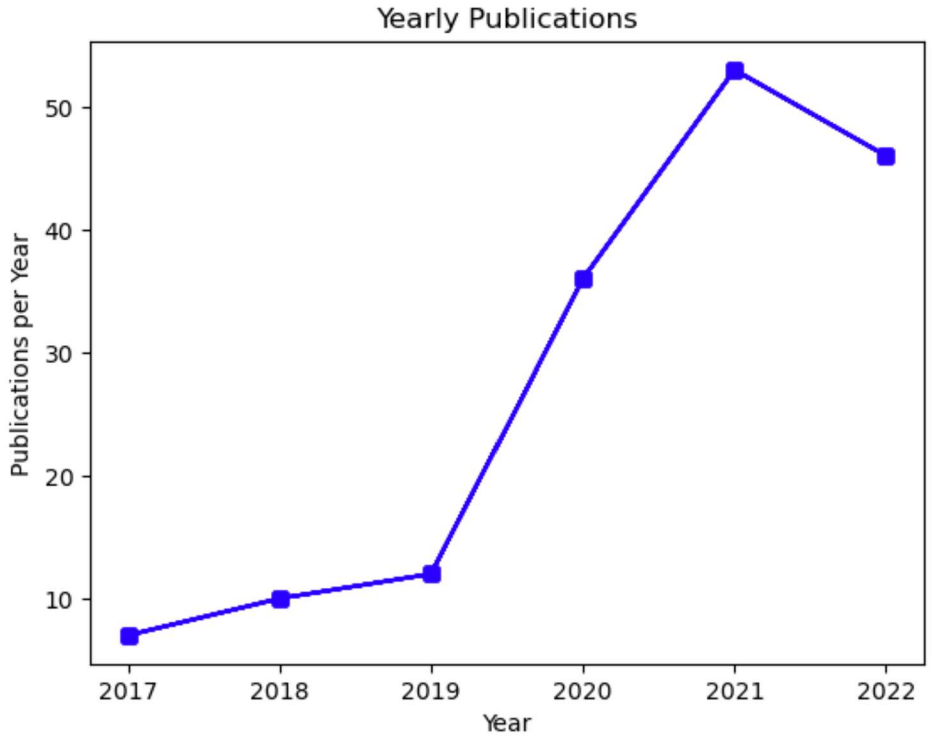
[Nathan Gilman, UFID: 34785552]

CIS4930: Introduction to Multimodal Machine Learning in Python
Individual Coding Assignment 01
Due Feb 07, 2023, 11:59 pm

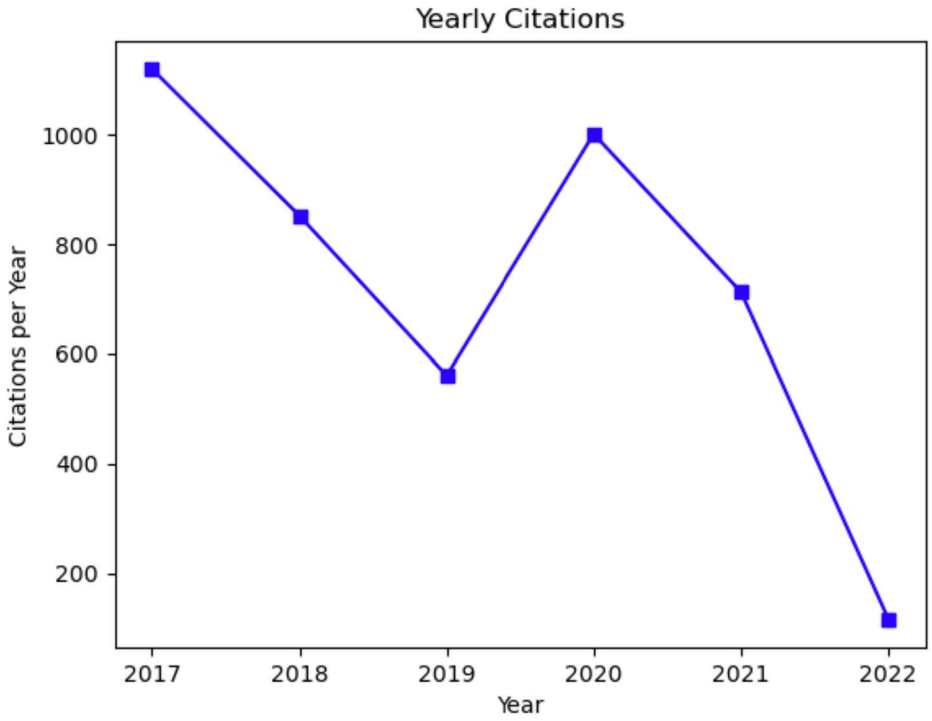
Jupyter Notebook: https://github.com/npgilman/CIS4930-HW1/blob/master/programming_assignment_1.ipynb

