# Identifying the Leading Risk Factors of Suicide Attempts

Courtney Stenstrom

Western Governors University

# Table of Contents

# Project Overview

## Summary

### Research Question

Using survey data from a social media platform, the goal of this research was to gain an understanding of what leads a person to make the decision to attempt suicide, specifically what social factors have the greatest influence on the decision. In the United States, suicide is the tenth leading cause of death across all ages. Globally it claims the lives of about 800,000 individuals every year.

### Project Scope

The scope of this research was to determine what factors may influence a person's decision to attempt suicide, or whether the particular social factors had a minimal influence on the decision overall.

### Solution Overview – Tools and Methodologies

#### *Tools*

The python programming language was the main source of cleaning and analyzing the data, which originated as an Excel file foreveralone.xlsx. This file contained records of individual survey responses. Additional Python libraries used for the analysis include Pandas, a well-rounded library for data analysis and Scikit-learn, an expansive Machine Learning library.

#### *Methodologies*

Multiple methodologies were used during this research and analysis. CRISP-DM was used for the overall planning. The project, analysis, and statistical methodologies are discussed in sections below.

# Project Plan

## Project Execution

### Project Plan

The overall goals and objectives for this project did not differ from the original proposal. The major goal was exploring and gaining a better understanding and knowledge about the influences that may lead someone to make the decision to end their own life. The planned method of accomplishing this was to use three different Machine Learning Algorithms to evaluate their ability of predicting whether a person would attempt suicide based on factors such as gender, age, employment, or depression[1].

### *Project Planning Methodology*

Cross Industry Standard Process for Data Mining, or CRISP-DM, was utilized for the duration of the project. These steps follow a natural cycle of initial understanding, data collection, preparation, and analysis through the end of the project evaluation and deployment. The overall structure of this methodology allows for the flexibility of Agile methodology or the structure of Waterfall methodology, depending on implementation. During this project, the task completion was structured, where one task was completed prior to the next being started. The actual methodology did not change from the original planned methodology.

[1] The full set of variables and their encoded values are in attached document

**Business Understanding:** Project goals were determined from the objectives and requirements of the project. These included

1. Determination of which factor or factors were most influential on a person's decision to attempt suicide.
2. Using the factors identified, to create a model able to predict an individual's risk of attempting suicide with an average accuracy of over 85%
3. Based on the outcome of analysis, suggest best steps for further action, for further research or program suggestions

**Data Understanding:** In this step the data structure was viewed and explored to gain a more in depth of the variables involved and to determine what cleaning and formatting issues would need to be addressed.

**Data Preparation:** This was when the data was prepared for the actual analysis. The preparation included encoding variables to numerical codes, which is necessary for analysis and the creation of three engineered variables. The steps will be discussed in further detail in a future section.

**Modeling:** The models were created during this stage, and included Decision Tree Classifier, Random Forest Classifier, and Logistic Regression.

**Evaluation:** Cross validation scoring of three folds was used to determine performance of each model. Cross validation helps determine the model's ability to predict new unseen data by withholding a portion of the training data for each fold.

**Deployment:** The last stage focused on the overall evaluation and finalization of the project as it ended.

### *Project Timeline and Milestones*

There were some timeline changes, which are summarized in the chart below.

| Milestone | Start Date | End Date | Duration | Change from Plan |
|---|---|---|---|---|
| Preprocessing and Cleaning of data | 4/25/2022 | 4/26/2022 | 2 days | + 1 day |
| Analysis of data via three ML algorithms | 4/27/2022 | 4/28/2022 | 2 days | + 1 day |
| Creation of Supporting Charts | 4/29/2022 | 4/29/2022 | 6 hours | - ½ day |
| Writing Report / Key Insights / Recommended Next Steps | 4/29/2022 | 4/30/2022 | 1 ½ day | - ½ day |

# Methodology

## Data Collection Process

### *Actual Data Selection vs. Planned Collection Process*

Due to the subject matter, the process to collect this type of data would need to consider numerous ethical, safety, health, and privacy precautions that were unrealistic due to time constraints and my own decision on whether I could be comfortable providing the assistance this high-risk population may require. The data selection process was the most difficult part of this, and many venues of information and statistics were explored until I found the dataset that was used.

The data is a preconstructed dataset composed of surveys done when the questionnaire was posted on the social media platform Reddit. The author remains unknown, but the survey data is hosted on Kaggle, a popular Data Science website and community.

### *Obstacles to Data Collection*

The obstacles in which I would have faced during data collection would have been focused on ethical concerns. As mentioned above, these include confidentially, gathering informed consent, and protocols for dealing with an at-risk population. Time constraints and confidence in my ability to successfully maneuver these with current resources did not allow for research to be conducted from the bottom up.

This led to the next obstacle, finding data that captured aspects in relationship to the decision. This is not always an easy task due to social stigma and the unfortunate reality that many who make this decision are no longer there to express the factors that led up to the decision. Many government-provided statistics are in the form of overall totals by region, or method used, which would not be conductive to the research question.

### *Unplanned Data Governance Handling*

There were no unplanned data governance issues encountered during this project. The data was collected and the dataset preconstructed, therefore I was not working with survey respondents or personally identifying information that would need to be handled with certain precautions.

Had this not been the case, or if further research is done that entails collecting new responses the issues and factors I would need to address and plan for include collecting / ensuring no collection of identifiable information (IP addresses, emails, names). If the data were to be collected it would need to be stored to where participants could be ensured of security. Informed consent would need to be gathered, as well as identifying those who were unable to give consent. There would need to be guidelines outlining confidentiality and steps to take if emergency situations are suspected.

### *Advantages and Limitations of Data Set*

The advantages of the dataset include its ability to portray the firsthand experiences of those who are going through, or have gone through, suicide attempts. The method of survey collection allowed responses from diverse backgrounds, whether graphically, culturally, or socioeconomically. The diversity this collection method allowed may be one of the greatest advantages, yet the inability to control for this brings the disadvantages of not being guaranteed a sample that represents the population.

The survey collection being anonymous and executed from a standpoint that would be more relaxed than a fully structured research study also has its advantages, namely respondent honesty, and candidacy.

The dataset also allowed for an analysis of a topic that I am extremely passionate about but would have had difficulty in executing from a time and resource availability standpoint.

