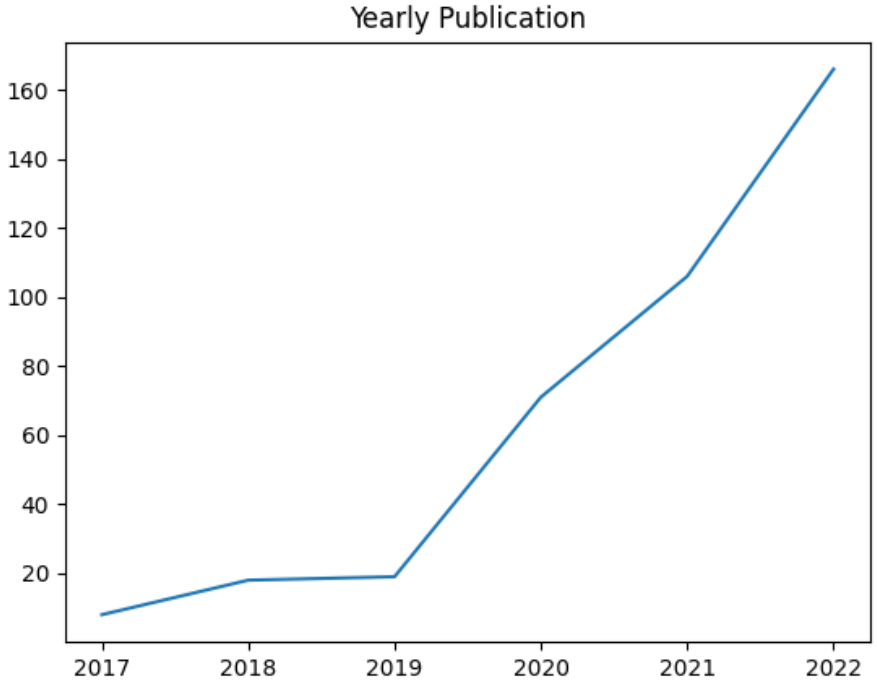
Avary McCormack 99121093

CIS4930 Individual Coding Assignment

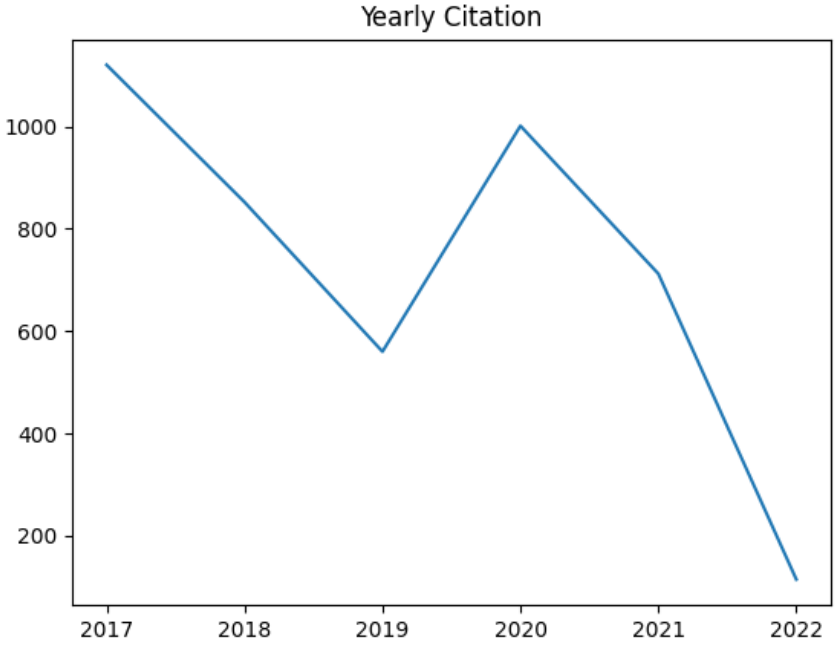
Spring 2023

Python Fundamentals

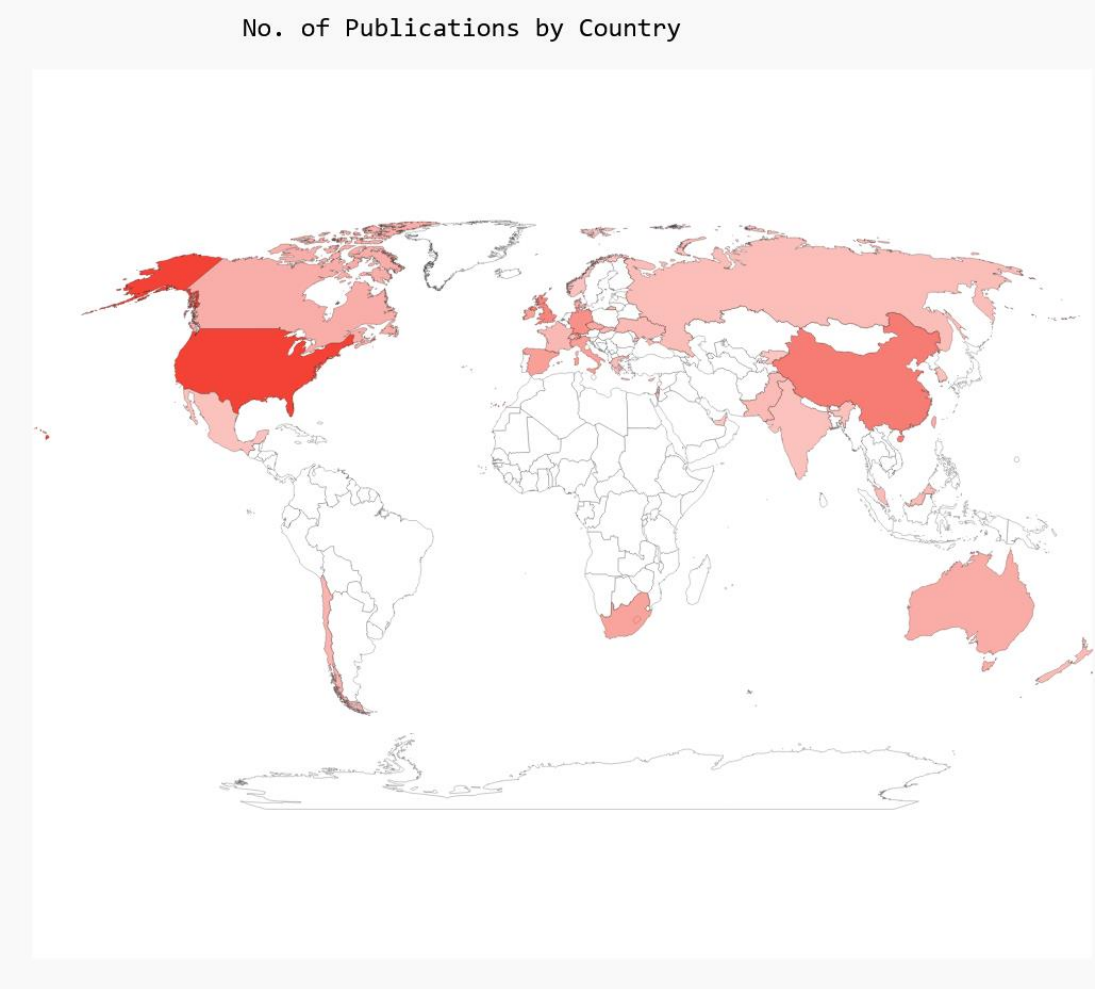
1. As depicted by the graph below, the number of publications increased as time went on. This shows that the topic of VR/AR in education has become more popular with each year.



2. Taking the sum of each year's number of citations per article produces the graph below. Citation either means citations that the article has or how many times that article is cited. Based on the results, it can be assumed that it is referring to how many times that article has been cited. Since articles from 2017 are the oldest. It makes sense that they would be cited the most.



3. In the browser, this an interactive map that allows users to see the number of publications each country published when hovering the mouse over the respective country.
   The US has the most publications which is obvious from the darker shade of red shown.



4. Top 5 institutions that have the most published articles:
   These results were found by taking the number of unique articles affiliated with an institution. Meaning if a certain article had 3 authors all from the same institution, this would only count as 1 publication for that particular institution. Because the dataset is quite small, all institutions had only 1 or 2 publications.