Python Fundamentals (20 pts)

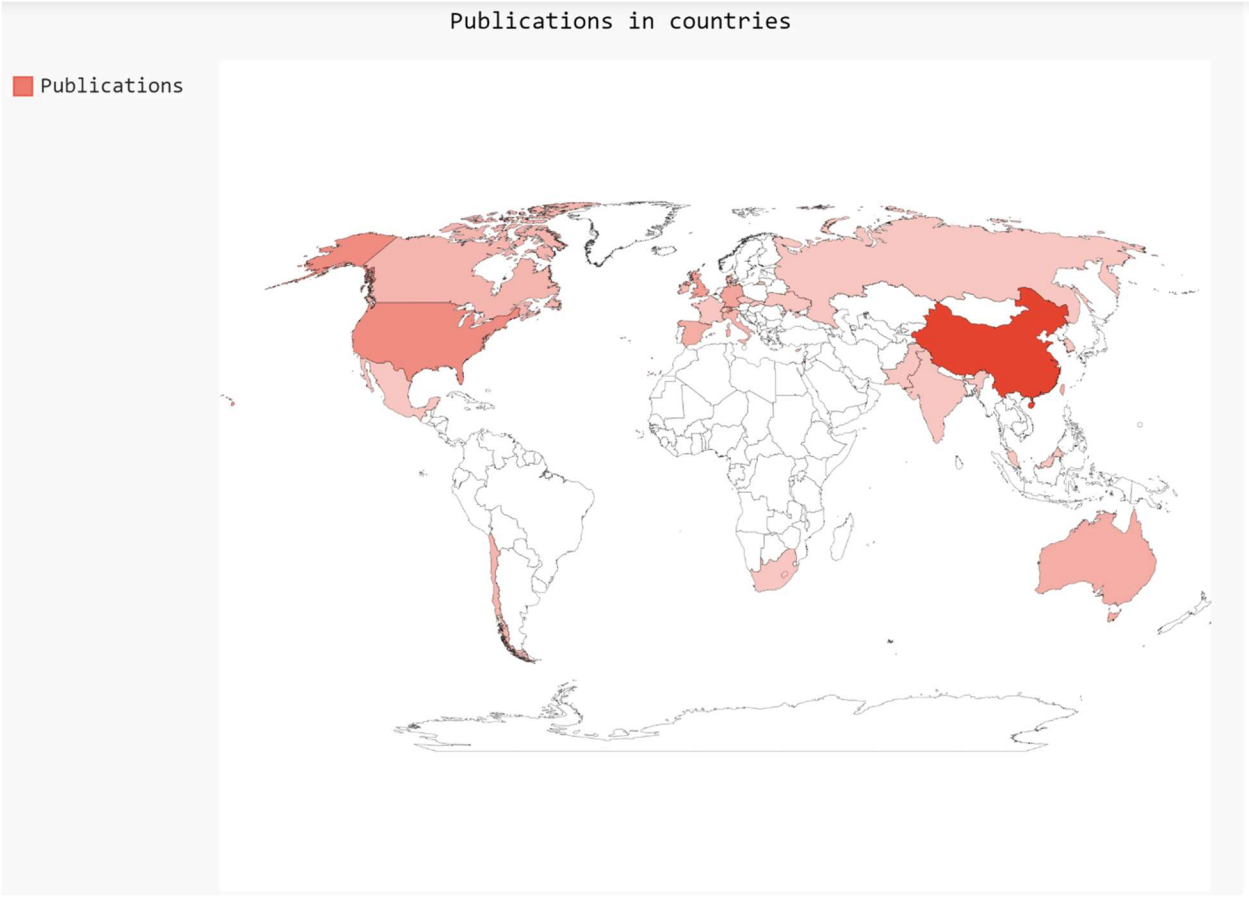
1. Plot the yearly_publication figure, in which the x-axis is the year, the y-axis is the number of articles published during that year.