Some of the biggest disadvantages of the data do go hand in hand with its advantages. Due to having no control of collection of data, and not having any ability of rapport with the respondents presents challenges. With every hope of honesty and candidacy, there comes the risk of the opposite. This is apparent with the responses that are outwardly false, for example the response mentioned in the previous proposal of the 12-year-old transgender boss who had plastic surgery already. For every response that is obviously false, one can only assume there are others that seem realistic but are not, but this can be true with any survey collection dealing with people.

## Data Extraction and Preparation Process

Data preparation was straightforward, following the steps outlined below:

1. Importing the excel file into the Juypter notebook environment, using Python and the Pandas library to do so.
2. As the Variance Inflation Factors, Chi Square, and Cramer’s statistic were calculated, it was clear there was a dependency among some of the variables, specifically gender and sexuality. The solution I originally implemented was to create a new variable that combined the two columns, with the new values now encoded.
    a. Running tests on both versions (with variables combined or separated) did not change model performance in either a negative or positive manner. I went with the variables in two distinct columns for simplicity.
3. Encoding values to be in numeric form. This is necessary for many machine learning models. Most of the values were encoded manually and are described in attached documentation. There were two notable exceptions to this, the column of lists pertaining to a respondent’s attempts at improving themselves / their situation and what help the respondent wanted / would be open to.
    a. Both columns were constructed from a list of options that were able to be checked, with as many or as few that were applicable. This created columns that were not necessarily useful for analysis in this

case, as many recorded values had one or two occurrences due to the high number of combinations available.

b. Two variables were created to replace these. The first column created was from the column ‘improve_yourself_how’ which dealt with activities the respondent attempted already to improve themselves or their life in some way. The new column is a numeric value of how many items were listed, or how many things a person tried to improve.

c. The second column followed the same format and described what the respondent would want help with from those around them. This encoded into a binary format, dependent on if they were open to help (had anything listed) or not (None).

---

Examples from the column ‘improve_yourself_how’

| **Original Value** | | **New Value** |
|---|---|---|
| None | ➔ | 0 |
| Joined a gym / go to the gym, join clubs /social clubs / meet ups, other exercise | ➔ | 3 |

Examples from the column ‘what_help_from_others’

| Original Value | | New Value |
|---|---|---|
| None | ➔ | 0 |
| Set me up with a date, wingman / wingwoman | ➔ | 1 |
| Date coaching | ➔ | 1 |

# Data Analysis Process

## Data Analysis Methods

This project utilized a predictive analytical method. The project goal was to be able to identify the likelihood of a future outcome (predicting whether someone would attempt suicide) based on the historical information available (the factors influencing survey respondents and their attempted suicide, or lack of any attempt). Through this process the following were utilized:

1. Chi Square Test of Independence
2. Cramer’s V
3. Variance Inflation Factor
4. Precision, Recall, F1, and AUC Scores
5. Cross Validation Score

## Advantages and Limitations of Tools / Techniques

The primary tools utilized for analysis included Python, and the libraries Pandas and Scikit-learn. These libraries are the reason Python was the best option for analysis, and the three in combination offer an extensive array of data analysis solutions. Python is a widely known, easily readable language but does have some limitations. These include slower performance when compared to other languages and it has a higher rate of memory consumption. These would become obstacles when dealing with large datasets and did not affect this project.

Frequency tables and chi square tests were used to test independence or dependence of variables. Cramer's V was used in conjunction with these to measure the strength of the relationship. Dependence within variables can lead to overfitting, and significantly impacts the statistical significance of the outcomes.

These tests are non-parametric tests, which analysis on data that is not normally distributed. Parametric tests require continuous data, and many of the variables are categorical. The limitation of these tests manifests as a lack of statistical power compared to their parametric counterparts. For example, the chi square test does not provide any insight into which category may be greater than or less than the other, nor does it give any indication of the degree of difference is, as Pearson's Correlation does. It strictly is saying if the variables are independent from each other.

### Decision Tree Classifier

Decision Trees require far less data preparation than other models and they can handle both numerical and categorical input. They are easy to implement and understand but may also create trees that are over complex, leading to overfitting.

### Random Forest Classifier

Advantages and disadvantages of this classifier are like those of Decision Trees. While a Decision Tree combines decisions, a random forest is combining several decision trees. Therefore, Random Forests can be slower than their single counterpart and can be significantly more complicated. What may be lost in speed is typically made up with accuracy and precision. This model also protects itself from overfitting, since each decision tree is using only a few predictors to build each tree, resulting in decorrelation that may exists otherwise.

### Logistic Regression

Logistic Regression was the third model I decided to use, due to its efficiency with training, simplicity in implementation and interpretation, and the binary outcome variable. The model's disadvantages include being prone to overfitting if the independent variables are related to each other. It also is limited in the assumption of linearity between independent variables and the dependent variable.

## Application of Analytical Methods

The following steps were taken after data cleaning and encoding were completed.

1. Functions were defined that would output the frequency counts and results of the chi square test of independence.
    a. Variable combinations were analyzed through these functions to determine what associations may exist.
2. Variance inflation factors were determined for the variables. This is a measure of how much multicollinearity may exist between variables. This in addition to the chi square tests were helpful in determining dependencies that exists which could influence certain algorithms.
3. The three Machine Learning models were initiated. The models being used are
    a. Decision Tree Classifier

   b. Random Forest Classifier

   c. Logistic Regression

4. Using the module cross_val_score from the scikit-learn library, each model was assessed. This module takes the data and splits it into *n* splits of data, also known as folds, while reserving a fold as a validation set.

   a. I used three folds for these tests and gathered the individual scores. This was repeated for precision, recall, and F1 as well as the average of these scores.

5. With a general idea of which models were performing best, I initialized the three models again, trained them on the entire training set, and tested them. Precision, recall, f1 score, as well as AUC score were calculated.

