Classifying Recidivism using LSI, ASUS, and Machine Learning

Gavin Sugita[1], Lipyeow Lim[1], Julie Takishima-Lacasa[2], Michael Endres[2]
Department of Information & Computer Sciences[1], Social Science Research Institute[2]
University of Hawaii Manoa

Abstract

Recidivism is the likelihood that an individual that has been released from incarceration will commit second crime that will lead to their arrest. The PSD (The Department of Public Safety) for the state of Hawaii spends millions of tax payer dollars each year to reform inmates and monitor parolees. Parole officers and case managers are responsible for the monitoring and recording of parolees and probations but the limited workers and funding often overburdens each public servant with too many cases. This can lead to burn out and the decrease in overall productivity. Often the public servant is forced to decide how to allocate their time most efficiently. The goal then becomes to help parole officers and case managers make an informed decision of who may require more monitoring by classifying released individuals on their likely hood to recidivate. Currently the standard for classification in the field of psychology is the LOGIT model. This paper proposes alternative solutions and compares predictive recidivism accuracy with known assisted learning algorithms to the LOGIT model. With the use of cross validation and an operation implementation, the random forest in particular is found to predict recidivism about 8% better than the LOGIT model which may allow the case workers of PSD to better select which individuals require more assistance.

Keywords AI, Assisted Learning, Recidivism, LSI, ASUS

1 Introduction

Upon release from incarceration, each individual has a certain chance to commit another crime and return to jail or prison. This is called recidivism and is defined as the tendency for a convict/criminal to reoffend after release back into the public. Predicting recidivism is an important tool to protecting the public from repeat offenders as well as determining the most efficient way to allocating limited government resources like manpower. Often overburdened caseworkers who are given too many active cases are forced to make a decision on which parolees require more attention than others.

Currently risk assessment tools like Level of Service Inventory (LSI) and statistical models like the LOGIT model can be used to estimate the risk level a released individual presents and focus the limited manpower on individuals that need it. These models have found limited success by observing and classifying recidivism on individual features. Upon increasing the number of features during testing, the current models are determined to be inadequate in give definitive conclusions.

This paper will explore the use of the assisted learning algorithms, SVM, decisions tree, random forest, and ANN in comparison to the LOGIT model on the risk assessment tool, LSI, and substance abuse survey, ASUS, datasets to predict recidivism rates and increase the efficiency of caseworkers.

Evaluation of each algorithm will be based on cross validation and operational results over recidivism rates over 1 and 3 year time spans. The cross validation will be created using the LSI and ASUS reports from the PSD by randomly dividing the reports into 4 even quadrants. Each slice will be used as part of the teaching dataset 3 times and the testing set once through an exhaustive rotation of all 4 slices. Every fourth test the reports will again be randomly divided into 4 new even quadrants for the next cross validation. Operational results reflect a single sampling based on time and how this system could actively operate. In this case the most recent quarter of the reports were used as a test case while all others were used to teach the algorithms. This is as close to a practical use scenario that could be accomplished given this data. The year intervals of 1 and 3 were chosen due to the significance within PSD reports accepted durations to measure recidivism.

Of the 4 assisted learning algorithms proposed the decision tree, random forest, and ANN were found to achieve a higher accuracy prediction rate for recidivism on the LSI than the control group of the LOGIT model. The linear SVM performed very similar to the LOGIT model and achieved very similar results. The accuracy of algorithms using the ASUS reports was homogenous and performed universally poorly with accuracy results around 50%. The LSI was found to be a stronger predictor of recidivism than the ASUS. Accuracy trends generally decreased from the 1 year to the 3 year recidivism prediction.

The random forest algorithm, using the LSI reports, achieved the highest performing accuracy in both the cross validation and the operational test. The average accuracy was between 58% - 61% . Predicting human behavior is a difficult task due to the multitude of variables. In addition the accuracy of the reports may also play a part as some individuals may not always be truthful during the administration of the tests or human error occurs when data is input.

This paper will start with a related work summary of current methods to predict recidivism followed by background information about the SVM, decision trees, random forest, and ANN. This will go into a brief overview of what these machine learning methods are and how they work. The method section will outline the way the experiment was carried out to insure proper data cleaning, implementation of algorithms, cross validation, and operational testing. Pre-processing goes over in detail the decisions made to clean the data and which features were available and which features were selected for training. In the

results section the findings are displayed in a general accuracy format then a confusion matrix as each quadrant can mean different outcomes for the efficiency of workforce and safety of the public. A conclusion and future work is added to show what was accomplished through this work and what possible solutions to problems faced as well as limitations for these methods. Finally the references include background information on current recidivism prediction papers as well as links to the sci kit library that was used to create the learning machines.

2 Related Work

There are many works into the field or recidivism prediction. The paper "The prediction of criminal recidivism" by Frank Urbaniok 2006 [8] used LOGIT models to classify recidivism rates of sex offenders and violent crimes. Their method simplified the search space into single features like age and nationality and were able to classify on each group to have a relative recidivism percentage. The main purpose was to classify a notable difference between sex offenders and violent offenders and which features like age or foster care will affect the recidivism rate. They were able to definitively classify a difference between sex offenders and violent crime but found the individual feature classification inconclusive.

In The paper "Using Logistic Regression Modeling to Predict Sexual Recidivism" by Grant Duwe [9] uses the classification tool Minnesota Sex Offender Classification Tool (MnSOST) as feature space selection to run a LOGIT model to classify high-risk sex offenders. This paper's main goal is to test the effectiveness of the MnSOST tool. They determined that the tool was able to predict recidivism with around 79.6% accuracy within a small portion of the population (2,315 sex offenders) and cross validation n = 220. The sample group is much more specialized and smaller than the population this paper will use.

These two papers found success when using the LOGIT model with sex offender subset of the prison population. Sex offenders are shown to have one of the lowest recidivism rates at around 13%.[11] This population is relatively homogeneous with a few easy to detect outliers. In the prediction of criminal recidivism the results show a larger variation in violent crimes when compared to sex offenders.[8] In using logistic regression modeling to predict sexual recidivism the model was able to pick out only extreme cases as possible recidivism candidate while most were uniformly classified to non-recidivate.[9] This paper will explore a higher dimensional search space over the general and drug using prison populations. This will give a much more diverse sampling of the whole prison population.

In PSD there are two rick factor assessments reports used to classify inmate risk level. The first is a well-known assessment, the LSI, which is used throughout the nation. The second is an assessment based upon the history of the inmate during their incarceration. The recidivism rates are calculated in a joint effort between probation, parole, and PSD.[4] These reports are able to show the history and detailed trends of recidivism within the population. These reports show the recidivism history based on features like county, age, gender, ect. Unfortunately, there are very few predictive metrics and mostly report on current findings and trends extrapolated from past data.

The paper "Short- and long-term recidivism prediction of the PCL-R and the effects of age: A 24-year follow-up" by Mark Olver [10] uses another classification metric called the Hare Psychopathy Checklist as the search space to classify recidivism. The invers correlation between the age and recidivism is the main conclusion of this study. Other contributed factors such as violence and follow up time were also explored though findings were inconclusive.

In general, the current works within this field focus on statistical metrics to classify recidivism based on individual features. These have all found promising results but are unable to rectify the problem incorporating these results into a single concrete probability. When the feature space became too large by adding additional information, the conclusions became unclear. This shows the limits of current statistical analysis pertaining to recidivism.

3 Background

3.1 Learning Algorithms

Learning algorithms are tools used to classify large amounts of data into groups based on specific features. These methods are especially powerful when there is a high dimensional search space and inferences into the raw data are difficult. There are two major classifications for learning algorithms, assisted and non-assisted learning methods. Non-assisted learning algorithms, like clustering, are used to identify patterns within large datasets without any known knowledge of the features of the data. These methods are good to find relevant information about data that looks completely random. Assisted learning algorithms are given both independent and dependent data. The independent data will consist of all variables/features that are in the dataset. The dependent data is the target goal for the classification. Since the data provided from PSD consists of the recidivism rates (target goal) and dates, assisted learning algorithms were selected to predict recidivism rates in this study.

3.2 Assisted learning algorithms

The assisted learning methods used to predict recidivism in this study include support vector machine (SVM), decision tree, random forest, a simple neural network, and the LOGIT model as the control metric. Visual representations of the models can be found in figure 1. The initial support vector machine used a linear support vector classifier. A linear support vector machine attempts to classify items in the training set in accordance to its target with the maximum margin or difference between the classification groups. The ground truth, also known as the target, is the known outcomes of the training set. In this experiment the target is the recidivism data. By separating the recidivating and non-recidivating individuals based on the feature selection the SVM is able to determine what features within the variable list are strong indicators of recidivism. The linear model is then tested on a test

case, a group of samples that the SVM does not know the outcome, to determine the predictive value of that model. The decision tree is also based on the classifier model. The decision tree model determines which of the features in the training set best separates the data set based on the ground truth. Each subset is then subjected to the same process until an optimum solution is achieved. This method of assisted learning allows for diagrams to show users which characteristics are most important to determining the ground truth. The trained model is then subject to a test case in which the model does not know the ground truth, the model predictions are compared to the ground truth of the test case to determine the predictive capabilities of the model. The random forest uses the decision tree model to create multiple decision trees; the mode of the population is taken as the representative of the population. This method is used to prevent overfitting. The neural network architecture is built on one hidden layer with 4 neurons. Artificial neural networks use three types of layers, an input layer, a hidden layer or layers, and an output layer. The input layer is the training/test cases and the output layer is either the ground truth or the predictive value. The hidden layer is where weights or the values of the input are changed to better map to the output layer. The assisted learning methods are part of the python package scikit and the R programming language. Multiple assisted learning methods were selected to test imperially which would have the highest predictive value. Implementations on both python and R allowed for double checking of methods.

