 were removed because they were reported on different scales (like ALCDAY5) and HTIN4, HTM4, WTKG3 were retained as they were computed using the measures of inches, meters, and kilograms respectively.

Some re-categorization also was performed as some variables had 53+ categories. EXRACT11 and EXRACT21 were found to have 90+ categories and so, a threshold of 60% or 45 categories was taken based on frequency of values in the category. A new category of 46 or “other” was created with all frequencies made up of less than 60% . Finally, all levels from all categorical data which were unused or empty, were dropped.

**Github reference**: Lines 6:793 in dataCleaning.R

## Step 2: Imputation

Imputation by median was performed for continuous/numerical variables using the user defined function impute(). Median was chosen due to the continuous variables having skewed distributions. Imputation by mode was performed for categorical variables using the user defined function mode(). Outlier imputation was excluded from this analysis as the numerical data lay on a discretized scale and thus, wasn’t prone to inconsistencies. Categorical data was not corrected as responses also followed a discretized scale and thus, treating outliers may introduce bias and rid the data of any meaningful insights.

**Github reference**: Lines 801:983 in dataCleaning.R

## Step 3: Feature selection

Two feature selection (filter and wrapper based) techniques were used to find the best predictors: chi square, and recursive feature elimination

#### Chi square: Chi square feature selection technique was performed using the FSelector package and the chi.squared() function. The following predictors highlight the top 75% (129) of predictors chosen:

**Table 1**: Features selected using chi square technique

|  |  |  |  |
| --- | --- | --- | --- |
| MENTHLTH\*  X\_MENT14D\*  DECIDE\*  POORHLTH\*  PHYSHLTH\*  GENHLTH\*  X\_LMTSCL1  X\_PHYS14D\*  EMPLOY1\*  DIFFALON \*  X\_RFHLTH  X\_LMTACT1\*  X\_LMTWRK1\*  DIFFWALK  HAVARTH3\*  X\_DRDXAR1  DIFFDRES  INCOME2\*  X\_SMOKER3\*  X\_INCOMG  X\_ECIGSTS\*  X\_ASTHMS1  CHCCOPD1\*  ECIGARET\*  ASTHMA3  X\_LTASTH1  MEDCOST\*  X\_CASTHM1  X\_RFSMOK3  MARITAL\* | X\_BMI5\*  SMOKE100  HIVTST6\*  X\_AIDTST3  SEX\*  BLIND\*  RENTHOM1\*  CADULT\*  FC60\_\*  MAXVO2\_  X\_BMI5CAT\*  X\_PACAT1\*  X\_PA150R2  TOLDHI2  X\_RFCHOL1  X\_PAINDX1  FVGREEN1  X\_PA300R2  X\_AGEG5YR  EXERANY2  X\_TOTINDA  HTIN4\*  GRENDA1\_  X\_CURECIG  X\_AGE80  HTM4  EXRACT11\*  FRUIT2  DIABETE3  X\_FRUTSU1  EXRACT21\* | FRUTDA2\_\*  PA1VIGM\_  X\_PAREC1  X\_MINAC11  EDUCA  PA1MIN\_  ACTIN11\_  PAMIN11\_  X\_STATE\*  BPHIGH4  CHCKIDNY  CVDSTRK3  HIVRISK5\*  X\_MICHD  PAFREQ1\_  X\_EDUCAG  greenIntake\*  PAVIG11\_  X\_AGE\_G  X\_FRTLT1A  X\_RFHYPE5  X\_AGE65YR\*  CVDCRHD4  X\_MINAC21  X\_RACE\*  METVL11\_  EXERHMM1  fruitIntake  PAMIN21\_  PADUR1\_ | X\_IMPRACE\*  VEGETAB2  POTATOE1  X\_MRACE1  WTKG3  DEAF\*  X\_PA30021  ALCDAY5  X\_PRACE1\*  ACTIN21\_  METVL21\_  X\_VEGESU1  PERSDOC2\*  PAVIG21\_  X\_PASTAE1  vegIntake  EXEROFT1  CVDINFR4  PREDIAB1  X\_RACEGR3  FRENCHF1  FRUITJU2  STRFREQ\_  POTADA1\_  X\_RACE\_G1  X\_DRNKWEK  FRNCHDA\_  X\_PASTRNG  STRENGTH  PNEUVAC3\*  DRNKANY5\*  DROCDY3\_\*  X\_DUALUSE  potatoIntake  VEGEDA2\_  X\_VEGLT1A  X\_RACEG21 |

### \*Feature appears in all models which may signify importance

*Recursive Feature elimination*: Recursive feature elimination technique was performed using the rfe() function using random forests. The data was trained on 10 samples in cross validation style and the following predictor was selected: ACTIN11\_. Since, only one predictor was selected, this feature elimination method was not explored further. This may further be supported by the study cited in Burcu, Malecki and, Engelman(2018) which concluded that RFE for random forests may not be the best feature selection method for high dimensional data.