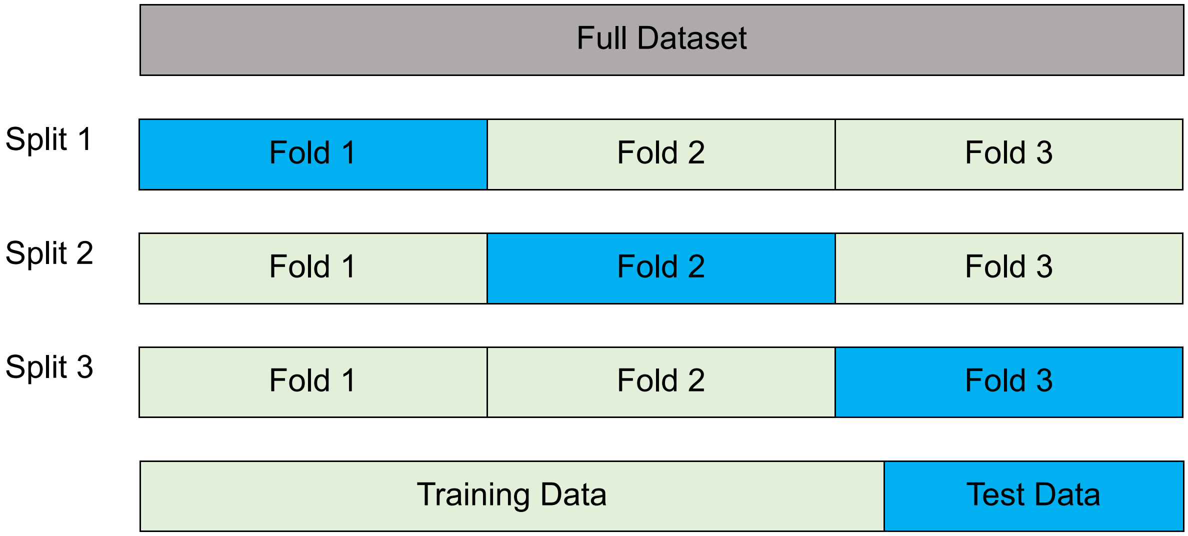
6. Selecting one of the models that was consistently performing better than the other two, I used GridSearch from the sklearn library to cycle through parameters testing for the best possible combination. I took this time to look at some other information that was interesting, such as the models feature importance scores. The top five features were friends, age, depressed, total improvements, and education.

# Results

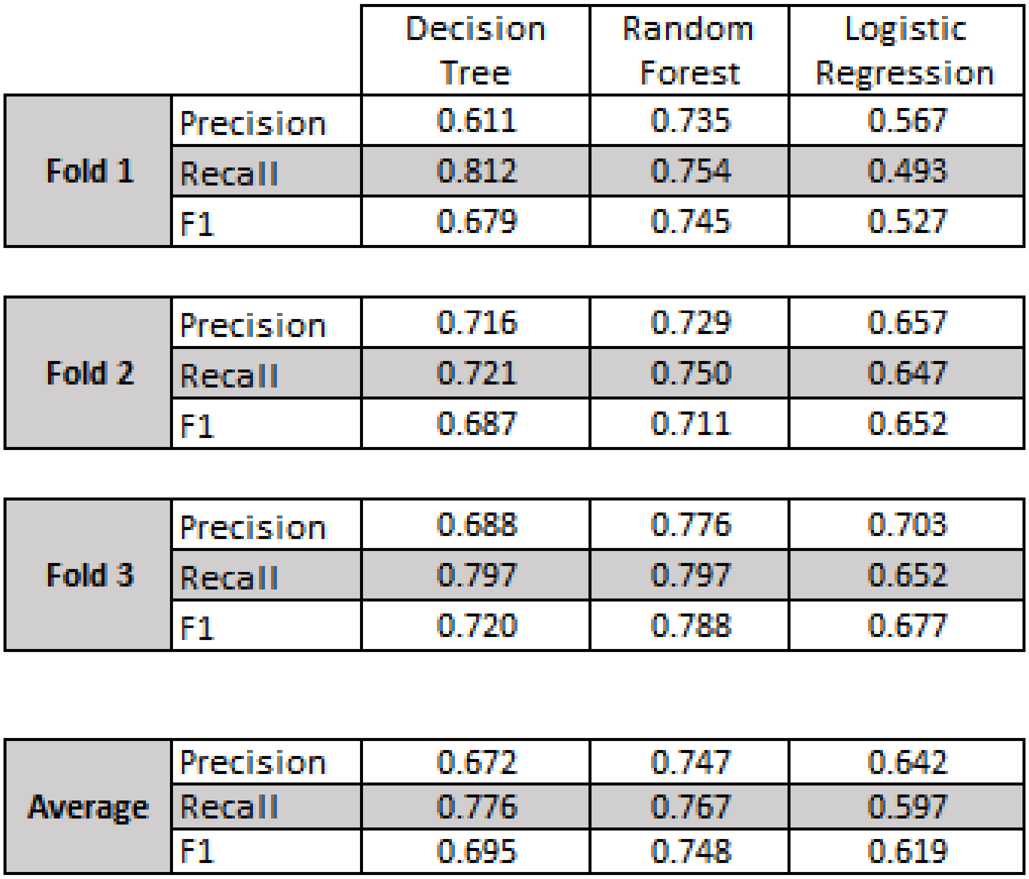
## Project Success

### *Statistical Significance*

As a method to prevent overfitting, the models were trained using cross validation scoring, separating the data into three folds. Models are trained and tested through multiple cycles, illustrated below:

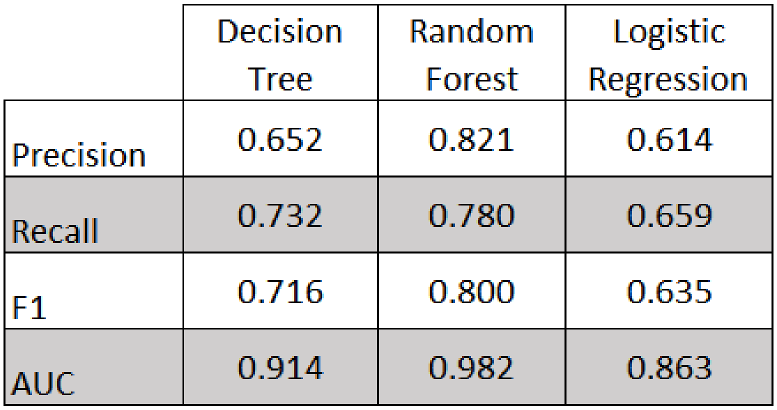


This was done for the scores of precisions, recall, and f1. The scores were averaged, giving an idea of which model would have the best overall performance.