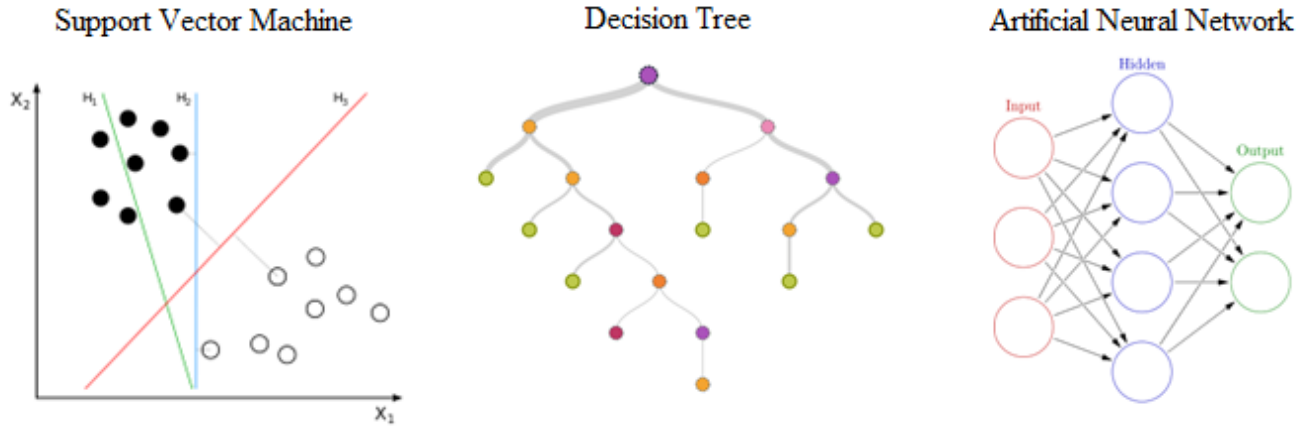


Figure 1:Visual representations of a Support Vector Machine, Decision Tree, and Artificial Neural Network

4 Method

4.1 Objective

The LSI (Level of Service Inventory) and the ASUS (Adult Substance Use Survey) are two questionnaires that assist corrections facilities to determine risk levels for inmates and drug treatment plans respectively. These questionnaires will be used as our feature lists to determine recidivism. Recidivism prediction over 1 and 3 years are the target goals for this study. Accuracy of these predictions will be scored into four fields. The algorithm predicted recidivism and the person did recidivate in the time duration, a true positive answer. The algorithm predicted recidivism and the person did not recidivate, a false positive answer. The algorithm did not predict recidivism and the person recidivated in the time duration, a false negative answer. The algorithm did not predict recidivism and the person did not recidivate, a true negative answer. The confusion table was used due to the nature of the implications each of these groups can mean for real life problems. People that are flagged for recidivism for example, can receive more observation once released or on parole. They could be subject to more frequent drug testing or required to go to more alcoholics anonymous meetings. This level of observation can lead to preventative measures before a crime is committed. In addition this increased observation also comes with an increase cost of manpower and time, resources that are finite. True positives and true negatives can properly focus resources on people that are more prone to recidivating. This increases safety for the general public while saving tax payer money. The false positive grouping represents a waste in resources as these people were marked for recidivating but did not in the time duration. As long as this number isn't too large or expensive, a certain amount of these cases are accounted for when determining department funding. The last group of false negative are the group of individuals that were classified as not recidivating but actually committing a crime within the time duration. This group represents a danger to the general public. Unlike the false positive group in which a small population is acceptable, the false negative group should have a population of zero as each person in this group was able to have less monitoring to commit another crime.

4.2 Recidivism Data

The data that is used in this study is provided by the Hawaii Department of Public Safety (PSD). There are 4 data tables that hold the information that we used to train and test the assisted learning algorithms to predict recidivism rates. Initially the data from PSD is not fit to train assisted learning machines. The cleaning process and final product for each of the data tables is detailed in the Data Processing section. (Section 5)

4.3 LSI Data

The cleaned LSI was randomly split into quarters. The ¾ slice was used for training while the ¼ slice was used for the testing set. The final cleaned LSI consisted of 31,813 entries. When split,

this gave 7,953 entries in the test set and 23,860 entries for the training set. Each training and testing set were rotated to get 4 tests for each random sampling. Sampling was done at random for every 4th trial. This processing method allows for cross validation for each sample.

4.4 ASUS Data

The cleaned ASUS was taken and randomly split into quarters. The ¾ slice was used for training while the ¼ slice was used for the testing set. The clean completed ASUS data consists of 8003 unique entries which were divided into a 6003 entry training set and a 2000 entry testing set. For cross validation purposes each training and testing set were rotated to get 4 tests for each random grouping. Sampling was done at random for every 4th trial. This processing method allows for cross validation for each sample.

4.5 Training

The training and testing for the LSI and ASUS were done separately for years 1 and 3 recidivism rates. The SVM, decision tree, random forest, ANN, and LOGIT model were trained on ¾ of the data and tested on the last ¼ and checked with the ground truth to calculate correctness. The training and testing was conducted 20 times for each assisted learning machine for 1 year and 3 year recidivism rates. The random forest generated n=100 trees in order to create the mode selection.

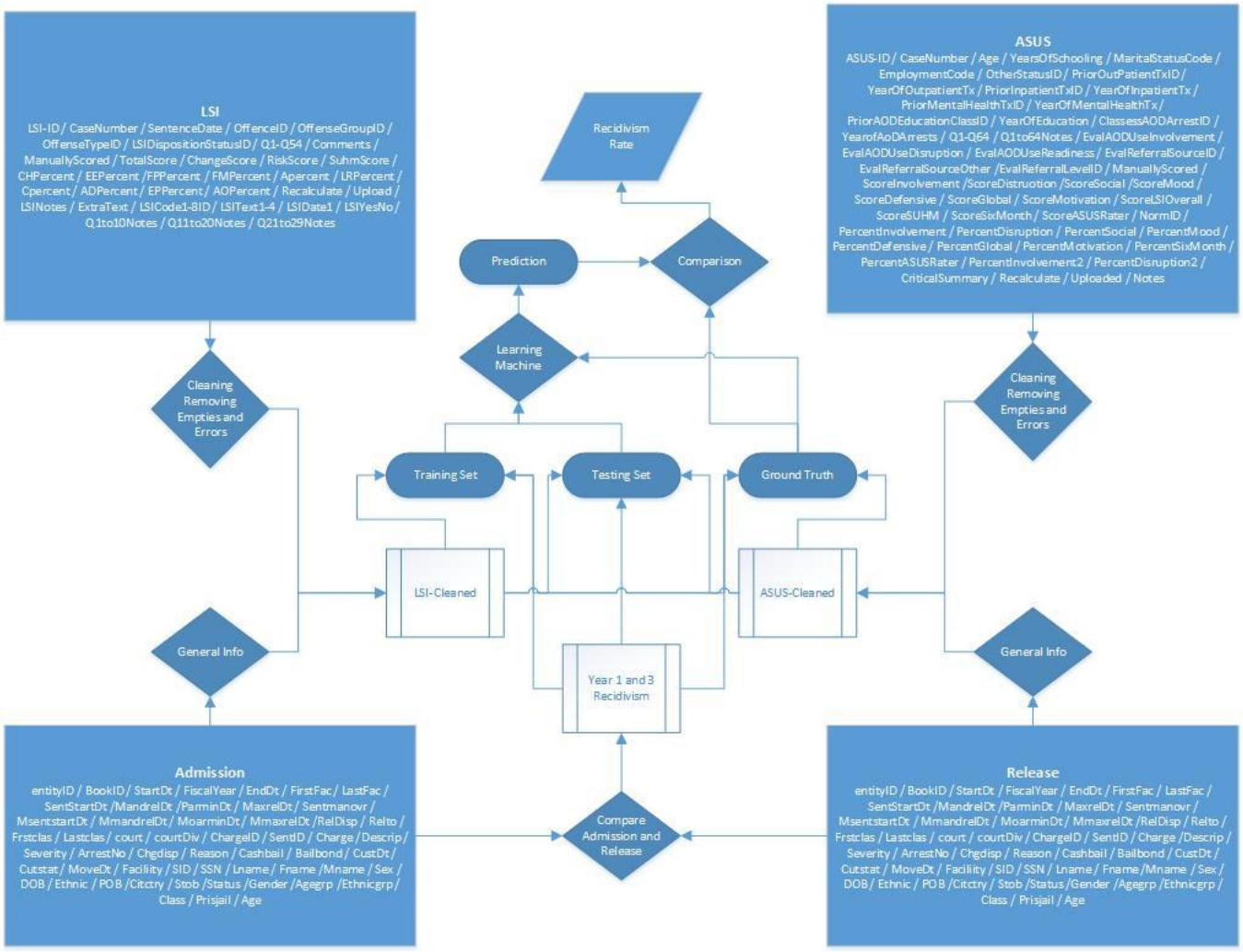


Figure 2: Flow chart of the Data Pre-processing

5 Data Pre-Processing

5.1 Data Details- Initial and End Goals

The assisted learning algorithms and models are unable to process the raw data received from PSD. The formatting of the data and removal or erroneous entries are outlined visually in figure 2. The detailed explanation of cleaning methods and meaning of the features within each dataset is listed below.