Number of Publications	Author Affiliation
2	University of Copenhagen
2	Malaysia University of Science and Technology
2	University of Bristol
2	Fudan University
2	University of Management and Technology

5. Top 5 researchers with most h-index:

	h-index	Author Name
117	95.0	Ulrich Trautwein
102	63.0	Nicolas Molinari
130	59.0	George S. Athwal
140	33.0	Maria Luisa Lorusso
147	33.0	Vicente A. González

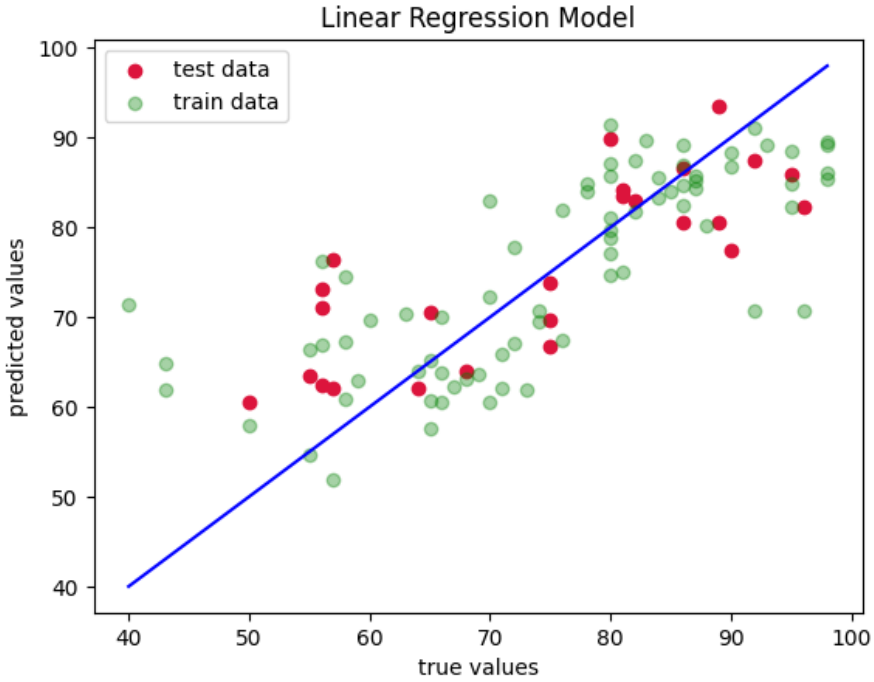
Regression

1. When splitting the dataset into the corresponding training and test sets, there can be some variety in the R^2 score and prediction results. One of the lowest R^2 that I saw was in the negatives and one of the highest was 0.69. However, I added a variable to ensure the test/train sets always resulted in the same, and the following results will reflect from those results.

The R^2 score of the linear regression model was 0.614.

This model was used to predict a set of test and training data. The test and training data predictions were graphed in a scatterplot along with the actual values.

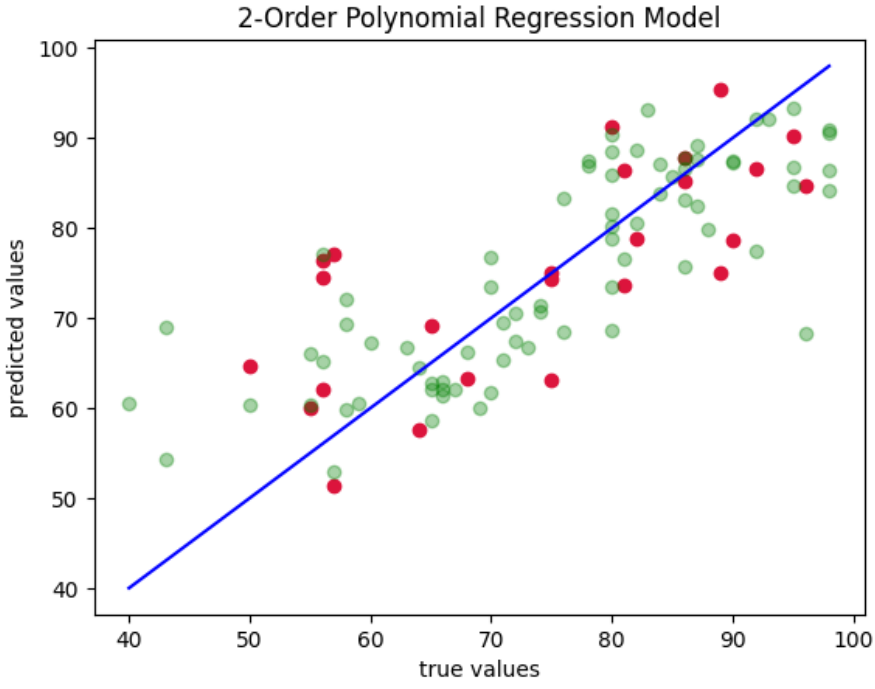
The test data is the unbiased result of the model.



We will also look at the 2-order polynomial regression model.