**Github reference**: Lines 11:15, 470:483 in firstModel.R

## Step 4: Class imbalance

Class imbalance was addressed using three techniques: oversampling, undersampling, and a combination of both. The analysis was performed using the ROSE package and the ovun.sample() function. Parameter of p=.5 was applied to sample probability of the majority or minority class respectively. Class imbalance techniques were applied to the training dataset of 70%. Note, depending on the sampled dataset, the percentage of training set data shifted from the 70% baseline, but the test dataset stayed at 30% of total data set.

**Github reference**: Lines 37:50 in firstModel.R

# Step 5: Model results

### Random forests using Chi Square Feature selection

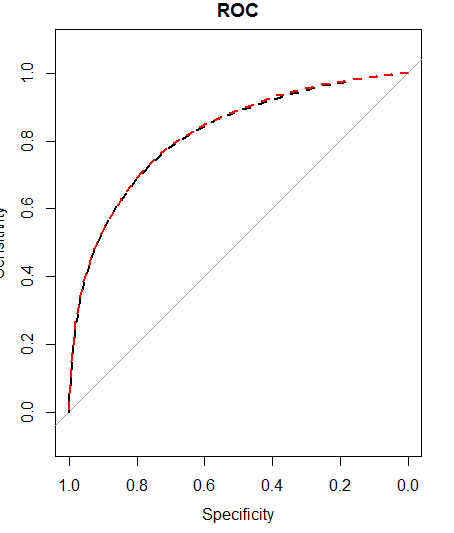
Random forests were performed using the randomForest package and function. The parameters of sampsize and mtry were utilized in the training model. Sampsize() was used to improve the speed of the model and address the highly unbalanced data set. In other words, the lowest frequency of the two classes were inputted in sampsize(). Mtry() was set to number of features times .1 given the high dimensionality of the data. Undersampling was the least accurate of all sampling methods and thus, was dropped from further analysis. Area under the curve stayed consistent across all models.

**Table 2**: Results for all random forest models using oversampled, undersampled or both sampling techniques

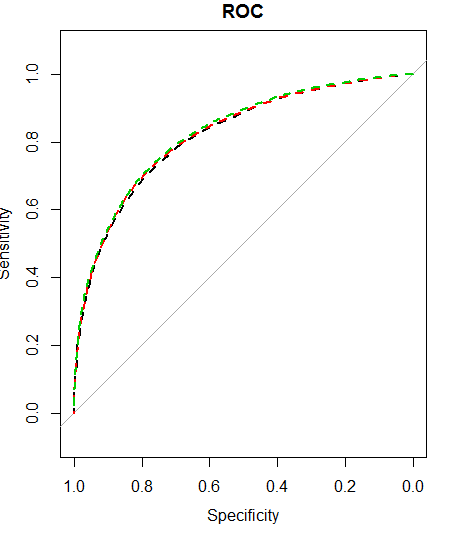
|  |  |  |  |
| --- | --- | --- | --- |
| **Metrics** | Ntree = 50  Mtry = 13 | Ntree = 100  Mtry = 13 | Ntree=500  Mtry = 13 |
| **Oversampling** | | | |
| Accuracy | 0.8373 | 0.839 | 0.8411 |
| Precision | 0.6065608 | 0.6145833 | 0.5477501 |
| Recall | 0.4868354 | 0.48106 | 0.49006 |
| F-measure | 0.5401432 | 0.5396856 | .54 |
| AUC | 0.1805761 | 0.1775014 | 0.1713477 |
| Kappa | 0.4428 | 0.4439 | 0.4531 |
| Time taken to train model | 4.665039 mins | 9.497573 mins | 43.74194 mins |
| Time taken to test model | 3.032154 | 4.526209 secs | 23.64517 secs |
| **Undersampling** | | | |
| Accuracy | 0.7173 | 0.7186 | - |
| Precision | 0.3897484 | 0.3906171 | - |
| Recall | 0.7788347 | 0.7849499 | - |
| F-measure | 0.5195173 | 0.5216459 | - |
| AUC | 0.1810658 | 0 0.1777066 | - |
| Kappa | 0.3493 | 0.3523 | - |
| Time taken to run model | 59.04148 mins | 1.749515 mins | - |
| Time taken to test model | 2.44164 secs | 3.49179 secs | - |
| **Both (Using SMOTE)** | | | |
| Accuracy | 0.7981 | 0.8015 | 0.8031 |
| Precision | 0.4888976 | 0.4903796 | 0.4986237 |
| Recall | 0.6358077 | 0.4955328 | 0.6461695 |
| F-measure | 0.5527579 | 0.5623806 | 0.5628884 |
| AUC | 0.1805615 | 0.1771185 | 0.1723112 |
| Kappa | 0.4252 | 0.437 | 0.4385 |
| Time taken to train model | 2.594229 mins | 5.116106 mins | 24.19785 mins |
| Time taken to test model | 2.489735 secs | 4.694327 secs | 18.1919 secs |