| | | Decision Tree | Random Forest | Logistic Regression |
|---|---|---|---|---|
| **Fold 1** | Precision | 0.611 | 0.735 | 0.567 |
| | Recall | 0.812 | 0.754 | 0.493 |
| | F1 | 0.679 | 0.745 | 0.527 |
| **Fold 2** | Precision | 0.716 | 0.729 | 0.657 |
| | Recall | 0.721 | 0.750 | 0.647 |
| | F1 | 0.687 | 0.711 | 0.652 |
| **Fold 3** | Precision | 0.688 | 0.776 | 0.703 |
| | Recall | 0.797 | 0.797 | 0.652 |
| | F1 | 0.720 | 0.788 | 0.677 |
| **Average** | Precision | 0.672 | 0.747 | 0.642 |
| | Recall | 0.776 | 0.767 | 0.597 |
| | F1 | 0.695 | 0.748 | 0.619 |

New models were initialized, each was trained and tested again on data split 75% for training and 25% for testing. Performance was consistent with what was previously seen during cross validation.



| | Decision Tree | Random Forest | Logistic Regression |
|---|---|---|---|
| Precision | 0.652 | 0.821 | 0.614 |
| Recall | 0.732 | 0.780 | 0.659 |
| F1 | 0.716 | 0.800 | 0.635 |
| AUC | 0.914 | 0.982 | 0.863 |

***Practical Significance***

I find it difficult to quantify practical significance of this subject. This is a far reaching, devastating social issue and the more it is spoken about, studied, researched, and actively fought against will hold a significant effect on the issue.

***Practical Significance – Effect Size***

While statistical significance communicates whether the result is due to chance or the associated factors, effect size communicates how much impact, or how strong the relationship between those variables are.

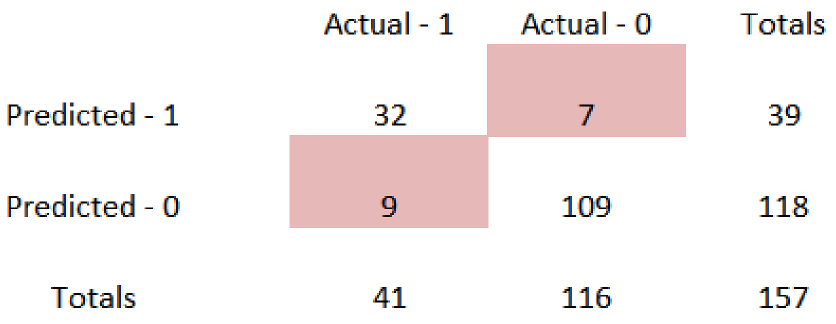
I took the predictions and test set, exported all into Excel, and examined the way the models answered against each other but also with itself.  For each model, I calcualted the phi coefficient (also known as Matthews Correlation Coefficient). This is a coorelation for binary outcomes in the range of -1 to +1, and takes into account True and False Positives as well as Negatives.

The formula for MCC is

$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative).As Davide Chicco explains in his 2017 paper, *Ten quick tips for machine learning in computational biology,* the MCC coefficient helps avoid over optomistic results from unbalanced datasets. To achieve a good score, a model is performing well in both positive and negative classifications. A score closer to 0 indicates mostly random predicition guessing.

Examing the Random Forest Classifier, the confusion matrix and associated metrics are:



| | Actual - 1 | Actual - 0 | Totals |
|---|---|---|---|
| Predicted - 1 | 32 | 7 | 39 |
| Predicted - 0 | 9 | 109 | 118 |
| Totals | 41 | 116 | 157 |

| | |
|---|---|
| Precision / Positive Predictive Value | 0.821 |
| Recall / True Positive Rate | 0.780 |
| F1 Score | 0.8 |

Calculating the MCC correlation coeffiecient:

$$MCC = \frac{(32 \cdot 109) - (7 \cdot 9)}{\sqrt{(32 + 7)(32 + 9)(109 + 7)(109 + 9)}}$$

$$MCC = \frac{3488 - 63}{\sqrt{39 \cdot 41 \cdot 116 \cdot 118}}$$

$$MCC = \frac{3425}{\sqrt{21887112}}$$

$$MCC = \frac{3425}{4678.36638}$$

$$MCC = 0.73209$$

This model scored higher than the other two, with their scores at 0.57306 and 0.50074 for Decision Tree and Logistic Regression respectively.

### *Overall Success*

With focus on Random Forest Classifier, the scores are documented in the following table. The three success measurements were AUC Score, Recall Score, and F1 Score, and this model performed well for two of them. The one measurement that fell slightly short of the predetermined goal was for the model’s recall, at 78% with the cutoff at 80%.

| Criterion/Metric | Score Necessary for Success | Actual Score |
|---|---|---|
| AUC Score | > 80% | 98.2 |
| Recall Score - Yes | > 80% | 78.0 |
| F1 Score | > 90% | 90.0 |

## Key Takeaways

### *Summary of Conclusions*

The battle against suicide is a battle not to be won overnight. The magnitude of factors that may influence the decision one makes is so immense, it will be impossible to effectively study them all at once.  There is no one size fits all option, and the most important thing will continue to be adaptive to social issues, as well as the needs of those facing crisis situations.

Social influences have been believed to have strong ties with mental health and overall happiness.  Social isolation tends to have negative effects on the well being of those isolated.

The current model supports this conclusion, with the number of friends being the most important factor in predicting suicide attempt risk. Depressed feelings are a strong indication, which even though is expected, it is important to not the error in logic to think all depressed people will attempt suicide.

Other factors that may play a bigger role in prediction include total improvements, which may be explained by an overall change in one's mood and opinion.  An outlook where one is seeking new or different ways to enrich life may overall have a more positive outlook, leading to milder depressive symptoms or the desire to reach out during that moment of crisis.

### *Effective Storytelling*

The Chi-square and Cramer’s V results were depicted in a correlation heat map. The format allows relationships to be determined quickly, the two colors representing whether there is or is not a relationship. For those with a relationship, I added the Cramer’s V category (weak, moderate, or strong) in which it fell in to. Overall, this visualization takes its simplicity and delivers a clear message about the variables involved.

The cross-validation scores are represented as grouped bar charts in addition to table format. The bar chart is an ideal way to communicate which model scored higher and lower with a quick glance while the information laid out in table format is an organized structure that presents the precise scores, detail that may have been lost otherwise.

AUC Scores are line graphs which, when grouped together for comparison, allows one to see where the line sits compared to the other models without having to look between three separate charts, which would be particularly difficult if two of the curves were in the same spot. As a visual, we can see even minute differences.

### **Findings-based Recommendations**

Recommendations for further research include a focus a social isolation, therefore those who have no social support structure or those who don’t have a sufficient support. There are several existing rating scales to measure isolation, which could successfully be integrated into a study or survey environment.  Some options to consider are the UCLA Loneliness Scale, the Duke Social Support Index (used to measure social support of the elderly), or the CEL's 3-question scale.