The R^2 score for the 2-order polynomial regression model was 0.522.

Like before, the test and train predicted vs true values were graphed on a scatterplot.



In this case, the linear regression model outperformed the 2-polynomial regression model.

2.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    SUS   R-squared:                       0.593
Model:                            OLS   Adj. R-squared:                  0.571
Method:                 Least Squares   F-statistic:                     27.39
Date:                Sun, 12 Feb 2023   Prob (F-statistic):           5.25e-17
Time:                        15:18:52   Log-Likelihood:                -362.39
No. Observations:                 100   AIC:                             736.8
Df Residuals:                      94   BIC:                             752.4
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const         93.0282      5.541     16.788      0.000      82.026     104.031
Purchase       1.3412      3.676      0.365      0.716      -5.958       8.641
Duration      -0.0002      0.010     -0.025      0.980      -0.020       0.019
Gender         0.8367      1.971      0.425      0.672      -3.076       4.749
ASR_Error     -1.4254      0.401     -3.553      0.001      -2.222      -0.629
Intent_Error  -2.0092      0.439     -4.572      0.000      -2.882      -1.137
==============================================================================
Omnibus:                        6.969   Durbin-Watson:                   2.023
Prob(Omnibus):                  0.031   Jarque-Bera (JB):                8.115
Skew:                          -0.378   Prob(JB):                       0.0173
Kurtosis:                       4.173   Cond. No.                     1.27e+03
...
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.27e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

First, we will look at the OLS Regression Results. Some key things to look at:

- The R^2 score is 0.593, which means that 59.3% of our dependent variables (SUS) can be explained using out independent variables.
- The F-Test is 27.39 and F-test prob is 5.25e-17. which means there is an extremely small probability that a simpler model could perform with a higher F-statistic. ~ we reject the null hypothesis that “all independent variables have no effect on the dependent variable”.

Coef – if X rises 1 unit, how much does Y rise?

- We can see that is purchase rises 1, SUS rises 1.3412
- When Gender rises 1, SUS rises 0.8367
- And so on

t – t farther from 0 => stronger the evidence against the null hypothesis (the corresponding variable is not significant)

p>|t| - low p => reject null and conclude that variable is statistically significant

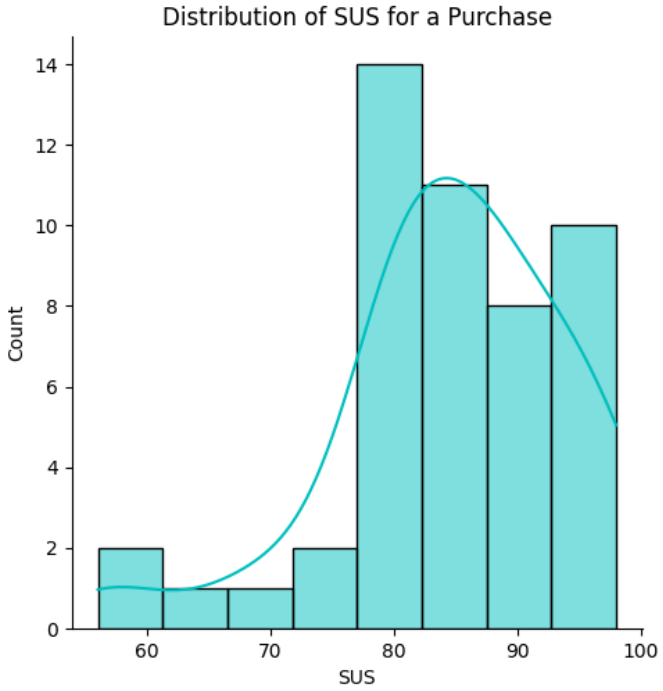
- Only ASR_Error and Intent_Error are statistically significant in changing SUS.