5.1.1 LSI Cleaning

The LSI consists of questions ranging from education level to family ties to determine the risk level of an inmate. This risk level is used to help predict how likely an inmate will respond poorly with a problem. For this reason the LSI scores have an impact on determining if an inmate will be accepted by the parole board. The LSI is mandatory for all incoming and outgoing inmates but can also be given after a violent incident or to determine the status of

a parolee. This has only been the case since 2013. Prior to 2013 the LSI was only given if the incarceration facility or an agency requested it. For this reason the number of recent LSI reports within the last 4 years makes up a large portion of the total LSI reports. The LSI table given from PSD consists of 109 items and 106,314 entries. Figure 3 shows the 109 elements of the LSI report.

LSI Content

LSI_AssessmentID			CaseNumber		SentenceDate		OffenseID		OffenseGroupID		
OffenseTypeID		LSI_DispositionStatusID			Q1		Q1b	Q2	Q3	Q4	Q4b
Q5	Q6	Q7	Q8	Q8b	Q9	Q10	Q11	Q12	Q13	Q14	Q15
Q16	Q17	QHomePens		Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25
Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36	Q37
Q38	Q39	Q40	Q40b	Q41	Q42	Q43	Q44	Q45	Q45b	Q46	Q47
Q48	Q49	Q50	Q50b	Q51	Q52	Q53	Q54	Comments			
ManuallyScored			TotalScore		ChangeScore		RiskScore		SuhmScore		
CHPercent		EEPercent		FPercent		FMPercent		APercent		LRPercent	
CPercent		ADPercent		EPPercent		AOPercent		Recalculate		Uploaded	
LSI_Notes		ExtraText		LSI_Code1ID		LSI_Code2ID		LSI_Code3ID			
LSI_Code4ID		LSI_Code5ID		LSI_Code6ID		LSI_Code7ID		LSI_Code8ID		LSI_Text1	
LSI_Text2		LSI_Text3		LSI_Text4		LSI_Date1		LSI_YesNo1			
Q1To10Notes		Q11To20Notes		Q21To29Notes							

Figure 3: The features/information within the LSI

5.1.2 LSI Details

- The LSI_AssessmentID (LSI-ID) is a unique identifier for each LSI given. This is unique to each test and therefore unique to a time, place, and individual. Since an inmate can take the LSI multiple times, one inmate can have multiple LSI-ID that corresponds to each test taken. These identifiers is always in the format of "DOCH-LSI-#", where # is the number of the LSI. Currently there are roughly 100,000 cases and each new case is incremented by 1. Though the entries count is 106,314 there are some cases that were tests or duplicates.
- Case Number is given to a depending on which facility gives the LSI and who proctors it. This number is unique to the test but the format will differ depending on the facility. Example of A5011248 maybe a case number in facility 1 but 09-1-0857 will be the case number given in facility 2 while yet another test may not have a case number as it was done as a follow up through parole.
- SentenceDate is the date at which the inmate was incarcerated. In this field there are many blanks as some tests were administered before incarceration and others were completed while on parole.
- OffenseID and OffenseGroupID are fields in which additional information about the inmate offence can be noted. These two areas are free note areas and include comments like which gang affiliations and gang rank the inmate may have.
- OffenceTypeID is a numerical ID to show what type of crime the person was sentenced for, example of a 3 can represent robbery 1 while a 10 can be assault 3.
- LSI_DispositionStatusID is the general status of the inmate while taking the test. This shows if the inmate is cooperative or not.
- Q1 – Q54 are the individual answers to each question given within the LSI. Many are yes or no questions but some do have multiple choice. Questions like Q1b and Q45b allow for clarification. For example if Q45 asked have you done drugs, and the inmate responds with a yes, then Q45b may list a number of commonly used drugs like methamphetamine or marijuana. These questions maybe left blank if they would answer no to initial question.
- Comments are the notes left from the proctors of the test. This field is mostly blank with some comments like LSI Reassessment completed.
- ManuallyScored is a true of false question to show if the LSI results were calculated by hand or where done by a computer.
- TotalScore is the summed score from all the LSI. This is what will be used to determine the risk level of the inmate and what score the inmate will receive.
- RiskScore is the risk level determined by the TotalScore. The RiskScore is grouped into 3 categories, high, medium, and low. This score will help to determine where the inmate will be housed and at what security level. High RiskScore means the inmate is more prone to causing disturbances then a low RiskScore.
- SuhmScore is the recommended treatment level based only on the LSI. This includes drug treatment and counselling to help reform the inmate. The next few items are percent scores that help to determine the status of the inmate for a given portion of the LSI. Each score

is given as a fraction and goes from 0-1. The percent scores are independent of one another.

- CHPercent is the criminal history percent score. This represents if the person is a repeat offender or first offence and what the crimes were. A theft 1 will be weighted less than a murder 1.
- EEPercent is the education/employment percent score. This represents level of education and if they inmate had a job at the time of assessment.
- FPercent is the financial percent score. This represents the financial security the inmate has at the time of assessment.
- FMPercent is the family/marital percent score. This represents the familial support, the closeness to relatives and spouse, the inmate has at the time of the assessment.
- APercent is the accommodation percent score. This represents if the inmate has adequate housing at the time of assessment.
- LRPercent is the leisure/recreation percent score. This represents how much they enjoy their life.
- CPercent is the companions percent score. This score is similar to the family score but focuses on friends and others in the inmate's life.
- ADPercent is the alcohol/drug percent score. This represents the alcohol and drug addictions that the inmate may have.
- EPPercent is the emotional/personal percent score. This represents the emotional wellbeing of the inmate and helps to determine any psychological ailments like chronic depression, anxiety, and thoughts of suicide.
- AOPercent is the attitude/orientation percent score. This represents how the inmate views the world, does the inmate believe the world is hostile and unfair towards him/her.
- Recalculate is to determine if the scores were recalculated. This was usually done to insure accuracy of the calculations done.
- Uploaded is to determine if the LSI was uploaded form a secondary facility. This is false for every LSI administered.
- LSI_Notes is another section where the proctor can add additional notes like parole dates and additional criminal background.
- ExtraText is another note space if the proctors need to make a second note or notes become too long.
- LSI_Code1ID, LSI_Code2ID, LSI_Code5ID, LSI_Code6ID, LSI_Code7ID, and LSI_Code8ID does not currently have data within the field nor a key to show what data should be there.
- LSI_Code3ID represents the agency that is conducting the LSI. This is filled with a numerical value representing the group.
- LSI_Code4ID represents the county at which the LSI was given.
- LSI_Text1, LSI_Text2, and LSI_Text3 does not currently hold any data.
- LSI_Date1 does not currently hold any data. The date of the LSI given will be pulled from another source table.
- LSI_YesNo1 holds only false, documentation on this question is missing.
- Q1To10Notes, Q11To20Notes, and Q21To29Notes does not currently hold any data.

Additional information on the questions of the LSI will be listed in the appendix.

5.1.3 ASUS Cleaning

The ASUS test is specifically tailored to determine the drug treatment programs an inmate will receive in within the incarceration facility or while on parole. Unlike the LSI which is given to each inmate the ASUS is a specialized questionnaire that only targets people with drug/alcohol addiction problems. This means there are fewer individuals with an ASUS than a LSI. The ASUS table initially contains 164 items and 86,946 entries. There exist duplicate entries and fields of missing data. Figure 4 shows the elements of the ASUS report.