### 

**Figure 1**: Random Forest - ROC curve for ntree = 50, mtry = 13(dotted black curve) . ntree=100, mtry = 13 (red curve), ntree=500,mtry=13(green curve) for oversampled dataset.



**Figure 2**: Random Forest - ROC curve for ntree = 50, mtry = 13(dotted black curve) . ntree=100, mtry = 13 (red curve) for undersampled dataset.



**Figure 3**: Random Forest - ROC curve for ntree = 50, mtry = 13(dotted black curve) . ntree=100, mtry = 13 (red curve), ntree=500,mtry=13(green curve) for both under and oversampled dataset.

**Github reference**: Lines 53:485 in firstModel.R

### Logistic regression using Chi Square Feature selection

Logistic regression model was performed using the glm() function. 10 fold cross validation was used with the function A baseline model was run to select the best features depending on the significant p values. The selected features were then included in another final logistic regression model which was evaluated. Over sampling and both(over and under sampling) were used to address the imbalance and same method of feature selection was applied to both. Under sampling was not applied due to its low performance in random forest model.

***Table 3:*** *Features selected based on p values in first model using oversampled dataset:*

|  |  |
| --- | --- |
| MENTHLTH\*  X\_MENT14D\*  DECIDE\*  POORHLTH\*  PHYSHLTH\*  GENHLTH\*  X\_LMTSCL1  X\_PHYS14D\*  EMPLOY1\*  DIFFALON\*  X\_LMTACT1\*  X\_LMTWRK1\*  HAVARTH3\*  INCOME2\*  X\_SMOKER3\*  X\_ECIGSTS\*  CHCCOPD1\*  ECIGARET\*  ASTHMA3  MEDCOST\*  X\_CASTHM1  MARITAL\*  X\_BMI5\*  HIVTST6\*  SEX\*  BLIND\*  RENTHOM1\*  CADULT\*  FC60\_\*  X\_BMI5CAT\*  X\_PACAT1\*  TOLDHI2  FVGREEN1  HTIN4\*  GRENDA1\_ | HTM4  EXRACT11\*  FRUIT2  DIABETE3  EXRACT21\*  FRUTDA2\_\*  ACTIN11\_  X\_STATE\*  CHCKIDNY  HIVRISK5\*  PAFREQ1\_  greenIntake\*  X\_FRTLT1A  X\_AGE65YR\*  CVDCRHD4  X\_RACE\*  EXERHMM1  fruitIntake  PADUR1\_  X\_IMPRACE\*  WTKG3  DEAF\*  X\_PRACE1\*  PERSDOC2\*  vegIntake  PREDIAB1  X\_RACEGR3  FRENCHF1  X\_DRNKWEK  PNEUVAC3\*  DRNKANY5\*  DROCDY3\_\*  VEGEDA2\_ |

### \*Feature appears in all models which may signify importance

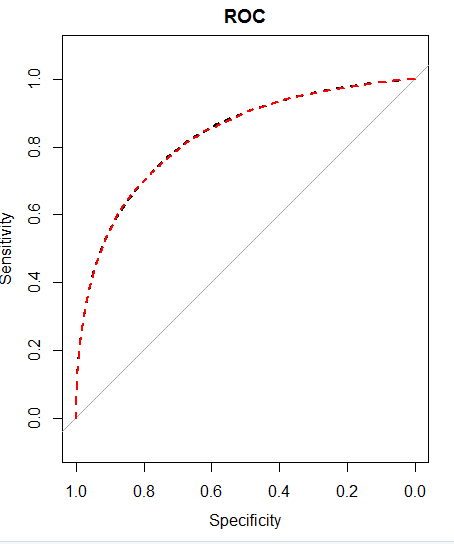
#### **Table** **4**: Features selected based on p values in second model using both over and under sampled dataset:

