COURSE 18-785: DATA, INFERENCE & APPLIED MACHINE LEARNING

ASSIGNMENT 4

Nchofon Tagha Ghogomu

ntaghagh

October 14, 2024

LIBRARIES:

The following libraries were used:

- Pandas[1]
- Numpy[2]
- Pyplot from Matplotlib [3]
- Stats from Scipy
- Seaborn
- Linear model from Sklearn
- Linear Regression from Sklearn Linear model
- Logistic Regression from Sklearn Linear model

- import pandas as pd # Pandas
- import numpy as np # Numpy
- from scipy import stats # Scipy - for statistics
- import matplotlib.pyplot as plt # Matplotlib - for ploting
- import seaborn as sns # Seaborn - for ploting
- from sklearn import linear_model
- from sklearn.linear_model import LinearRegression
- from sklearn.linear_model import LogisticRegression
- from sklearn.metrics import confusion_matrix, classification_report
- from sklearn.model_selection import train_test_split
- import statsmodels.api as sm
- from tabulate import tabulate

Programming Language:

- Python

INTRODUCTION

This assignment was made of 4 practical questions that portray real application of data analytics on world data. This assignment was centred around:

- Linear regression with one or multiple explanatory data
- Model fitting and estimation.
- Hypothesis test for correlation.

This assignment had one open study for us to assess and study the trend in the transport domain of transportation.

SOLUTIONS

Question 1: Statistical learning

1.1 Rule-based approach to decision-making. Is any domain knowledge required to establish a rule? Support your answer with an explanation.

a. Determine the issue:

To implement a rule-based approach, we need to clearly identify and define the issue. It is essential to detailly outline the problem that needs to be resolve or decided on. For example, to curb the spread of the Marburg virus the medical team in Rwanda needs to quickly and efficiently test individuals for Marburg.

b. Establish rules/test cases:

Based on experience, data, and other sources of credible information, the guideline for identifying this issue is defined. In our example, based on the symptoms demonstrated by bearers of this virus, a list of test cases is identified. This could include checking for fever, temperature, vomiting, and contact with infected patients. This will certainly come from the professional experience of the team.

c. Implement the rules:

Subsequently, these test cases are applied on samples and real-world scenarios. In this case study, the medical team will use the established rules during the screening of people for this disease and the results are noted.

d. Test the rules and evaluate:

Finally, these set of rules is evaluated to determine its effectively and based on the results, the rules can be revisited and upgraded for better results. In the case study, the medical team identify the shortcomings of the roles and works upon them.

Another example of a rule-based approach is educational recommendation systems that recommend specific contents for students based on their data and needs.

Domain knowledge is very essential as they guide the team in determining this rule. In the above case study, domain knowledge would be understanding the symptoms of the virus and knowing how to test for them.

1.2 Over-fitting, Simple vs Complex Models.

Over-fitting refers to a situation where the model memorises the training data too well but fails to perform on out-sample data. In some cases, the model tends to fit noise, making it overly complex and less reliable. Overfitting inhibits generalisation.

From the standpoint of parsimony, I would choose the simple model with a single parameter. Simpler models are easily understood while more complex models would make it difficult to see how each parameter contributes to the overall model.

1.3 Two commonly used approaches to avoid over-fitting.

a. Regularisation: In this technique we prevent overfitting by penalising the model using a cost function to prevent/discourage models that tend to be too complex.

b. Cross-Validation: In this case, the data set is split into two: one part is used for training and the other is used to evaluate the model's performance.

1.4 Provide two examples of metrics used to evaluate the performance of a model and give a formula for each one. Give two examples of applications and appropriate metrics for each case.

Some Matrices used in evaluating a model include:

- *R-squared value:* It measures the amount of variance explained by the model given by the ratio of the explained variance with the total variance of the data. The R-squared value asses how well a regression model fits the data. An example application of this matrix will be creating a model to predict the price of a commodity given several parameters. If for example, we build a model with all parameters and another model with few parameters, the R-squared valued will help us asses the performance of the model. A value closer to one indicates a better model.

$$R^2 = 1 - \frac{\sum_i (y_{i-}\hat{y}_i)^2}{\sum_i (y_{i-}\bar{y})^2}$$

- Accuracy: This is a measure of how well a model gets a prediction right. It can be applied on logistic regression models. Of course, models with higher accuracy are better. For example, in the medical field, the accuracy, precision, and specificity can be used to evaluate drug performance.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

1.5 Why are benchmarks useful in machine learning and give two examples.

Benchmarks are crucial in machine learning as they help us evaluated and compare the performance of models to guide our selection. Benchmarking is important for:

- Deciding if a model meets validation criteria,
- Comparing between models and selecting optimal performance,
- Improving our models.