ASUS Content

ASUS_AssessmentID	CaseNumber	Age	YearsOfSchooling	MaritalStatusCode						
EmploymentCode	OtherStatusID	PriorOutpatientTxID	YearOfOutpatientTx							
PriorInpatientTxID	YearOfInpatientTx	PriorMentalHealthTxID	YearOfMentalHealthTx							
PriorAODEducationClassID	YearOfEducationClasses	AODArrestsID								
YearOfAODArrests	Q1	A1	Q1b	Q2	A2	Q2b	Q3	A3	Q3b	
Q4	A4	Q4b	Q5	A5	Q5b	Q6	A6	Q6b	Q7	A7	Q7b
Q8	A8	Q8b	Q9	A9	Q9b	Q10	A10	Q10b	Q11	Q12	Q12b
Q13	Q13b	Q14	Q14b	Q15	Q15b	Q16	Q16b	Q17	Q17b	Q18	Q18b
Q19	Q19b	Q20	Q20b	Q21	Q21b	Q22	Q22b	Q23	Q23b	Q24	Q24b
Q25	Q25b	Q26	Q26b	Q27	Q27b	Q28	Q28b	Q29	Q29b	Q30	Q30b
Q31	Q31b	Q32	Q33	Q34	Q35	Q36	Q37	Q38	Q39	Q40	Q41
Q42	Q43	Q44	Q45	Q46	Q47	Q48	Q49	Q50	Q51	Q52	Q53
Q54	Q55	Q56	Q57	Q58	Q59	Q60	Q61	Q62	Q63	Q64	
Q1To10Notes	Q11Notes	Q12To21Notes	Q22To31Notes	Q32To36Notes	Q37To41Notes						
Q42To50Notes	Q51To57Notes	Q58To64Notes	EvalAODUseInvolvement								
EvalAODUseDisruption	EvalAODUseReadiness	EvalReferralSourceID									
EvalReferralSourceOther	EvalReferralLevelID	ManuallyScored									
CriticalSummary	Recalculate	Uploaded	Notes								

Figure 4: The features/information within the ASUS

5.1.4 ASUS Details

- ASUS_AssessmentID (ASUS-ID) is a unique identifier of each ASUS test. This ID is unique to the test and therefore unique to the time, place, and individual it was given to. An individual inmate can take the ASUS multiple times to determine if there is improvement. This can lead to a single inmate having multiple ASUS-ID associated with him/her.
- CaseNumber is also a unique identifier for an ASUS test but is dependent on the facility and proctor who administered the test. If the test was administered through parole there is a possibility that a case number does not exist for that given ASUS. This leads to high levels of variability within this element. Example, facility 1 has a case number of cr00-1-2321 while facility 2 has a case number of 1/1/2484.
- Age is the age of the individual at the time of the assessment.
- YearsOfSchooling is the number of years of education or equivalent that the inmate has completed. If the inmate receives their GED then they have a completion of 12 years or a high school education.
- MaritalStatusCode is the marital status of the inmate, if they are single, married, divorced, widowed, or separated.
- EmploymentCode is if the inmate is employed at the time of assessment. This contained the information of unemployed, student, unemployed for 1-3 months, or unemployed for more than 3 months.
- OtherStatusID shows additional employment options such as student, retired, disabled, or homemaker.
- PriorOutpatientTxID is the reference to any prior outpatient drug rehabilitation treatments.
- YearOfOutpatientTx is the last year of outpatient drug treatment if any.
- PriorInpatientTxID is the reference to any prior inpatient drug rehabilitation treatments.
- YearOfInpatientTx is the last year of inpatient drug treatment if any.
- PriorMentalHealthTxID is the reference to prior mental health rehabilitation treatment.
- YearOfMentalHealthTx is the year of the last mental health treatment.
- PriorAODEducationClassID is the reference to prior alcohol rehabilitation treatments.
- YearOfEducationClasses is the year of the last alcohol rehabilitation treatment.
- AODArrestsID is the reference to the number of drug or alcohol related arrests.
- YearOfAODArrests the year of the last drug or alcohol related arrest.
- Q1-Q64 are the questions of the ASUS. These include in-depth questions of what type of drugs and/or alcohol the inmate has taken and how recently.
- Q1To10Notes, Q11Notes, Q12To21Notes, Q22To31Notes, Q32To36Notes, Q37To41Notes, Q42To50Notes, Q51To57Notes, and Q58To64Notes include comments and notes from the proctor for further clarification or out of place behavior.
- EvalAODUseInvolvement is a numerical score to classify the involvement the use of alcohol or drugs. The scale ranges from 0-9 where 0 is minimal and 9 is high.

- EvalAODUseDisruption is a numerical score to classify the disruption the use of alcohol or drugs has on the inmate's life. The scale ranges from 0-9 where 0 is minimal and 9 is high.
- EvalAODUseReadiness is the score of how readily available alcohol and drugs are for this inmate. This scale was recently changed from a numerical representation to a low, medium, and high scale.
- EvalReferralSourceID and EvalReferralSourceOther are the referral information that was not given.
- EvalReferralLevelID is the evaluators assessment and recommendation for the inmate, a numerical value represents a recommendation is given. For example 2 is a comprehensive assessment while 8 will be weekly therapy.
- ManuallyScored shows if the scores were calculated by hand or by computer.
- CriticalSummary notes for a summary.
- Recalculate is the scores were recalculated.
- Uploaded if it was an uploaded from offsite.
- Notes does not currently hold any data.

5.1.5 Admission and Release Cleaning

The admissions and release data holds information of the incarceration dates, facilities, and general information about each inmate. The comparisons of these two data sets will give the recidivism timelines for inmates. The two data tables hold very similar data and therefor can be explained in a single section. Figure 5 shows the elements of the admissions and release reports.

Figure 5: The features/information within the Admissions and Release dataset

5.1.6 Admission and Release Details

- Entityid and bookid is the unique identifier for the admissions and release data tables. Each one represents either an admissions entry or release entry.
- Startdt is the date at which person was scheduled to either be incarcerated or released.
- Fiscal Year is the fiscal year in which the start of the processing occurred.
- Enddt is the date in which the release or incarceration took place.
- Firstfac is the first facility that the inmate was incarcerated in.
- lastfac is the last facility that the inmate was incarcerated in.
- Sentstartdt, mandreldt, parmindt, maxreldt, and sentmanovr are remnants of the hard copies. These sections are duplicate sections or have been removed from recent reports.
- Msentstartdt is the date in which the incarceration time begins.
- Mmandreldt is the time required to be served.
- Mparmindt is the time until the inmate can be possibly paroled.
- Mmaxreldt is the maximum time the system can hold an inmate for their crime. These are all listed as dates at which the action will have to take place.
- Reldisp is the current stage of processing in which the inmate is subject to. This includes bail, probation, and incarceration.
- Relto shows any change in stage of processing that occurred during the admissions or release.
- Frstclas is the security risk assigned to the inmate at the start to the admissions or release process.
- Lastclas is the security rick assigned to the inmate at the completion of the admissions or release process.
- Court is the court that sentenced the inmate.
- Courtdiv is the court district that sentenced the inmate. Most court and courtdiv fields are empty, usually only filled out if the case became a federal case.
- Charged and sentid are secondary unique identifies for the admissions and release data set. Charge is the type of offence the inmate is charged with. The format of ###-#### is used where the first 3 numbers are the offence and the last 4 numbers are the severity, example of 708-0832 is for the crime of theft 3.
- Descrip is the description of the offence and lists the actual offence name, example of theft 3.
- Arrestno and caseno are used as unique identifiers for the judicial branch.
- Chgdisp is if the action of admissions or release has been completed. Reason serves as a notation area for why an admissions or release was not completed on time or before time.
- Cashbail and bailbond are the fields in which the inmate was given a bail of a specific amount during trial.
- Custdt is the date in which the inmate was apprehended.
- Facility is the facility that houses the inmate at the start of sentencing.

- Sid is the unique identifier give for each inmate upon their first arrest. The SID will remain with that individual for life.
- Ssn is the social security number of the inmate.
- Lname is the last name of the inmate.
- Fname is the first name of the inmate.
- Mname is the middle name or middle initials of the inmate.
- Sex is the physical sex of the inmate.
- Dob is the date of birth of the inmate.
- Ethnic is the ethnicity that best describes the inmate.
- Pob is the country of birth.
- Citctry is the country that the crime was committed in.
- Stob is the state of birth, in the cases of non-American places of birth this is left blank.
- Status is the current status location of the inmate, example 2_SFP is for parole.
- Gender is the gender that the inmate associates with. There are cases where sex and gender do not match as these are considered transgendered inmates.
- Agegrp is the age grouping the inmate falls into. They are listed from A to K where A are ages from 18-19 and K are ages greater than 65.
- Ethnicgrp is the group classification of the ethnicity of the inmate.
- Class is the classification of the inmate's offence.
- Prisjail is if the inmate resides in a jail or in a prison. Jails are typically for mild offences and for inmates with less than 1 year left in their sentence. Prisons are for higher security inmates and for long term incarcerations.
- Age is the calculated age of the inmate at the time of the admissions or release.

5.1.7 Data Goals- LSI and ASUS

The merged data will require the item level questions that the LSI and ASUS provide with the unique identifiers for each inmate within the study. The differences between the merged LSI and merged ASUS was the LSI-ID/ASUS-ID, the questions, and the scoring methods. The information of each inmate is the same between both data sets. Figure 6 shows the elements of the final report that will be used to train and test the algorithms.