The UCLA 3 – Item Loneliness Scale would be easiest to implement, as incorporating it into a survey would not be cumbersome. There are three questions, and each as three scores that may be assigned to it.

1. How often do you feel that you lack companionship?
2. How often do you feel left out?
3. How often do you feel isolated from others?

The rating scale is numbered 1 though 3, where a score of 1: Hardly ever, 2: Some of the time, and 3 : Often. These scores are added resulting in a single value within the range of 3 to 9. Typically, those scoring 3 -5 are labeled as “not lonely” while those with scores 6 – 9 are “lonely”.

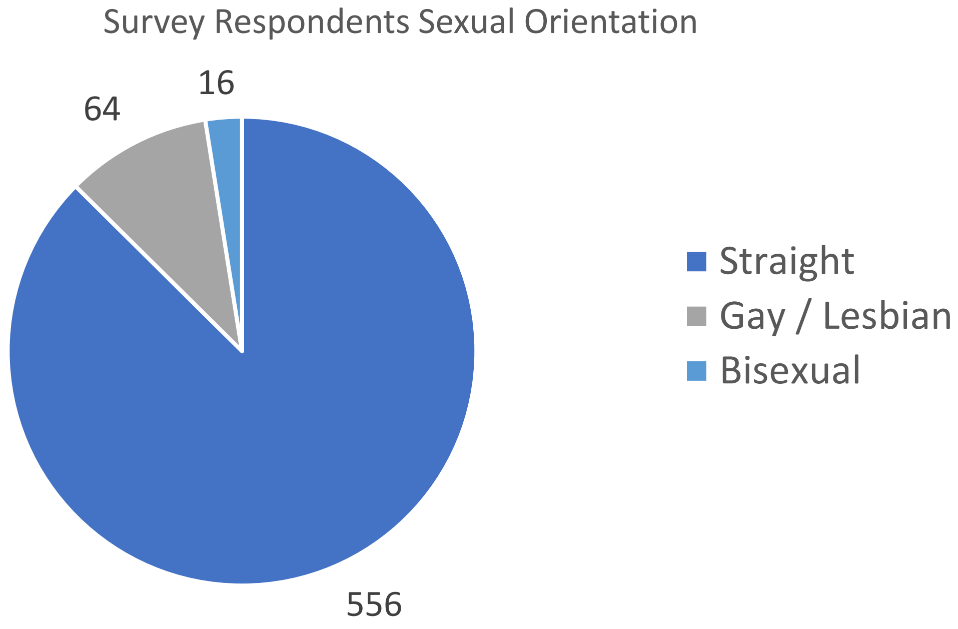
I like this scale for its simplicity, for both adding it to a section of a survey and its implementation. The scores can be totaled, which would create a discrete numerical variable that could be further used to explore from a different perspective.

I would also suggest securing a wider range of individuals who identify in different sexual identities. One limitation with this dataset would be the distribution of individuals and their chosen sexuality. As it can be seen below, there is little representation to groups other than those who are heterosexual, and as topics such as sexuality and gender identity become more mainstream and less of taboo subjects, it brings that likelihood that a larger number of people will be more open about talking with others concerning these areas. The current group breakdown by sexuality category is

Straight – 87.42%

Gay/Lesbian – 10.062%

Bisexual – 2.516%



Having a study which represents these identities could offer strategies and supportive measures for those who may face unique challenges.