For example, we want to develop a model on a given data set. We can build different models using neural networks, support vector machines, decision tree, to list a few. At this stage we evaluate the models using the same matrix as precision, accuracy, F1-Score, Mean Squared Error, to list a few. Based on these results, and or haven compared our model with a standard model, we will be able to rate a model's performance.

Another example could be in a medical lab, when we want to choose a model that is the most accurate in identifying a disease.

2. Machine Learning (25 points)

2.1 What is machine learning? Discuss its evolution over time and why is it popular?

Machine learning is a knowledge discovery process based on data and experience. The growing popularity of machine learning is backed by the fact that algorithms are now capable of understanding partens in complex set of data, that would otherwise be very difficult for humans. The daily discovery special use cases, and its application in diverse fields in practical ways have also contributed to its fame.

The concept of machine learning all started when Alain Turing brought in the idea of learning machines in 1950. Later in 1952, Author Samuel of IBM developed ELIZA, the first game-playing algorithm that prepared people for victory against a world champion. In 1957 The idea of neural network was introduced by Frank Rosenblatt with the discovery of the perceptron, a simple linear classifier. In the 1990, the concept of Artificial Intelligence came around as knowledge in computer science and statics were brough together to create different data-driven machine learning approaches. Then there was a rise in Big Data following a boom in volume, velocity and variety of data that could be used for research. There has been improvement in infrastructures, network and standard, under the umbrella of Open Data and IoT, that makes data more and more accessible for use.

In the last decade, there has been growing discoveries in this domain like Google's Alpha Go, Open AI's Cha GPT, DeepMind's Alpha Fold, to list a few. Newer machine learning approaches to solving complex problems are also being developed.

The popularity of Machine Learning is fuelled by:

1) Its applicability diverse domains: In diverse areas of life like health, education, weather, to list a few, ML can be used to understand hidden patterns to solve complex problems.

2) The abundance of data: More and more data is available for model training with the advancement of IoT and other easy means of data collection.

3) Technological Advancement: The development of CPUs and CPUs with great computing power has made ML become even more popular.

2.2 Give three examples of machine learning techniques that can be viewed as either supervised or unsupervised approaches.

Classification: This is a supervised machine learning technique that predicts the category or label of a sample input. To arrive at this stage, the model is trained with labelled or classified data.

Clustering: This is an unsupervised learning technique where the model groups together data of a given input based on feature similarity. As it is unsupervised, the algorithm is not fed with labelled data.

Linear regression: This is a supervised machine learning technique where and algorithm models the relationship between multiple values and fits are linear equation to it.

2.3 What is the difference between classification and regression?

A classification model would be able to identify to what category a given data belongs to (categorical output) while a regression model would help predict a value based on given input (numerical output). For example, a classification model will give us information as to weather an email is spam or not while a regression model will give us a forecast of the price of a given commodity in each period.

2.4 What is the difference between supervised learning and unsupervised learning?

In supervised learning, the model is fed with labelled data in which the desired output is known, and the model learns on it while in unsupervised learning, the model works solely on features. Common supervised learning algorithm include linear regression and neural networks while some unsupervised learning algorithms include the K-means, Principal component Analysis (PCA).

2.5 Give examples of successful applications of machine learning and explain what technique is appropriate and what type of learning is involved?

- Natural Language Processing: This is the application of ML to understand, generate, and interpret human language. This uses a supervised learning technique based on recurrent neural networks.