Goal - LSI & ASUS

SID	LSI-ID	fname	mname	lname	sname	date	reassessment	sentencing		maritalstatus		
address		currentstatus		firstfac	lastfac	bailbond		gender	agegrp	ethnicgrp		
prisjail	age	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24
Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36	Q37
Q38	Q39	Q40	Q41	Q42	Q43	Q44	Q45	Q46	Q47	Q48	Q49	Q50
Q51	Q52	Q53	Q54	TotalScore		ChangeScore		RiskScore		SuhScore		
CHPercent		EEPercent		Fpercent		FMPercent		Apercent		LRPercent		
Cpercent		ADPercent		EPPercent		AOPercent						

Figure 6:The features/information within the final dataset that will be used to train the learning algorithms

5.1.8 Data Goal Details

- SID is the unique identifier for an inmate given at the first incarceration and remains with the inmate for life.
- LIS-ID or the ASUS-ID is the unique identifier for each individual assessment taken.
- Fname is the first name
- Mname is the middle name
- Lname is the last name
- Sname is the surname.
- Date is the date the assessment was administered.
- Reassessment is if this was their first assessment for a repeat assessment.
- Sentencing is the branch that sentenced the inmate.
- Maritalstatus is the marital status of the inmate.
- Address is the last known address of the inmate; this is later used to determine homelessness.
- Currentstatus is the status of the person, if they are an inmate, parole, ect.
- Firstfac is the first facility the inmate was housed in.
- Lastfac is the last facility to house the inmate.
- Bailbond is the merge of bailbond and cashbail.
- Gender is the gender the inmate associates with.
- Agegrp is the age group the inmate falls into.
- Ethnicgrp is the ethnic group the inmate falls into.
- Prisjail determines if the housing was in a prison or a jail.
- Age is the actual age of the inmate at the time of the assessment. The rest are the questions and the calculations that were taken directly form the LSI and ASUS.

5.2 Merging & Assumptions

5.2.1 LSI & ASUS Cleaning

The LSI and ASUS data tables require cleaning and merging to get the information into a format that assisted learning machines can use. In order to do this the LSI and ASUS entries were checked for errors and duplicates. Errors were accounted for by checking missing input and input outside of acceptable parameters. The percent error within each entry was calculated, if the percent error was greater than 20% it was removed, else all

missing information or erroneous information was zeroed. Clarification questions were omitted from this error checking. Example of the clarification question, what type of drug did you most commonly use will only be filled in if the inmate previously said they do drugs. Duplication checking was done by comparing the LSI and ASUS scores within a specific SID. This gave all the tests and scores for a specific individual. The first thing tested is the date of the test. These tests are time consuming and will not usually be done multiple times a day. If there are multiple tests for the same individual on the same day a double check is initiated that will check the answers to all the questions. If the information of the two files are identical then the first test, the one with a lower LSI-ID number is discarded. If the information is close, only a difference in 5 questions or less, this is determined to be an error that the proctor was attempting to fix but did so by adding another test case. In this case the first test is discarded. If the differences in the test questions were greater than 5 then there is a larger errors in the data and both entries are discarded.

5.2.2 LSI & ASUS Linking to Identification

Four linking files that contain the SID to name, LSI-ID to name, SID to name, and ASUS-ID to name were used to match the tests with the SID of each inmate. Since names were used as a linking methods between SID to LSI-ID and SID to ASUS-ID fuzziness of 2 errors were allowed within the sum of the first and last name. The middle initial was required to be exactly correct after punctuations and whitespaces are removed. This was required since exact matching of names would not work due to white spaces, caps, or punctuation after the middle initial or surname. This gave use the temporal map of the inmate testing records. The SID was used with the admissions and release data tables to gain the general information about each person. This included name, date of birth, and address.

5.2.3 Admissions & Release – Recidivism List

The years or recidivism were calculated by the final dates as these were the times the action of admissions and release take place. The difference between the closest recorded admissions after a release is calculated and the release entry is segregated into a table for recidivism under 1 year, under 3 years, or did not recidivate within 3 years. The closest LSI and/or ASUS test that precedes the release date is taken and added to the LSI/ASUS 1 year, LSI/ASUS 3 year, or LSI/ASUS 3+ year recidivism tables. If the person did not recidivate they are added to the 3+ year recidivism table. If the person did not have a release prior to admissions they are assumed to be a first time offender even if an LSI may have been recorded prior to the admissions date, this will stop the guessing into which recidivism group to place the person. This projects overall goal is to improve jail/prisons then only the LSI and ASUS scores that were obtained within incarcerations will be relevant. LSI and ASUS can be administered during parole but these entries are disregarded as not pertinent information to this study.

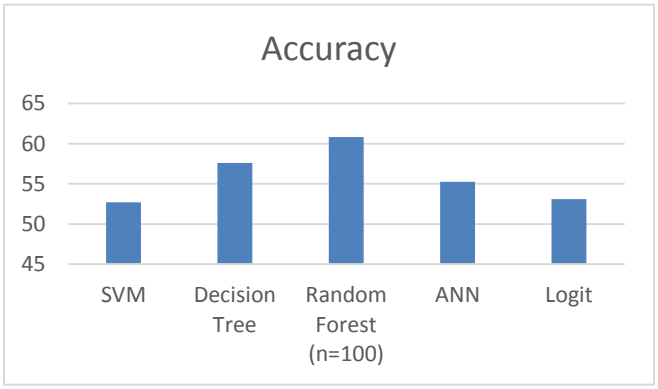
6 Results

6.1 LSI Cross Validation Results

The LSI data showed a slight predictive capability when the learning algorithms of SVM, decision tree, random forest, and ANN are used. The SVM performed the weakest of the 4 in both the 1 year and 3 year recidivism data with 52.705% and 51.25% accuracy respectively. The linear model tended to do well in some case but very poorly in others. Since the training data was selected randomly this caused some of the predictions to fall under 50% accuracy. The ANN performed with 55.23% accuracy for 1 year recidivism prediction and 53.78% accuracy for 3 year recidivism prediction. The decision tree performed with 57.61% accuracy for 1 year recidivism prediction and 55.17% accuracy for 3 year recidivism prediction. The most accurate predictive model was the random forest which had 60.835% accuracy for 1 year recidivism and 59.765% accuracy for 3 year recidivism.

There is a decrease in predictive accuracy between the 1 year and 3 year recidivism population. This occurred over all assisted learning machines and is believed to happen due to the possible changes that may have occurred since the last LSI upon release. There is an increased chance that the individual stopped attending AA meetings or lost their job and now slipped back into old criminal habits. Unfortunately any LSIs that are proctored during parole fall outside of the scope of the PSD data. These LSI reports can give additional snap shots of the wellbeing of the individual after release and can be incorporated in a later study.

The year 1 and 3 accuracy shows the random forest algorithm achieving the highest general accuracy at 60.835% and 59.765% respectively, though the decision tree was second with an accuracy of 57.61% and 55.17% respectively. This alludes to overfitting as one of the problems for this dataset. This is shown in figure 7 and 8 for the general accuracy and error of 1 year recidivism and 3 year recidivism respectively. The algorithms of decision tree and ANN also achieved a higher average accuracy then the LOGIT model. The linear SVM is known to have similar performance to the LOGIT model which was observed in these tests.



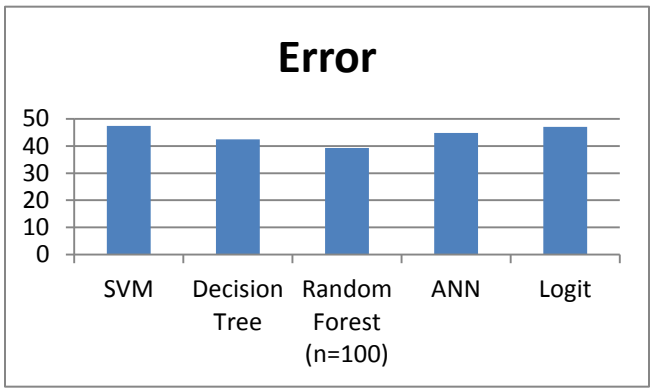
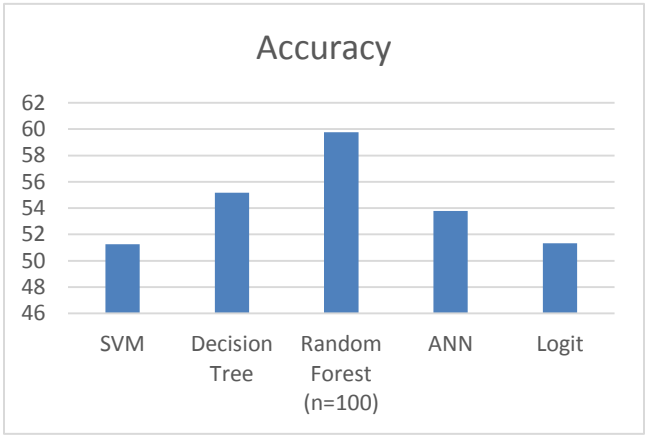


Figure 7: LSI year 1 Recidivism Accuracy – General (Accuracy and Error are computed in percentage)n is the number of trees generated



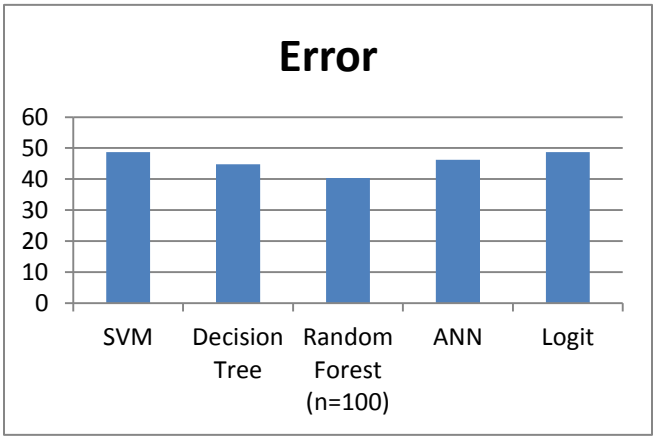


Figure 8: LSI year 3Recidivism Accuracy – General(Accuracy and Error are computed in percentage) n is the number of trees generated

The confusion matrix shows the true positive (predicted to recidivate and recidivate), true negative (predicted to not recidivate and do not recidivate), false positive (predicted to recidivate and do not recidivate), and false negative (predicted to not recidivate and recidivate) to determine the consequences for each decision. This is depicted in figure 9 and 10 for 1 year recidivism and 3 year recidivism respectively. The year 1 breakdown shows that the random forest was better in all aspects of the confusion matrix but showed the most difference in determining true positives and false negative. The overall similar accuracy of the false positives and the low false negatives shows that the random forest was move conservative, meaning a larger group was classified to recidivate then necessary. This is completely acceptable as the false positives are accepted to happen and are accounted for in practice. False negatives are not acceptable as these are possibility preventable crimes that were allowed pass the algorithm.