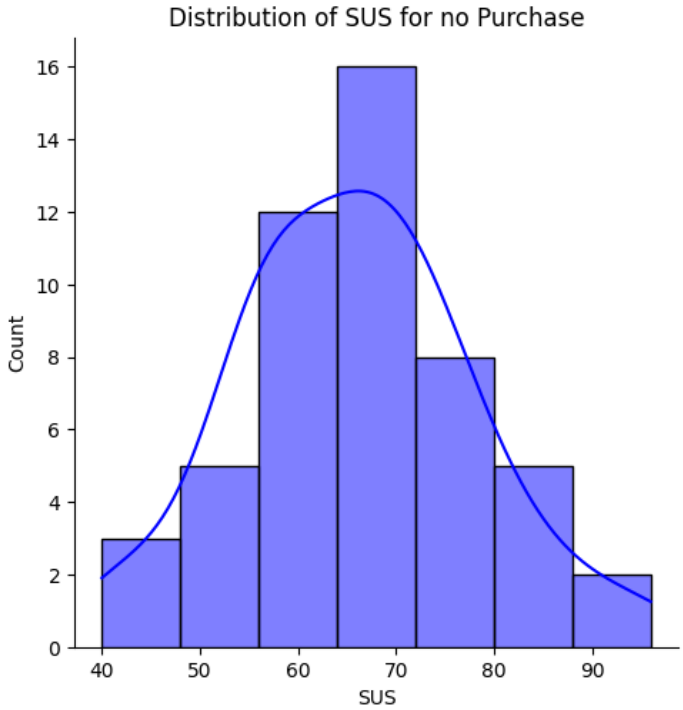
Intent_Error	-0.693675
ASR_Error	-0.662405
Duration	-0.006631
Gender	0.111523
Purchase	0.661931
SUS	1.000000

This is the results of a pair-wise correlations. (Pearson: standard correlation coefficient). +/- 0.7 means strong correlation. Negative values denote the variables are negatively correlated.

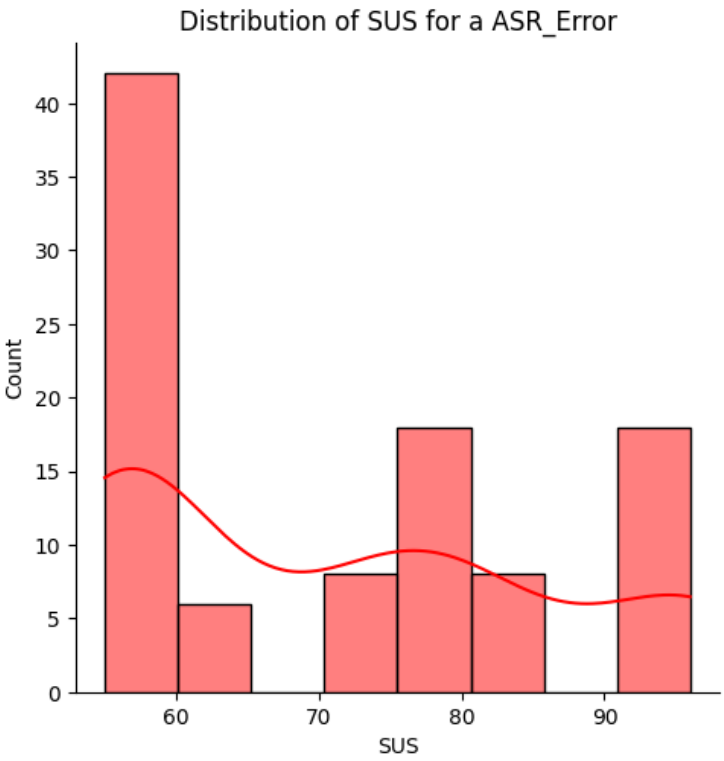
Intent_Error and ASR_Error have moderate negative correlations. Purchase has a moderate positive correlation.

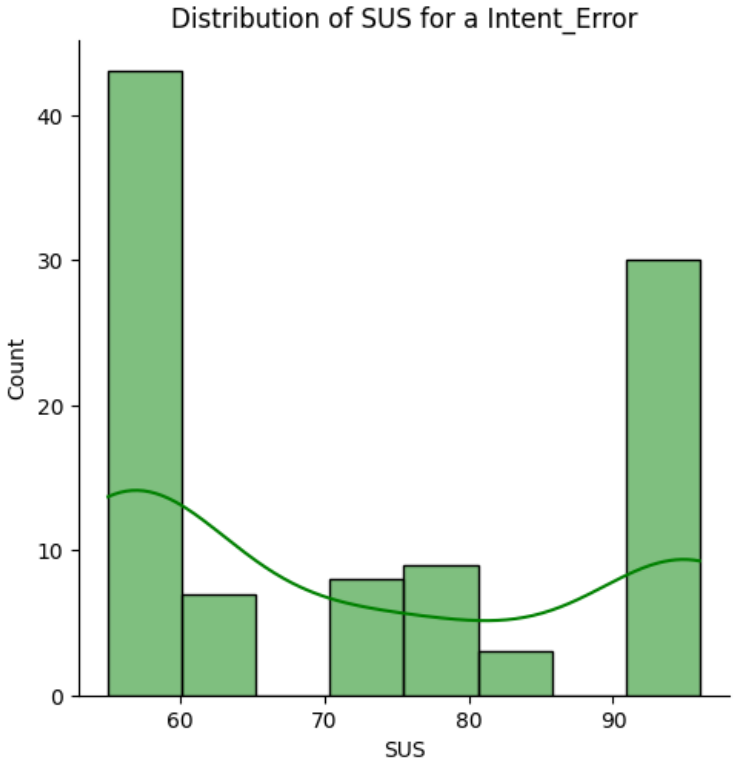
The features that are significant are Intent_Error, ASR_Error and Purchase. All 3 have a moderate correlation.
The features that are insignificant are Duration and Gender.
Below are the distributions of the significant featres with respect to SUS.