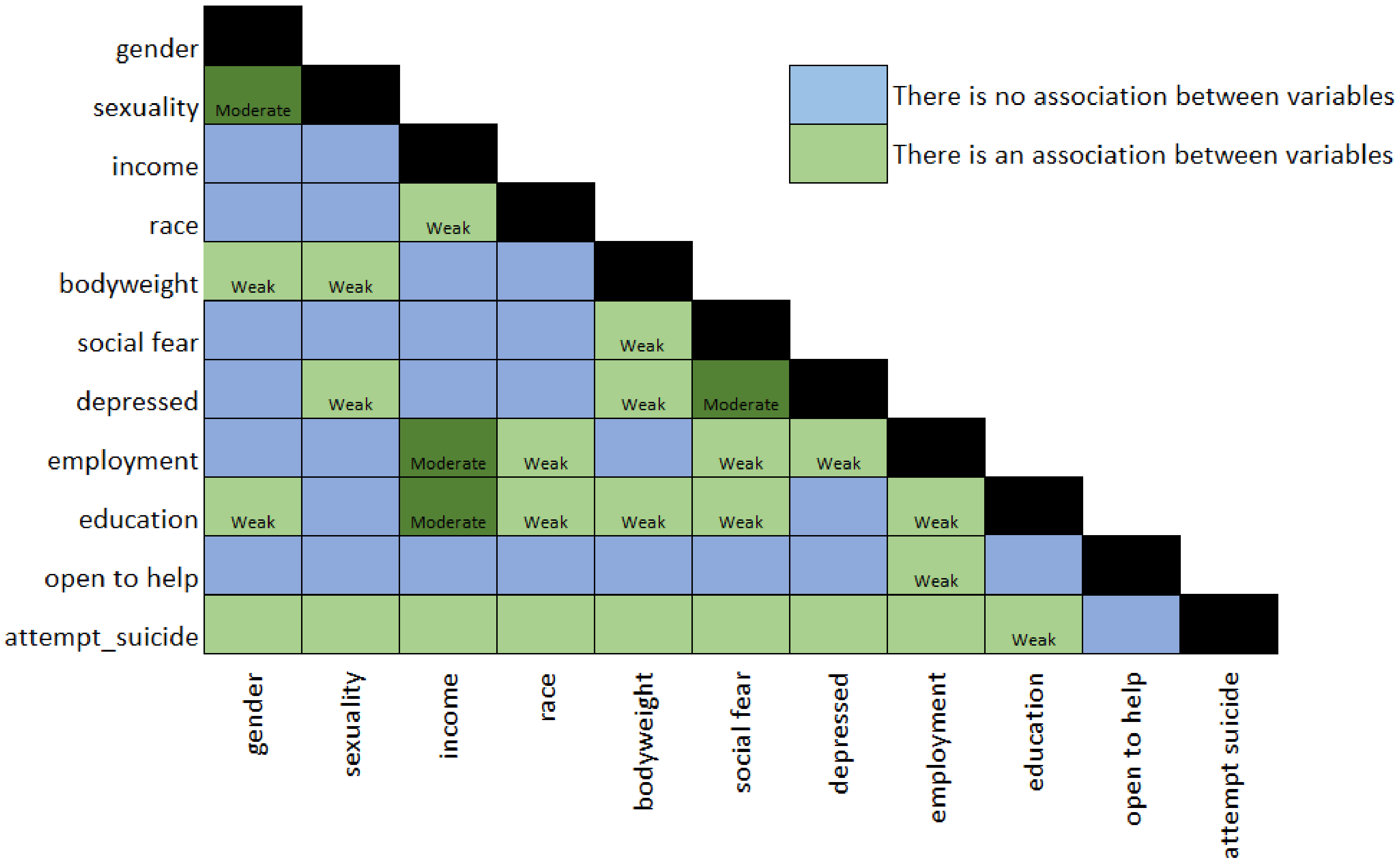
# Appendices

## Evidence of Completion

The following will be submitted as evidence of project completion

1. Jupyter Notebooks with Cleaning, Analysis Code, and Methods
2. Visual Representation of the Random Forest Classifier
3. Excel spreadsheets with scores and charts