Figure 9: LSI year 1 Recidivism Accuracy – Confusion Matrix(Accuracy and Error are computed in percentage) n is the number of trees generated

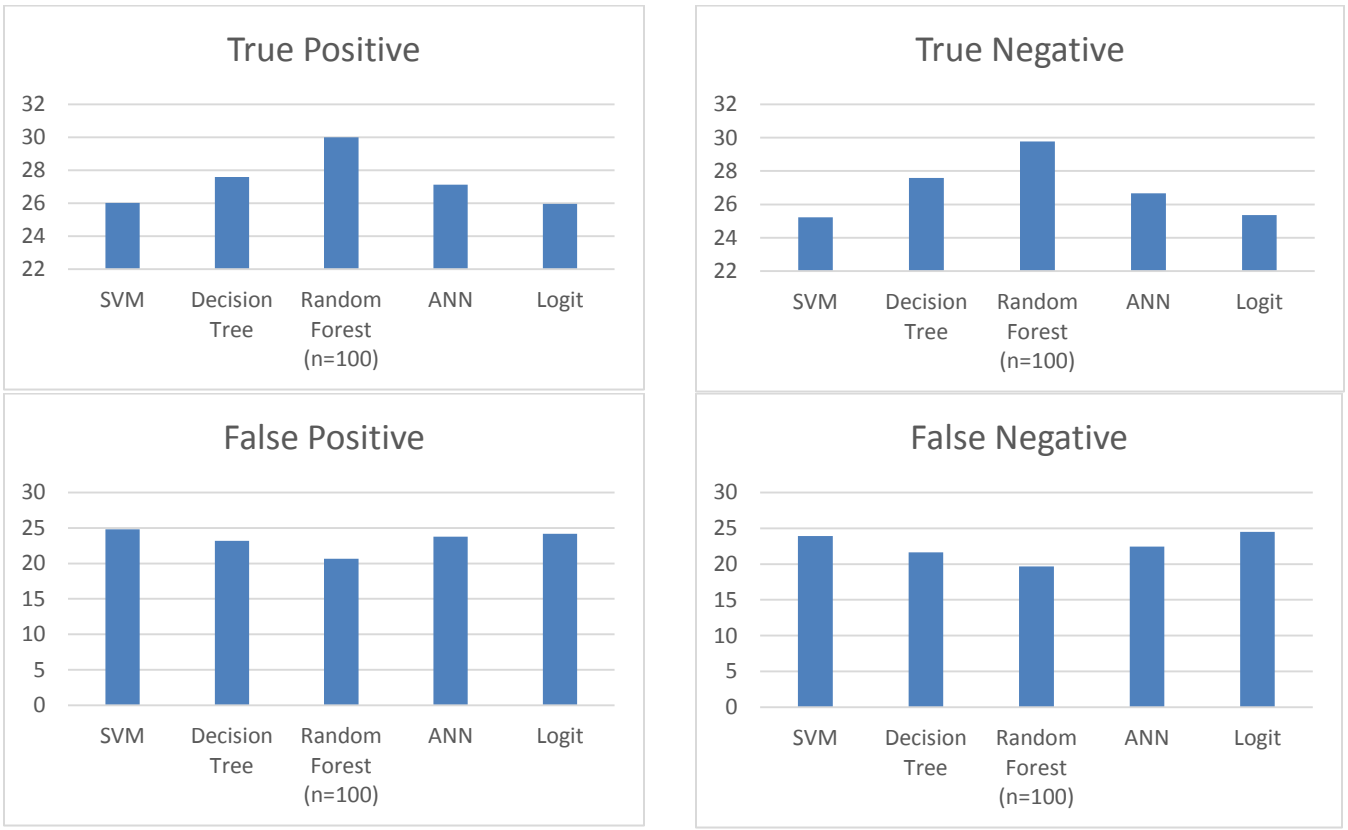


Figure 10: LSI year 3Recidivism Accuracy - Confusion Matrix (Accuracy and Error are computed in percentage) n is the number of trees generated

6.2 LSI Operational Data & Results

The operational results are based on the real world results that can be derived from a temporal sampling of the data. The LSI dataset was sorted by date administered then the test case was selected as the last 7,953 items within the dataset. The earlier entries were used as the dataset to teach the algorithms. This method of selection was chosen as any strict temporal selection would imbalance the dataset due to an increase in administration of LSI over recent years. By selecting the last 7,953 items the operational results would result in a testing and learning set size that is very similar to the cross validation runs and thus most comparable.

The operational results of the assisted learning algorithms preformed slightly less accurate then the cross validation average. One possible reason is that the most resent LSI questionnaires have not have the opportunity to recidivate yet. In some of the most recent cases the prediction says they will recidivate within 1 or 3 years but that duration has yet to be completed. For theses the individual has not recidivated but has only been out of incarceration for a few months. The accuracy of these predictions can increase as time and sampling continues.

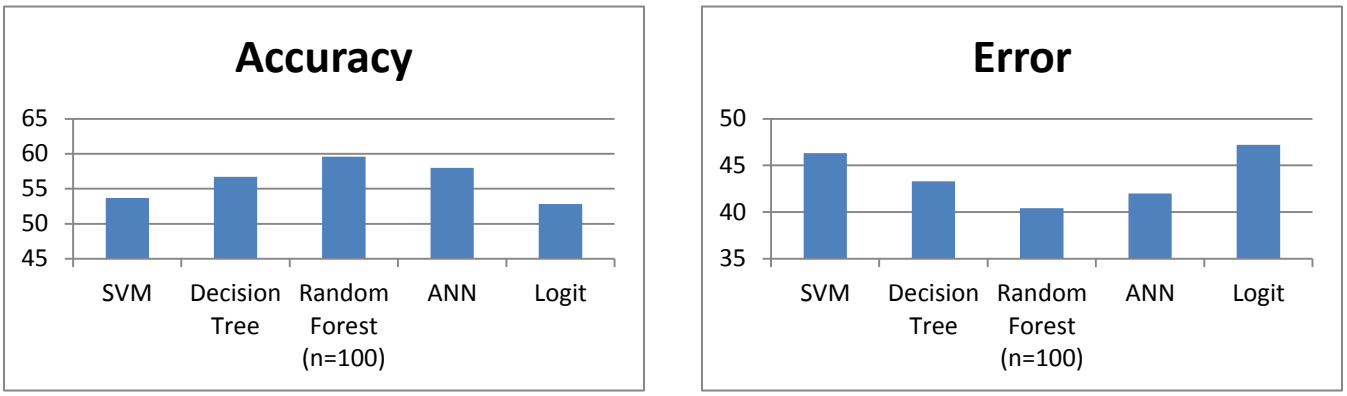


Figure 11: Year 1 LSI Operational Accuracy – General (Accuracy and Error are computed in percentage) n is the number of trees generated

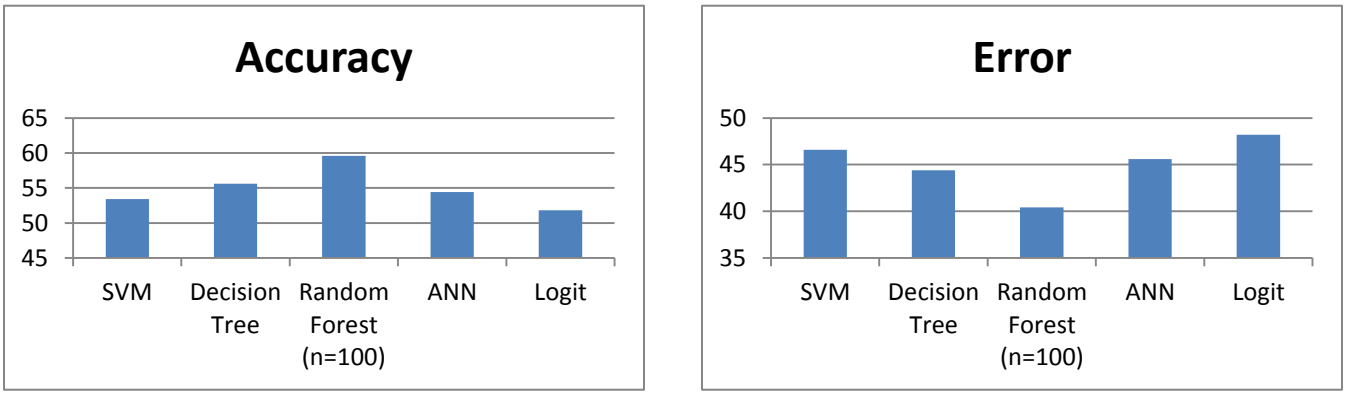
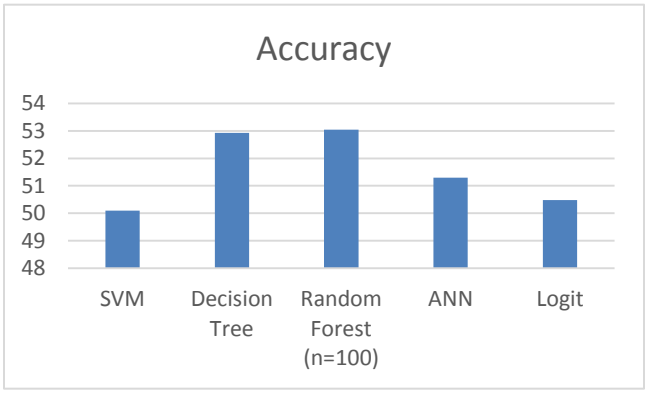