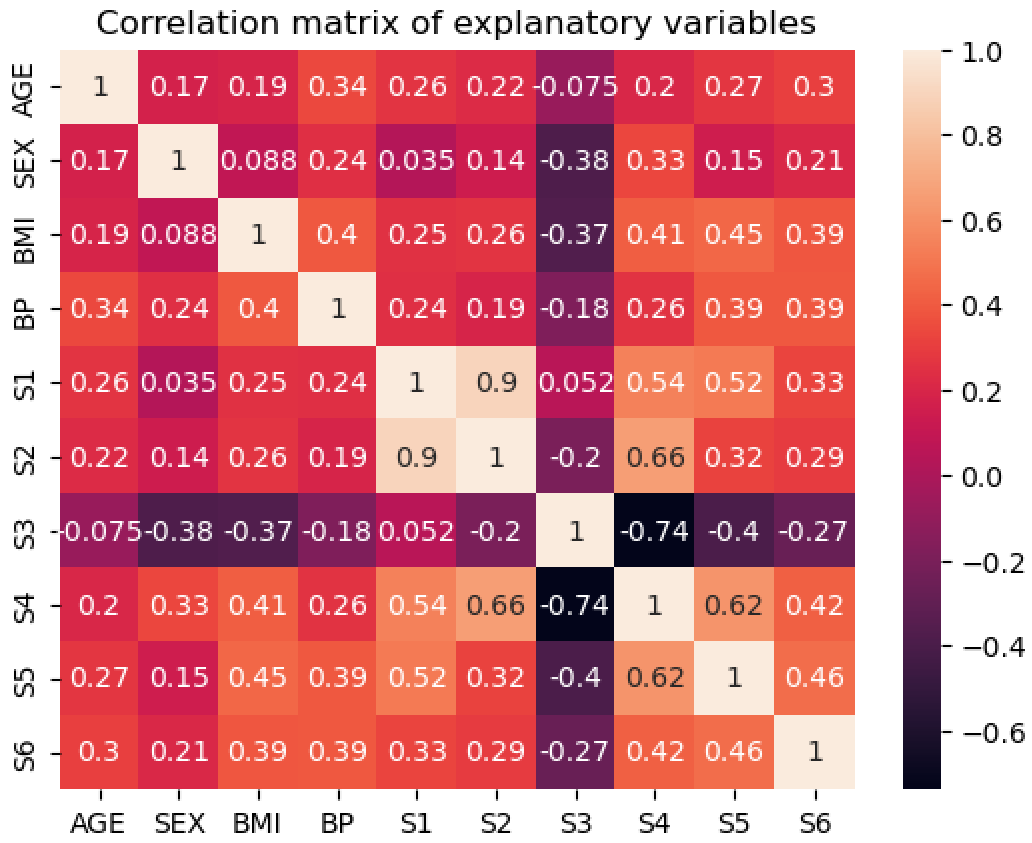
- Recommendation Systems: These are application of ML that are used to recommend special content or product to users. It is being employed by big tech firms like Netflix and Amazon. You can also find recommender system application on E-commers websites. It is based on a collaborative filtering and mix factorisation technique and uses unsupervised learning.

- Auto pilot algorithms: Now, we have self driving cars and other unmanned vehicles. The technic used to achieve this is convolutional neural networks and Deep learning. It also uses supervised learning.

3. Diabetes data (25 points)

3.1 Correlation Metrix

The correlation matrix shows the correlation between the explanatory variables, with values ranging from -1 to +1. A +1 indicates a high positive correlation, while a -1 indicates a high negative correlation. A 0 indicates an absence of correlation between the variables. Using the pandas *.corr()* function, we observe the following correlation between the variables.



Correlation matrix of explanatory variables

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6
AGE	1	0.17	0.19	0.34	0.26	0.22	-0.075	0.2	0.27	0.3
SEX	0.17	1	0.088	0.24	0.035	0.14	-0.38	0.33	0.15	0.21
BMI	0.19	0.088	1	0.4	0.25	0.26	-0.37	0.41	0.45	0.39
BP	0.34	0.24	0.4	1	0.24	0.19	-0.18	0.26	0.39	0.39
S1	0.26	0.035	0.25	0.24	1	0.9	0.052	0.54	0.52	0.33
S2	0.22	0.14	0.26	0.19	0.9	1	-0.2	0.66	0.32	0.29
S3	-0.075	-0.38	-0.37	-0.18	0.052	-0.2	1	-0.74	-0.4	-0.27
S4	0.2	0.33	0.41	0.26	0.54	0.66	-0.74	1	0.62	0.42
S5	0.27	0.15	0.45	0.39	0.52	0.32	-0.4	0.62	1	0.46
S6	0.3	0.21	0.39	0.39	0.33	0.29	-0.27	0.42	0.46	1

- We observe the greatest positive correlation between the blood serum measurements S1 and S2, and the greatest negative correlation is between the blood serum measurements S3 and S4.
- On the wider scale, S3 demonstrates negative correlation with other serum measurements.