2. Plot the yearly_citation figure, in which the x-axis is the year, the y-axis is the total number of citations during that year.



3. Plot the figure of the number of publications across countries. You may use any available python libraries, such as pygal_maps_world, geopandas, or others



4. What are the top 5 institutions that have the most published articles in this area?
   The top 5 institutions are *Fudan University*, *University of Bristol*, *University of Management and Technology*, *University of Copenhagen*, and *Malaysia University of Science and Technology*.

5. Who are the top 5 researchers that have the most h-index in this area?
   The top 5 researchers by h-index are *Ulrich Trautwein*, *Nicolas Molinari*, *George S. Athwal*, *Maria Luisa Lorusso*, and *Vicente A. González*.

Regression (40 pts)

1. Show the statistical results of your trained regression model.

```
                            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:                Tue, 14 Feb 2023   Prob (F-statistic):           5.25e-17
Time:                        11:57:42   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.
```

   After fitting a linear regression model, I found the following.

```
The R square score of linear regression model is:  0.602540864943556
```

2. What features are significant? What features are insignificant?
   The features that are significant are *Purchase*, *ASR_Error*, and *Intent_Error*. The features that are insignificant are *Gender* and *Duration*.

3. Were the results what you expected? Explain why or why not, for each feature.
   The results are what I expected and here are my reasons for each feature:
   a. Purchase: This is a significant feature. This makes sense because the user's satisfaction would be directly linked to the outcome of interacting with Siri. The goal of the interaction is buying a ticket, thus if a user is able to make a purchase it is reasonable they would be satisfied.
   b. ASR_Error: This is a significant feature. This makes sense because if the user experiences issues with Siri not being able to understand their speech, they would not have a good experience. This feature is inversely proportional to satisfaction ratings.
   c. Intent_Error: This is a significant feature. This makes sense because if Siri understands the speech, but does not correctly identify what the user desires to do the interaction will be unsatisfactory.
   d. Gender: This is not a significant feature. It makes sense that gender is not significant because it has little to no impact on the interactions with the device.
   e. Duration: This is a significant feature. It makes sense that duration is the least significant feature because I believe the user would focus more on interactions with the device and the outcomes of those interactions rather than the duration of time interacting.

4. What does the model suggest is the most influential factor on SUS? Explain what tells you this is the most influential factor statistically.
   The model suggests that Intent_Error is the most influential on SUS. This is indicated by the higher magnitudes for pairwise correlation and OLS regression coefficients for Intent_Error.

5. What are the potential reasons for these factor(s) being significant predictors of SUS?
   Similar to question 3, it makes sense that Intent_Error is the most influential feature. This makes sense because if Siri understands the speech, but does not correctly identify what the user desires to do the interaction will be unsatisfactory.

Classification (40 pts)

[Report on following pages]

[Problem 3 Report]

1. Problem Statement

Task-oriented dialogue systems are becoming increasingly popular and integrated into modern applications. Siri and Cortana are prominent examples of this trend. These systems are powerful and versatile, able to complete a variety of tasks. For example, a user is able to converse with a task-oriented dialogue system and purchase an airline ticket. Despite their many use cases, these systems sometimes struggle to correctly understand speech or user intent.

Given a specific dataset of user feedback, we can create a classification model that predicts if the system was used to purchase an airline ticket based on the variables of *ASR_Error*, *Intent_Error*, *Gender* and *Duration*. This is useful because it will help us understand the relationship between system errors and system usability.