# Tables, Charts, Visualizations

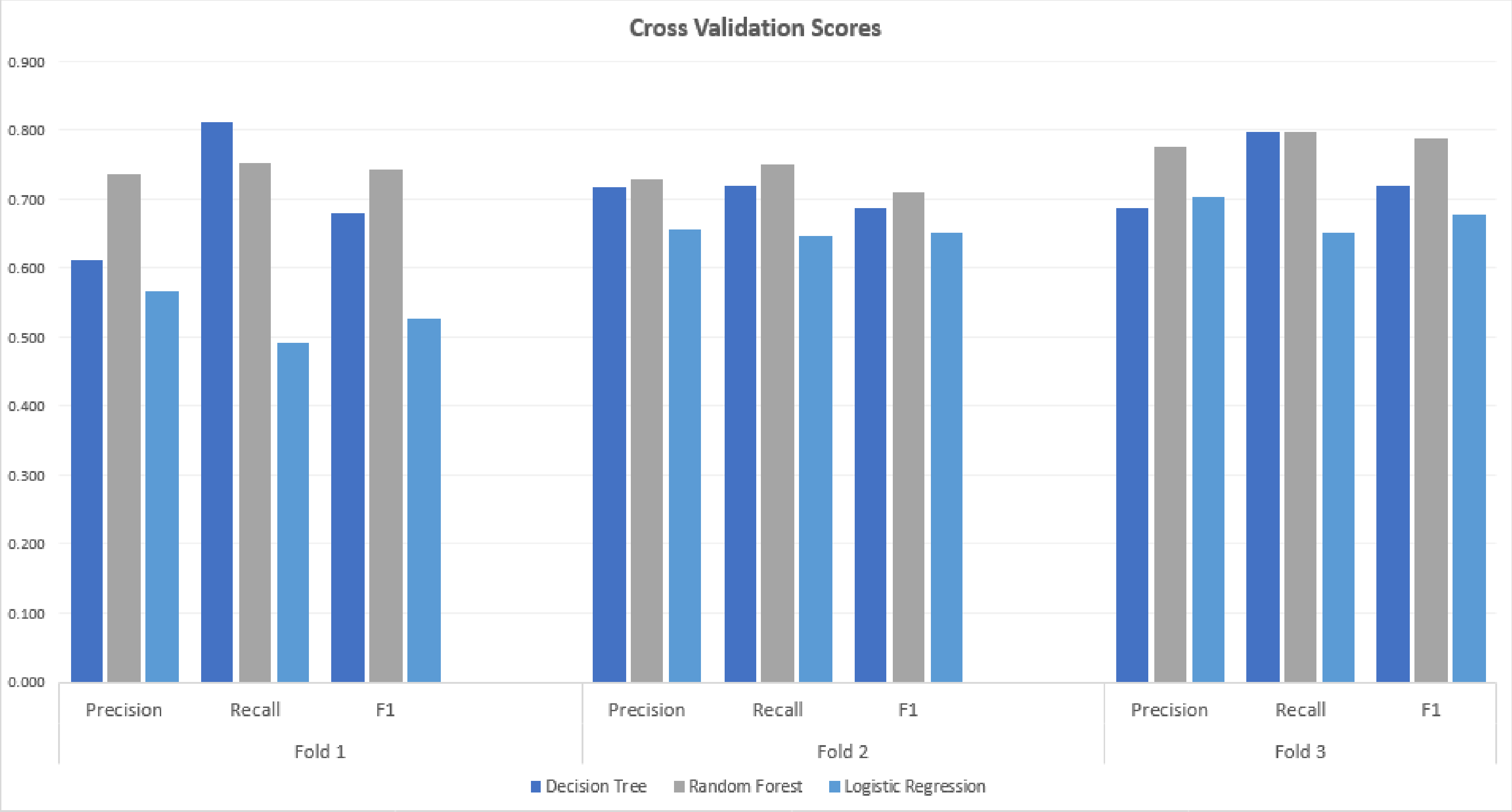


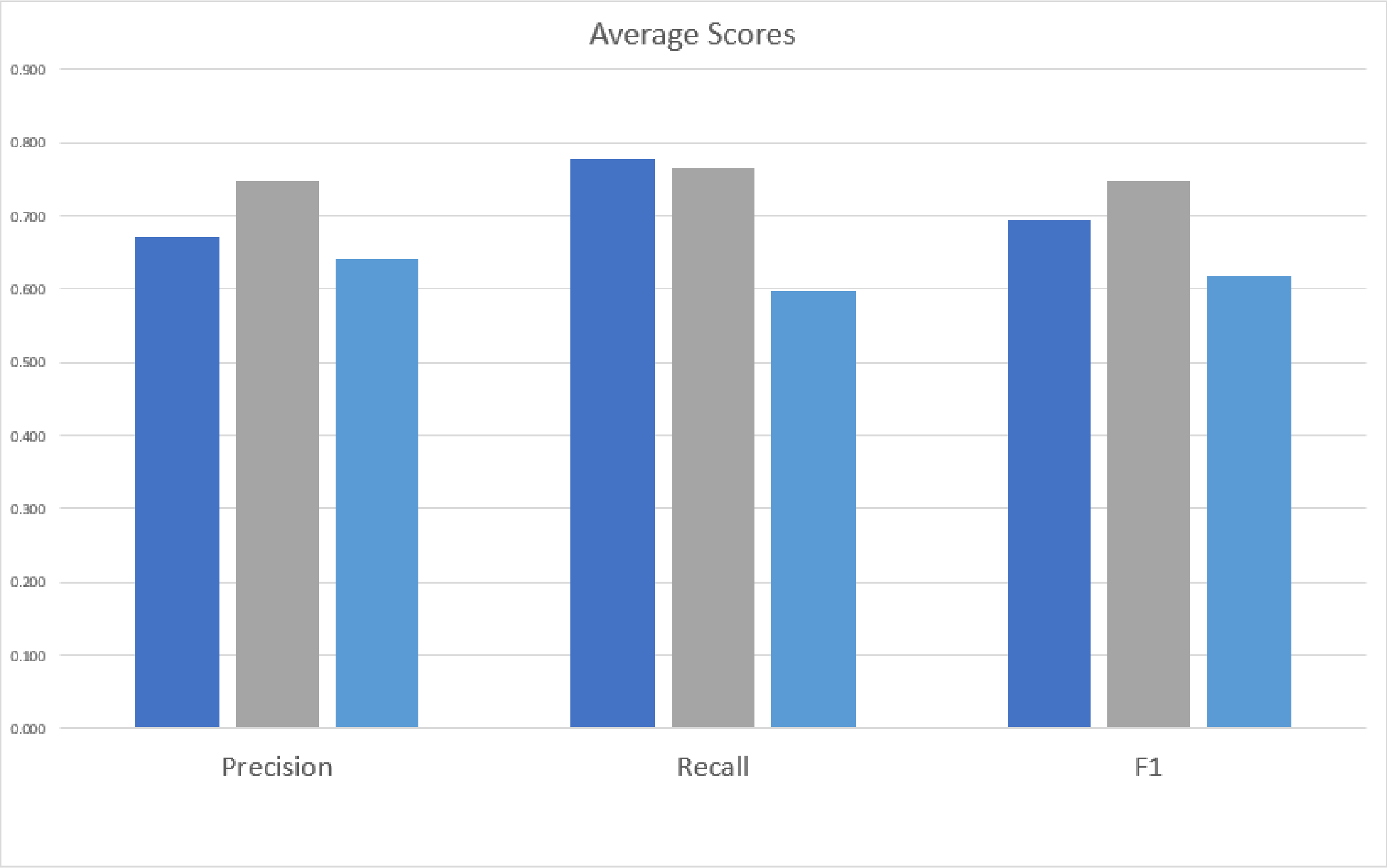
🟦 There is no association between variables

🟩 There is an association between variables

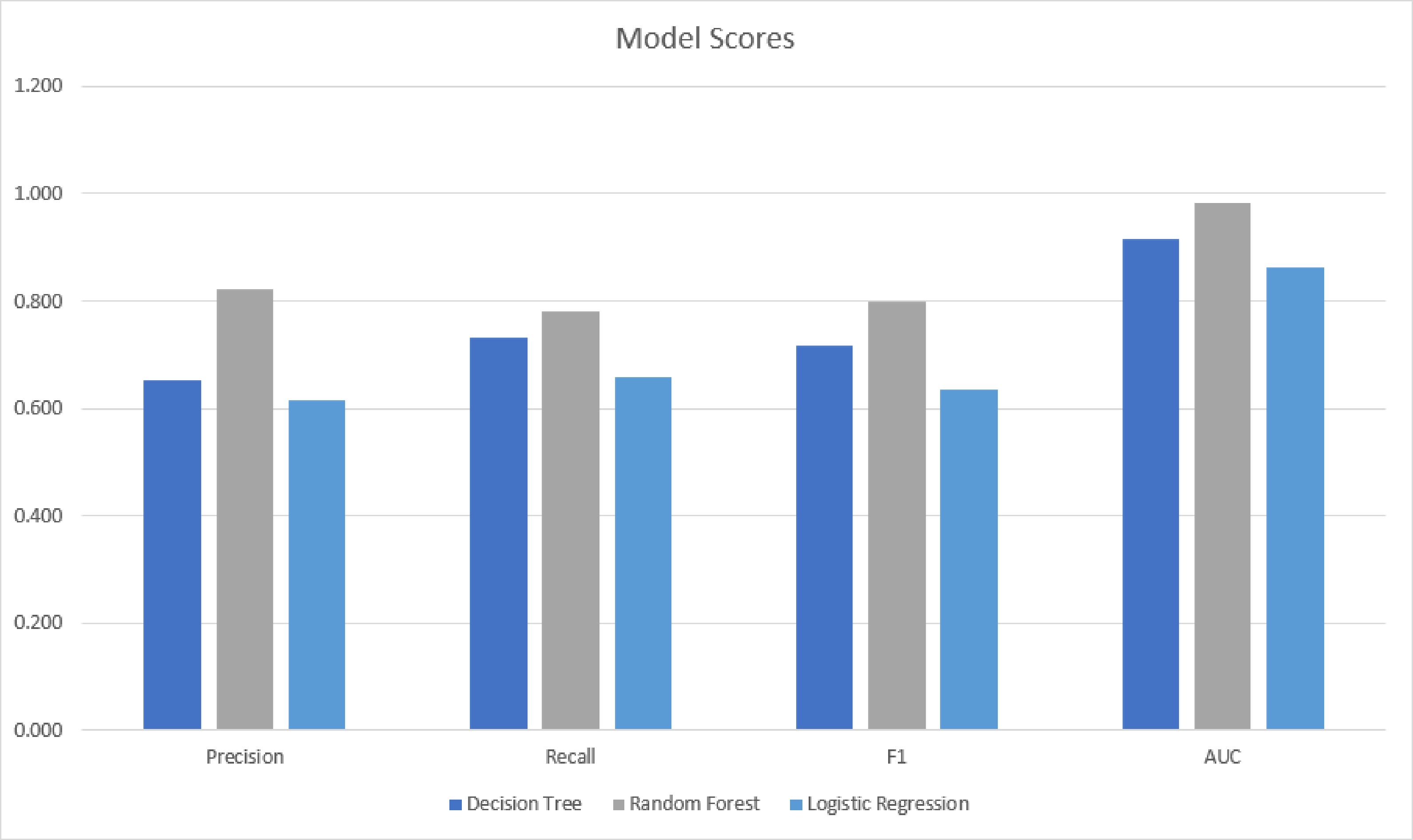
| | gender | sexuality | income | race | bodyweight | social fear | depressed | employment | education | open to help | attempt suicide |
|---|---|---|---|---|---|---|---|---|---|---|---|
| gender | ⬛ | | | | | | | | | | |
| sexuality | 🟩 Moderate | ⬛ | | | | | | | | | |
| income | 🟦 | 🟦 | ⬛ | | | | | | | | |
| race | 🟦 | 🟦 | 🟩 Weak | ⬛ | | | | | | | |
| bodyweight | 🟩 Weak | 🟩 Weak | 🟦 | 🟦 | ⬛ | | | | | | |
| social fear | 🟦 | 🟦 | 🟦 | 🟦 | 🟩 Weak | ⬛ | | | | | |
| depressed | 🟦 | 🟩 Weak | 🟦 | 🟦 | 🟩 Weak | 🟩 Moderate | ⬛ | | | | |
| employment | 🟦 | 🟦 | 🟩 Moderate | 🟩 Weak | 🟦 | 🟩 Weak | 🟩 Weak | ⬛ | | | |
| education | 🟩 Weak | 🟦 | 🟩 Moderate | 🟩 Weak | 🟩 Weak | 🟩 Weak | 🟦 | 🟩 Weak | ⬛ | | |
| open to help | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟩 Weak | 🟦 | ⬛ | |
| attempt_suicide | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 Weak | 🟦 | ⬛ |