As we can see, the difference in Purchase vs no Purchase is slight in that a successful Purchase leans toward higher SUS scored, and no purchase is more heavily centered.





ASR_Error and Intent_Error's negative correlations are evident in these graphs.

3. Intent_Error is the number of times the system failed to classify the user's intention/speech act. ASR_Error is the number of times Siri fails to recognize the user's speech.

   It makes sense the ASR_Error and Intent_Error are negatively correlated with SUS scores. Siri performing poorly leads to worse SUS scores.

   Purchase denotes whether a customer purchased a ticket or not by using Siri. It makes sense that this is positively correlated since a successful purchase could lead to a higher SUS score.

   Gender and Duration do not have significant correlation in this case. It was expected that gender does not have a significant correlation to SUS since it does not have much to do with how well Siri performs.

   It was also expected that Duration has no significant correlation since the duration of dialogue could mean several things. A short dialogue could either mean Siri did very well or it could mean Siri did terrible and the user gave up immediately.

4. Based on the Pearson Coefficients, Intent_Error seems to have the greatest correlation with SUS scores. Intent_Error has -0.69 which is the farthest score from zero, denoting it's correlation strength.

5. Some potential reasons for Intent_Error being the most significant predictor lie in the purpose of this data itself. Users are attempting to buy tickets using Siri. The users are then asked to take a survey which involves scoring their interaction with Siri. Obviously, if the system could not classify what the user wanted or said, Siri would also not be able to help the user or may provide wrong information. In contrast, ASR_Error, which seems similar, is aimed at the number of times Siri fails to recognize the user's speech. This is different because Siri may just ask a user to repeat themselves, while Intent_Error misclassifies and potentially delivers incorrect information.

1. Problem Statement

In a dataset, researchers have collected information regarding the user interaction with Siri while purchasing a ticket. Given several factors: ASR_Error, Intent_Error, Duration, and Gender, classify whether or not a user purchased a ticket using Siri. It is important to know the threshold standard in which Siri needs to perform in order to yield positive user experience and ticket purchasing. I solved this problem by training 4 machine learning classification algorithms and comparing the performance.

2. Data Preparation

The first step in data preparation that I took was reading the csv file into a pandas dataframe. Next, I returned the sum of na values in the dataframe and filled na values with 0. From there, I created two variables x and y. X received the Purchase data as a numpy array. For Y, I dropped the Purchase and SUS columns as they were not needed for this particular variable.
I also needed to scale the X dataset in order to split into training and test data.