Figure 12: Year 3 LSI Operational Accuracy – General (Accuracy and Error are computed in percentage) n is the number of trees generated

6.3 ASUS Cross Validation Results

The decision tree and random forest were the highest accuracy recidivism predictors. This shows that overfitting does not tend to decrease overall accuracy of the classification. The predictive power of the ASUS is very limited with predictive accuracy for all algorithms only a max of 1.5% over complete random selection. The decision tree, random forest, and ANN were able to achieve higher predictive results then the Logit model. The linear SVM is the only assisted learning algorithm that constantly underperformed the Logit model.

The overall accuracy was very similar for all methods. The random forest seems to do slightly better than other methods but only achieves 53.05% and 51.223% accuracy for years 1 and 3 respectively. This can be seen in figure 13 and 14 for 1 year and 3 year recidivism.



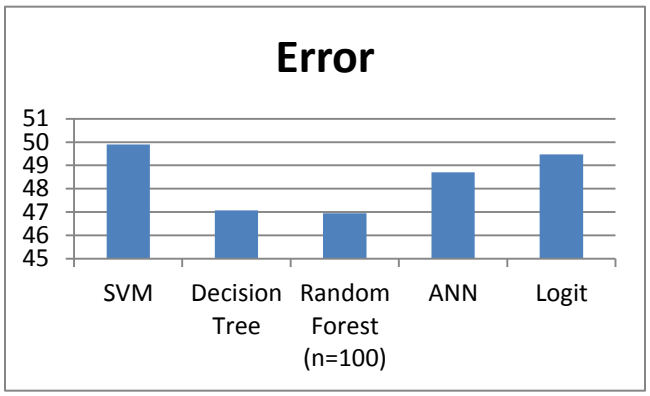
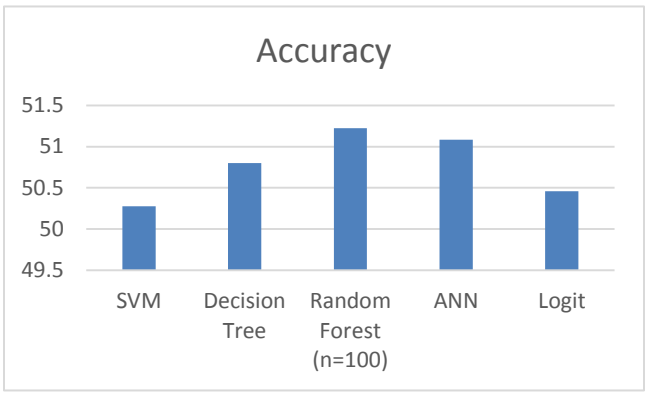


Figure 13: ASUS year 1 Recidivism Accuracy – General (Accuracy and Error are computed in percentage) n is the number of trees generated



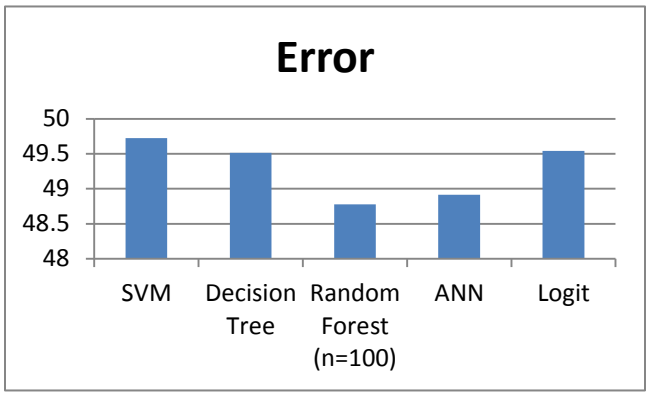


Figure 14: ASUS year 3 Recidivism Accuracy – General (Accuracy and Error are computed in percentage) n is the number of trees generated

The confusion matrix shows the true positive (predicted to recidivate and recidivate), true negative (predicted to not recidivate and do not recidivate), false positive (predicted to recidivate and do not recidivate), and false negative (predicted to not recidivate and recidivate) to determine the consequences for each decision. The decision tree and random forest have achieved slightly better accuracy then the other methods but the overall difference between all methods was only 1-2 percentages. The ASUS is primarily used to determine the level of treatment for drug and alcohol abuse individuals. The ASUS maybe one of many factors that determine recidivism but alone it does not appear to be a strong indicator of recidivism. The accuracy for all quadrants of the confusion matrix show very little difference in accuracy. This can be seen in figures 15 and 16 for the confusion matrix of the 1 year and 3 year recidivism. There may have some information that can help slightly select for recidivism but the indicators are just not strong enough in the ASUS. This could be attributed to people being less truthful about their drug habits or since the sampling size is smaller than the LSI proper patterns were unable to immerge.

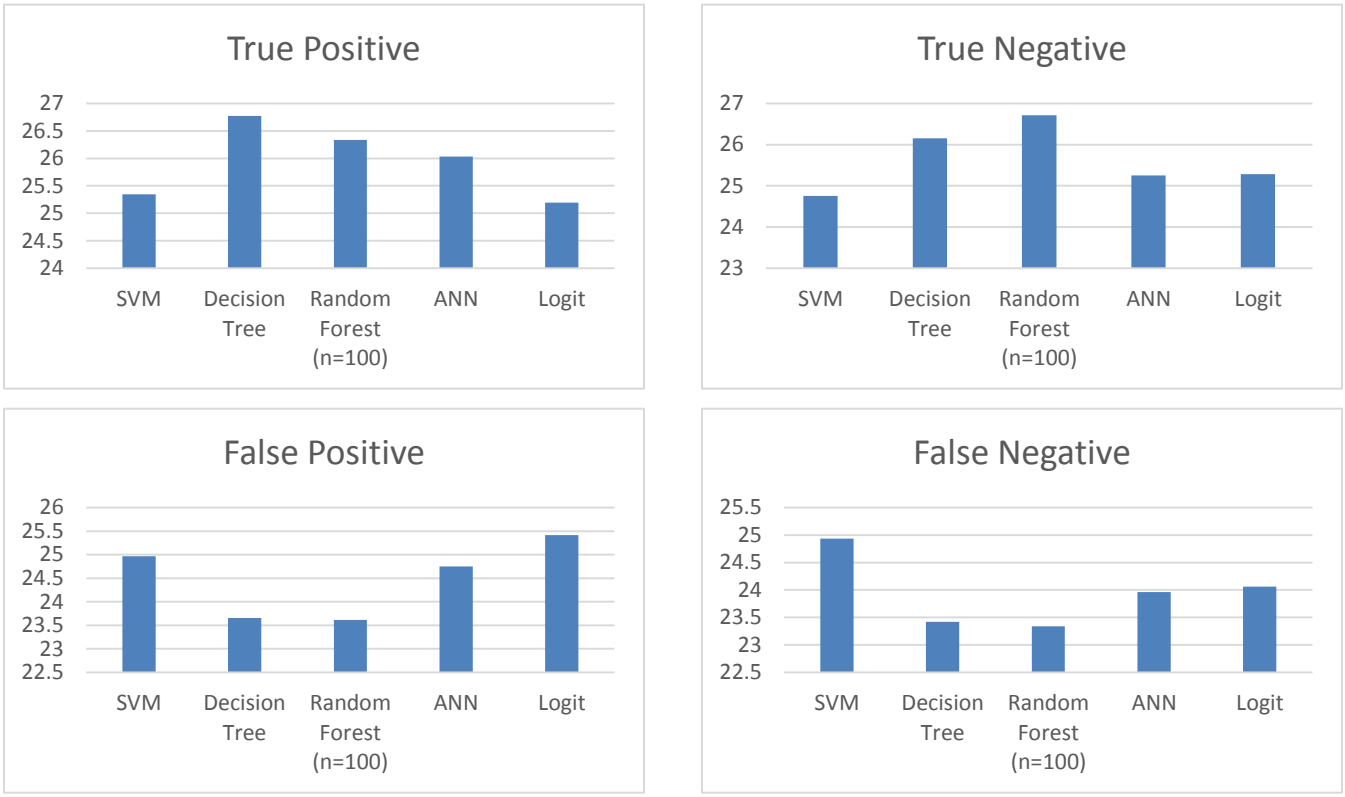
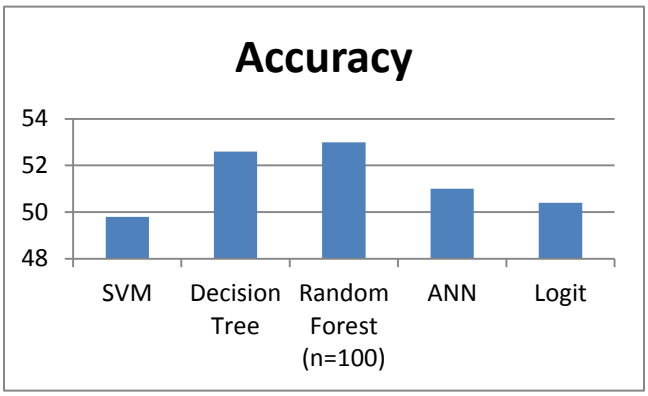