2. Data Preparation

To begin, I used the Pandas python library to read "data.csv" into a dataframe. I noted that a seventh column appeared in this initial dataframe. Since the column was soley NaN values and titled "Unnamed: 6", I removed it using the delete function. After that, I extracted the "Purchase" column to a separate dataframe since it was the dependent variable. This left me with two dataframes for the dependent and independent variables. I also removed the "SUS" column from the initial dataframe because it was not listed as an independent variable in the problem.

After diving my data into independent variables and dependent variables, I needed to split those further into training and testing data. Using the test_train_split() function, I divided the data into training and testing sets for later use. I decided to use an 80-20 split for the training to testing ratio because many popular resources online recommended it.

3. Model Development

- Model Training

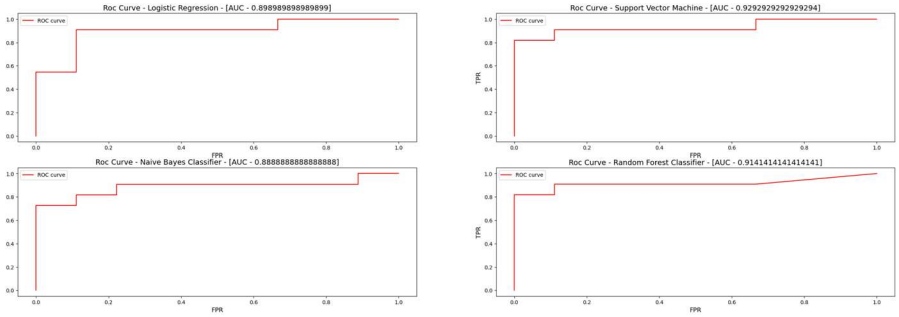
  In this problem, I needed to compare the applications of 4 types of classification algorithms: Logistic Regression, SVM, Naïve Bayes, Random Forest. Because of that, I implemented them all.

  I split the data into a 80 – 20 split, reserving 80% of the data for training and 20% in the test set. I did this because many popular resources online suggested that this was an optimal range. I was worried that the small data set might limit the accuracy and application of these models but I proceeded anyways.

  - *[Describe the training phrase, which may include what models did you select, how you split training/validation/test sets, training epochs, and any other parameters.]*

- Model Evaluation

  From the below graph, we can see that the Support Vector Machine classification performed the best from ROC curve. It has an area under curve value of 0.9292 whereas the next best classification algorithm in this example is the Random Forest Classifier which has an AUC value of 0.9141.



  However, the accuracy and f1 score of the Logistic Regression, Support Vector Machine and Random Forest Classifier were all the same with an accuracy of 0.85. This means that the model correctly predicted whether a purchase was made 85% of the time.

	precision	recall	f1-score	support
0	0.88	0.78	0.82	9
1	0.83	0.91	0.87	11
accuracy			0.85	20
macro avg	0.85	0.84	0.85	20
weighted avg	0.85	0.85	0.85	20

4. Discussion

The models have an accuracy of 80-85% when predicting if a purchase was made. I believe that the limited dataset limits the models to differentiate form each other. There were only 20 items in my y_test dataset. Perhaps with a larger dataset more accurate results would appear. I believe that if this was used in a real context it would need to be refined further.

Challenges:

Some challenges that I faced during the data preparation stage was my inexperience with the pandas library. Pandas is a powerful library that has many tools for cleaning the data. However, it took me a long time to find the correct tools to use. I was eventually able to clean the data properly by reading the documentation online and using the functions that I found to be appropriate for my use case.

Reflection:
This assignment was difficult to me because I have not implemented regression or classification models independently before so I was using the knowledge I used in class. Although there were many key points touched on in class, it took me some time to figure out and understand the modules covered in class.

5. Appendix

The link to my github code repository is https://github.com/npgilman/CIS4930-HW1. It contains a .ipynb file that I created through Jupyter Notebook and worked on the problems in, as well as 3 csv files that I loaded into dataframes to complete the homework.