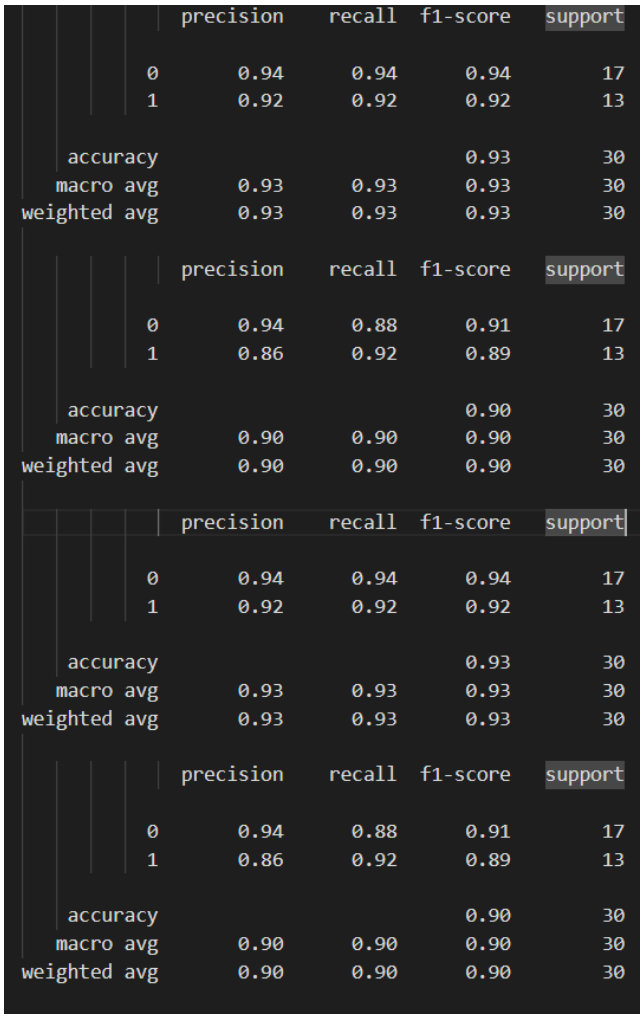
3. Model Development

- Model Training

  The models I selected were the Logistic Regression, the Support Vector Machine, the Naïve Bayes, and the Random Forest models.

  The training and test sets were split by a 30/70 split. 30% of the data became test data and 70% became training data.

- Model Evaluation



	precision	recall	f1-score	support
0	0.94	0.94	0.94	17
1	0.92	0.92	0.92	13
accuracy			0.93	30
macro avg	0.93	0.93	0.93	30
weighted avg	0.93	0.93	0.93	30

Logistic Regression
Accuracy is 0.93
Both f1-scores are high.

	precision	recall	f1-score	support
0	0.94	0.88	0.91	17
1	0.86	0.92	0.89	13
accuracy			0.90	30
macro avg	0.90	0.90	0.90	30
weighted avg	0.90	0.90	0.90	30

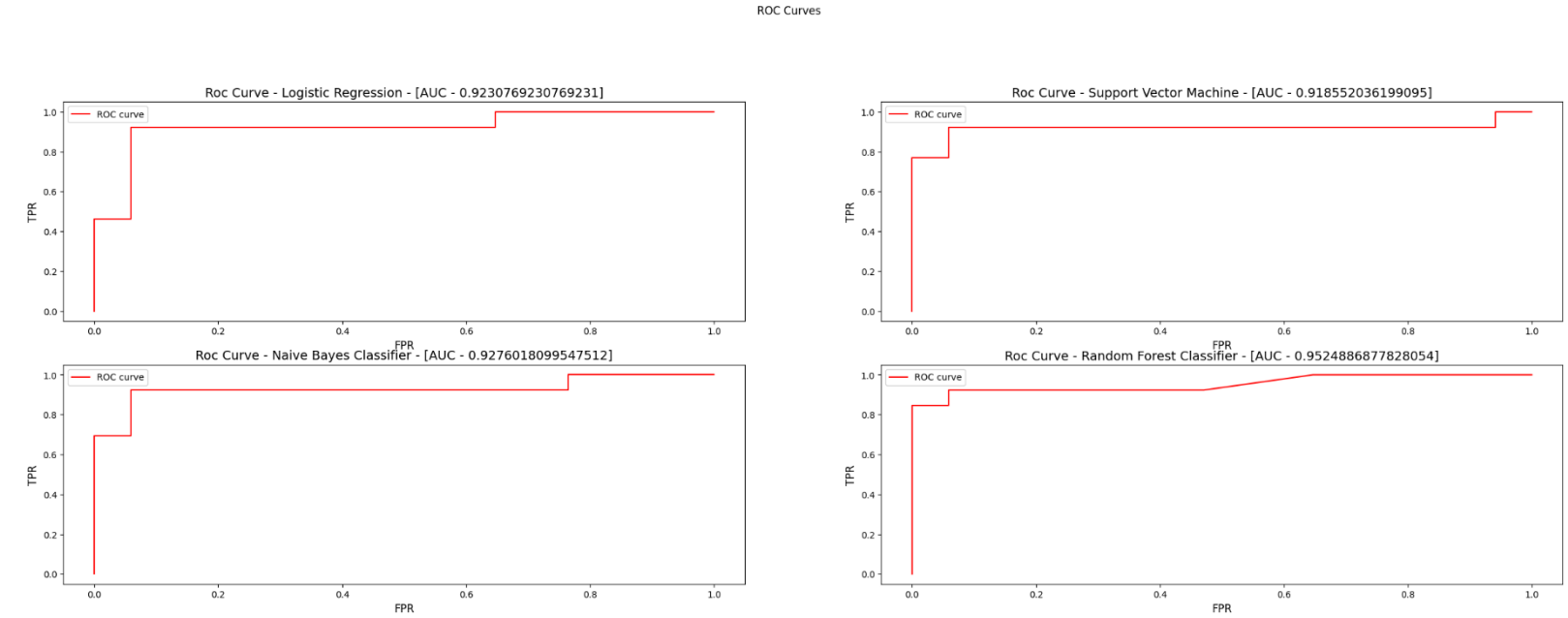
SVM
Accuracy is 0.90
Both f1-scores are high