|  |  |
| --- | --- |
| MENTHLTH\*  X\_MENT14D\*  DECIDE\*  POORHLTH\*  PHYSHLTH\*  GENHLTH\*  X\_PHYS14D\*  EMPLOY1\*  DIFFALON\*  X\_LMTACT1\*  X\_LMTWRK1\*  HAVARTH3\*  INCOME2\*  X\_SMOKER3\*  X\_ECIGSTS\*  CHCCOPD1\*  ECIGARET\*  MEDCOST\*  MARITAL\*  X\_BMI5\*  SMOKE100  HIVTST6\*  SEX\*  BLIND\*  RENTHOM1\*  CADULT\*  FC60\_\*  MAXVO2\_  X\_BMI5CAT\* | X\_PACAT1\*  X\_AGEG5YR  HTIN4\*  EXRACT11\*  EXRACT21\*  FRUTDA2\_\*  EDUCA  X\_STATE\*  BPHIGH4  HIVRISK5\*  greenIntake\*  X\_AGE\_G  X\_AGE65YR\*  X\_RACE\*  X\_IMPRACE\*  POTATOE1  X\_MRACE1  DEAF\*  X\_PRACE1\*  PERSDOC2\*  PREDIAB1  X\_RACEGR3  FRENCHF1  PNEUVAC3\*  DRNKANY5\*  DROCDY3\_\*  X\_DUALUSE |

### \*Feature appears in all models which may signify importance

#### **Table 5:** Results for all logistic models using oversampled and both sampling techniques

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Oversampling** | | **Both (Over and Undersampling)** | | |
| ***Accuracy*** | 0.7757 | | ***Accuracy*** | 0.7743 |
| ***Precision*** | 0.4540788 | | ***Precision*** | 0.4519805 |
| ***Recall*** | 0.706302 | | ***Recall*** | 0.7074911 |
| ***F-measure*** | 0.5527785 | | ***F-measure*** | 0.5515826 |
| ***AUC*** | 0.1704202 | | ***AUC*** | 0.1712385 |
| ***Kappa*** | 0.4124 | | ***Kappa*** | 0.4144 |
| ***Misclassification error*** | 0.2242667 | | ***Misclassification error*** | 0.2257333 |
| ***Time taken to train model*** | 56.81835 mins | | ***Time taken to train model*** | 36.82462 mins |
| ***Time taken to test model*** | 13.82866 secs | | ***Time taken to test model*** | 16.24051 secs |



**Figure 4**: Logistic Regression - ROC curve for first regression model(black curve) and second regression model(red curve)

**Github Reference** – Lines 489:639 in dataCleaning.R

#### Summary of results: Random forests were the most accurate models for this analysis. Some limitations of the random forests were that more often than not they were not as effective at identifying a true depression diagnosis(low recall scores), but they were more accurate at identifying a true non-depression diagnosis. The first set of random forest model using oversampling led to the most reliable results of all models, but there was no relief from the low precision and fairly low recall scores. It is interesting to note that undersampling led to some high true positives or recall scores, however accuracy suffered greatly when compared to the other models.

# Conclusions

There were some predictors which were intuitively significant across all models such as alcohol intake, diet, and pre-existing health conditions (denoted by ‘\*’). Other factors such as HIV risk and medical cost also proved to be significant. Perhaps, another iteration of this analysis can be run using these predictors to explore a more wholesome contribution to depression diagnosis and prevention methods. Also, given that random forests led to the most consistent scores, more iterations by adding more trees and tweaking the mtry parameter could be explored further for a more comprehensive list of predictors.

# Citations

1 Darst, Burcu F., Malecki, Kristen C., Engelman, Corinne D.(2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics*; London Vol. 19(Suppl 1). DOI:10.1186/s12863-018-0633-8

2 Jiarpakdee, J., Tantithamthavorn, C., & Treude, C. (2018). AutoSpearman: Automatically Mitigating Correlated Software Metrics for Interpreting Defect Models. *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. doi:10.1109/icsme.2018.00018

3 Khalilia1, M., Chakraborty2, S., & Popescu3, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*. Retrieved from <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-11-51>.

4 Wray, T., Luo X., Ke, J., Perez, A., Carr, D., Monti, P.(2019). Using Smartphone Survey Data and Machine Learning to Identify Situation and Contextual Risk Factors for HIV Risk Behavior Among Men Who Have Sex with Men Who are Not on PrEP. *Society for Prevention Research.* doi:tps://doi.org/h10.1007/s11121-019-01019-z