Figure 15: ASUS year 1 Recidivism Accuracy – Confusion Matrix (Accuracy and Error are computed in percentage) n is the number of trees generated

Figure 16: ASUS year 3 Recidivism Accuracy – Confusion Matrix (Accuracy and Error are computed in percentage) n is the number of trees generated

6.4 ASUS Operational Data & Results

In the operational run for the ASUS sample is similar to the LSI sample as the overall behavior is about what is found in the cross validation average but slightly lower. This again could be accounted for by some recently released inmates not having the time needed to recidivate. Unlike the LSI the ASUS in general had a lower recidivism prediction accuracy which bordered random chance. This can be seen in figures 17 and 18 for 1 year and 3 year operational recidivism.



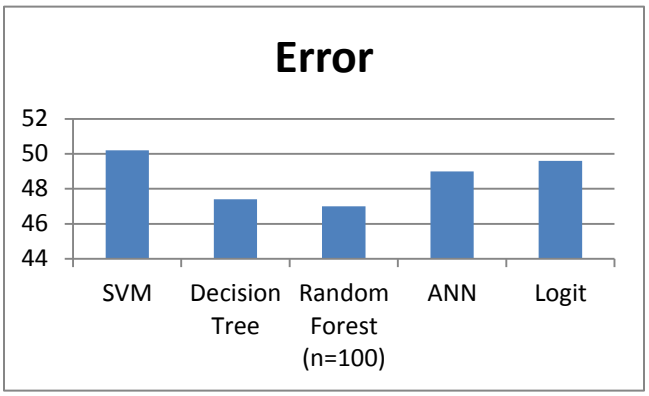
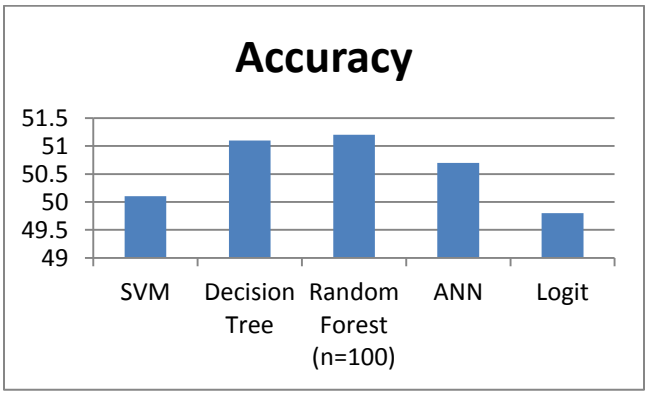


Figure 17: Year 1 ASUS Operational Accuracy - General(Accuracy and Error are computed in percentage) n is the number of trees generated



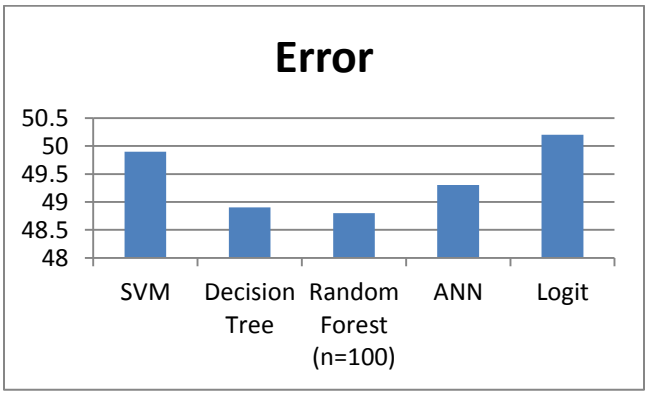


Figure 18: Year 3 ASUS Operational Accuracy – General (Accuracy and Error are computed in percentage) n is the number of trees generated

6.5 Overall Results

In computer science these percentages may appear very low and boarders on random chance. In the field of psychology prediction of human behavior has been close to impossible. Human behavior is made up of countless variables from learned experiences like actions and consequence to physical limitations like housing and employment. In psychology it is widely accepted that human behavior is extremely complicated and therefor predictive analysis tend to result in be somewhat inaccurate. The accepted form of predictive analysis in the psychology field is the logit model, logistic regression. The logit model is a probability model not necessarily a classifier so it is limited to how accurate the model can be as a behavioral predictor.

For the operational runs for both the LSI and ASUS, the run is based on a temporal division of the data and only run one time. This allows us to show a real world example of how the algorithms are able to perform but is bias in its results since only a single sampling was taken.

7 Conclusion

The average cross validation and operational run of each of the assisted learning algorithms showed the same or better predictive rates then the standardized LOGIT model, though the best performing algorithm, the random forest, was only able to correctly predict recidivism 60% of the time. The ASUS is general is less effective tool to predict recidivism than the LSI. This makes sense since the ASUS is a very specialized tool that focuses only on the drug habits of the individual while the LSI is able to give a broader criminal history. The duration of time after the questionnaires also negatively impacted the predictive quality of the models. Intuitively this also makes sense as these questionnaires are snapshots of the individual at the time of the test. As people change throughout time these past questionnaires

are no longer accurate representations of the individual. This only shows how hard it is to predict human behavior.

There are also the problems with the data that should also be addressed. There were duplicate and blank entries meaning there are currently some problems with the data entry that should be rectified. This experiment also relies on the assumption that the data collected is truthful and recorded correctly. Unfortunately there are people prone to lying which can corrupt the dataset with false information.

This project represents the first steps in the use and benefits of computer learning models in the field of psychology. The human psyche if based in a highly dimensional search space with huge numbers of variables, machine learning methods are perfect for these types of problems and can achieve higher accuracy of classification then current statistical models in these cases.

8 Future Works

The LSI and ASUS were run separately but the combination of the two could yield more accurate predictive results. The addition of general information like race and gender; the inclusion of more personal information like number of immediate family members or where they were born can assist in increase accuracy of predictions and classifications. Further fine grain approach to this problem can show which features are more representative to the accuracy of the models. Lastly these models were generated on close to the default parameters using scikit library, which gives a good general performance but can use fine tuning. This process will be assessed again by a more fine grain approach to the data analytics.

9 References

[1] Anon. scikit-learn. Retrieved April 01, 2018 from http://scikit-learn.org/stable/

[2] Anon. LSI-R. Retrieved April 01, 2018 from https://www.mhs.com/MHS-Publicsafety?prodname=lsi-r

[3] Anon.Retrieved April 01, 2018 from http://aodassess.com/assessment_tools/asus/

[4] Timothy Wong. 2007. State of Hawaii, FY 2013 Cohort 2016 Recidivism Update . *Interagency Council on Intermediate Sanctions* (June 2007), 1–24.

[5] Timothy Wong. 2016. Dashboard Indicators and Trends. *Interagency Council on Intermediate Sanctions* (October 2016), 1–30.

[6] Kunal Jain, Shubham Jain, Tavish Srivastava, and Analytics Vidhya Content Team. 2018. A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python). (April 2018). Retrieved April 01, 2018 from https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

[7] Arthur Aron, Elliot J. Coups, and Elaine Aron. 2013. *Statistics for psychology*, Boston: Pearson.

[8] Frank Urbaniok, Jérôme Endrass, and Astrid Rossegger. 2007. The prediction of criminal recidivism. (April 2007). Retrieved March 2018 from https://link-springer-com.eres.library.manoa.hawaii.edu/article/10.1007/s00406-006-0678-y

[9] Grant Duwe and Pamela Freske. 2007. Using Logistic Regression Modeling to Predict Sexual Recidivism. *Sexual Abuse: A Journal of Research and Treatment*, 24 (2007), 350–377.

[10] Mark Olver, Stephen Wong. 2015. Short- and long-term recidivism prediction of the PCL-R and the effects of age: A 24-year follow-up. *Personality Disorders: Theory, Research, and Treatment* 6 (2015), 97–105.

[11] Anon. Chapter 5: Adult Sex Offender Recidivism by Roger Przybylski. Retrieved April 23, 2018 from https://www.smart.gov/SOMAPI/sec1/ch5_recidivism.html