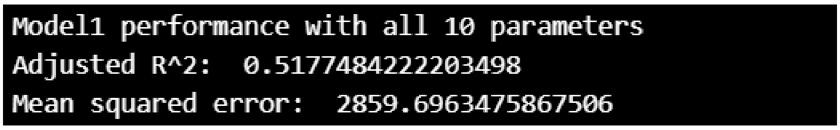
3.2 Coliniarity

Collinearity between predictors refers to a situation where the predictors are highly correlated. This can be identified when the calculated correlation between them tend to 1. In this exercise, we identify S1 and S2 as colinear features since their correlation, 0.9, is very close to 1.

Collinearity between features can cause high instability in their estimated coefficient value. This makes it highly sensitive to very small variations and can cause the model not to be generalisable.

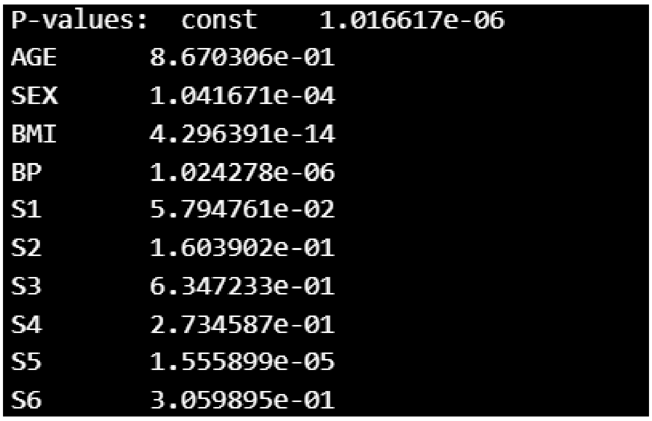
3.3 Model 1

Using the linear regression from the Sklearn linear model, Model1 is built. We obtain the following MSE and R-squared for Model1 as seen bellow. The and R-squared is acceptable as very high values could be indications of overfitting.



```
Model1 performance with all 10 parameters
Adjusted R^2:  0.5177484222203498
Mean squared error:  2859.6963475867506
```

To analyse the significance of every feature, can compute the p-values and values bellow the set 0.05 value, we will consider them as insignificant. Bellow are the p-values of all features.



```
P-values:  const     1.016617e-06
AGE      8.670306e-01
SEX      1.041671e-04
BMI      4.296391e-14
BP       1.024278e-06
S1       5.794761e-02
S2       1.603902e-01
S3       6.347233e-01
S4       2.734587e-01
S5       1.555899e-05
S6       3.059895e-01
```

From these results, we notice that not all features are significant. Also, this can partly be due to collinearity. Of features that demonstrate high collinearity, one of them can be drop since their variation is already represented by the collinear variable.

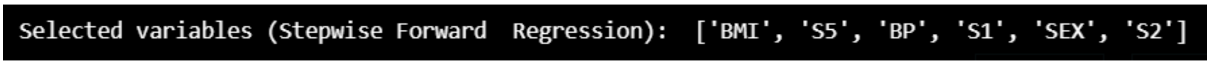
3.4 Forward Vs Backward Selection

These are methods use in selecting features for a model. In forward selection, we begin with an empty list of variables. Any feature which significantly adds to the overall R-squared value of a model is added to the list of variables, starting from the most significant. Any insignificant feature is left out. On the other hand, with backward selection, stepwise, we remove variables that don’t significantly contribute to the overall performance of the model till the point. These selection criteria yield different set of variables as variables tend to influence each other when in a model.

3.5 The Stepwise Approach

In the approach, a variable added or removed from the list of selected features is subject to scrutiny and can be removed or added respectively, steps down the process. For example, if a feature is added because its large individual power, its performance is evaluated again in the model and if its contribution is insignificant, it can be removed.

Using stepwise forward regression, we obtained the following features:



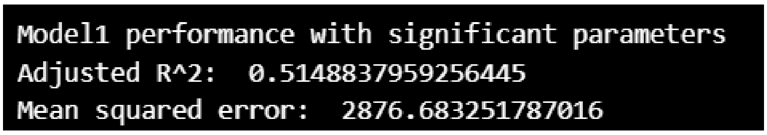
```
Selected variables (Stepwise Forward  Regression):  ['BMI', 'S5', 'BP', 'S1', 'SEX', 'S2']
```

The stepwise function works as such:

- In stepwise forward regression, the variable based on their individual strength are added one by one to the list, starting with that of greatest strength.
- At each step, all variables are evaluated on their individual contribution to the overall model.
- Feature that does not significantly contribute to the overall model are removed and the cycle continuous.