*Chi Square and Cramer's V Results 1*

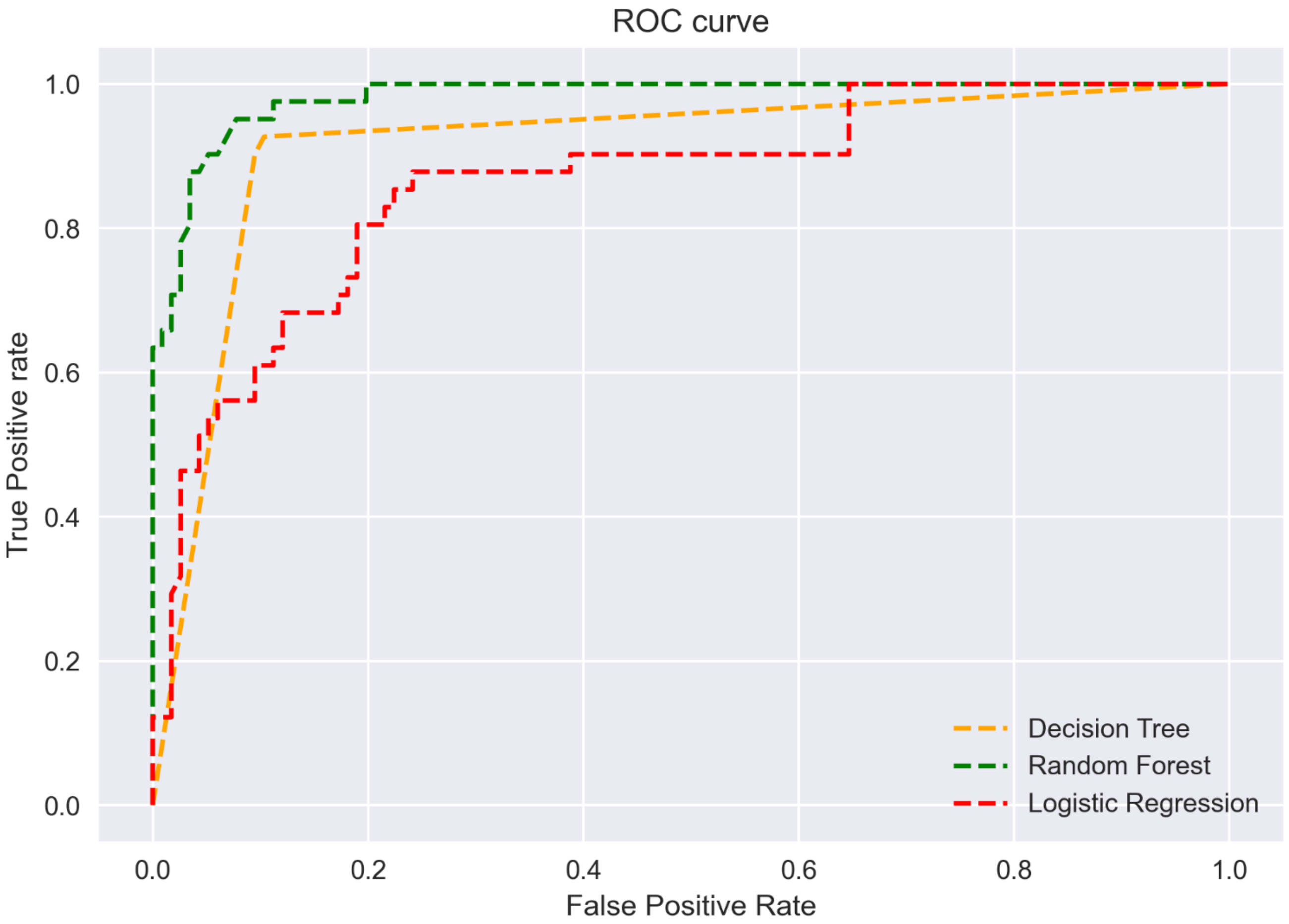

Cross Validation Scores
0.900
0.800
0.700
0.600
0.500
0.400
0.300
0.200
0.100
0.000
Precision
Recall
F1
Fold 1
Precision
Recall
F1
Fold 2
Precision
Recall
F1
Fold 3
Decision Tree
Random Forest
Logistic Regression



*Cross Validation Scores Averaged*



*Model Scores on Full Training Data*


ROC curve
True Positive rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Decision Tree
Random Forest
Logistic Regression

# Sources

Brownlee, J. (2020, September 1). *Hypothesis test for comparing machine learning algorithms*. Machine Learning Mastery. Retrieved May 1, 2022, from https://machinelearningmastery.com/hypothesis-test-for-comparing-machine-learning-algorithms/

U.S. Department of Health and Human Services. (n.d.). *Conducting research with participants at elevated risk for suicide: Considerations for researchers*. National Institute of Mental Health. Retrieved April 28, 2022, from https://www.nimh.nih.gov/funding/clinical-research/conducting-research-with-participants-at-elevated-risk-for-suicide-considerations-for-researchers#guidance