	precision	recall	f1-score	support
0	0.94	0.94	0.94	17
1	0.92	0.92	0.92	13
accuracy			0.93	30
macro avg	0.93	0.93	0.93	30
weighted avg	0.93	0.93	0.93	30

Naïve Bayers
Accuracy is 0.93
Both f1-scores are high

	precision	recall	f1-score	support
0	0.94	0.88	0.91	17
1	0.86	0.92	0.89	13
accuracy			0.90	30
macro avg	0.90	0.90	0.90	30
weighted avg	0.90	0.90	0.90	30

Random forest
Accuracy is 0.90
Both f1-scores are high



Here we can see the ROC curves and AUC scores for the 4 models.

Logistic Regression

AUC is 0.923

SVM

AUC is 0.918

Naïve Bayers

AUC is 0.927

Random forest

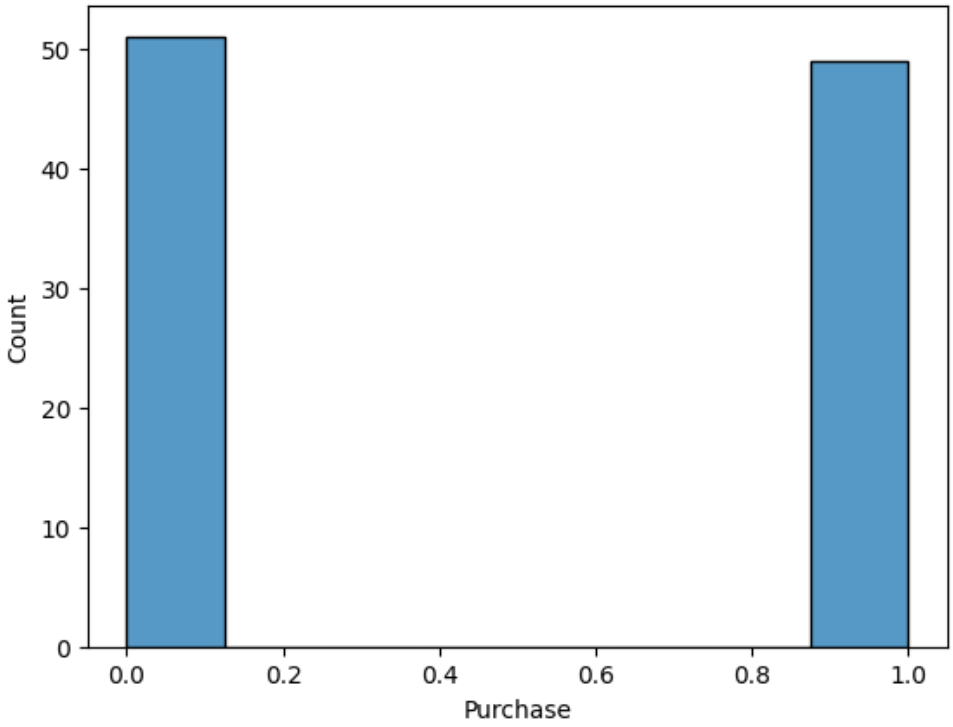
AUC is 0.952

From the AUC scores, we can determine that the Rain Forest Model performs the best.

Sidenote: we do not have to use SMOTE because the classification is not imbalanced.

(See →)

(and if we did use SMOTE, it yields the same results)



4. Discussion

All models performed very well with the Random Forest model performing the best.
A challenge I faced was figuring out the meanings and interpretations for the data. In order to solve them, I rewatched lecture videos and researched the interpretations online.

- *[Any reflections or thoughts on this assignment?]*
  There are so many ways to visualize data and fit models that it can become overwhelming; however, it is extremely interesting and rewarding to put all the data together in a cohesive manner.

5. Appendix

Src code - https://github.com/avary8/Linear-Logistic-Regression