Basically, a feature can appear significant but when added to the model with other features, its contribution might be minute. In this case, the feature is dropped.

The Mean Squared Error and R-squared value for this new model was found to be:



```
Model1 performance with significant parameters
Adjusted R^2:  0.5148837959256445
Mean squared error:  2876.683251787016
```

The R-squared value for the new model is almost the same as that of the model with all parameters. It is lower than Model1 by 0.0029. The MSE is of this new model slightly higher than that of Model one by a factor of 16. On a broader scale, both models are similar.

Conclusion

The model based on the best features from stepwise regression is roughly as productive as the model will all 10 parameters. However, this model is more parsimonious as it has fewer parameters but equally as productive.

4. Analyzing the Titanic data set (25 points)

4.1 Logistic Versus Linear Regression

They are both methods of supervised learning but differ in that logistic regression is used to predict binary outcomes while linear regression is used to predict continuous outcome. In this example, we want to predict survival or not based on the parameters. Other examples and application of logistic regression is in healthcare when we want to know weather or not a drug is effective. Some example scenarios when linear regression is applied includes when we are trying to predict stock prices or what GPA a student will have given the parameters.

4.2 Probability of Survival of a Passenger

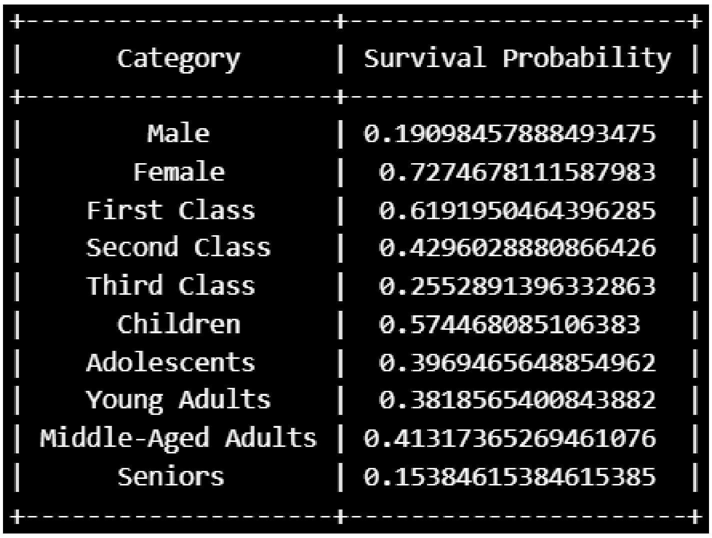
To calculate the survival probability of a passenger:

- The passengers were grouped by gender, passenger class, and age group.
- For each group, passengers were broken down into Male, Female for gender, 1st, 2nd and 3rd class for passenger class, and Children (0 - 12years), Adolescents (13 – 19years), Young Adults (20 – 34years), Middle Aged Adults (35 – 64) and Seniors (65 +)
- For each category, the probability of survival was calculated by dividing the number of survivors by the total number of people in each category. In each category:

$$P(survival) = \frac{Number\ of\ survivors}{Number\ of\ parsengers}$$

4.3 Probability table

The following result was obtained:



```
+--------------------+----------------------+
|      Category      | Survival Probability |
+--------------------+----------------------+
|        Male        | 0.19098457888493475  |
|       Female       |  0.7274678111587983  |
|    First Class     |  0.6191950464396285  |
|    Second Class    |  0.4296028880866426  |
|    Third Class     |  0.2552891396332863  |
|      Children      |  0.574468085106383   |
|    Adolescents     |  0.3969465648854962  |
|    Young Adults    |  0.3818565400843882  |
| Middle-Aged Adults | 0.41317365269461076  |
|      Seniors       | 0.15384615384615385  |
+--------------------+----------------------+
```

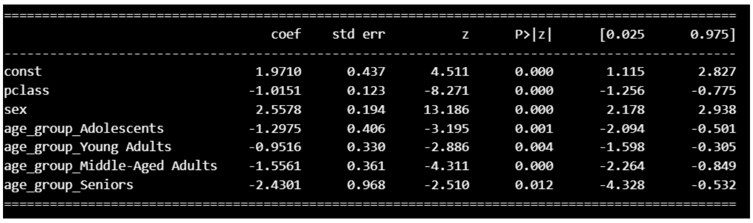
It is observed that being a female, one had the greatest chance of survival followed by being in first class, then being a child. The least probability of survival was in the Male category. On can obtain the combined probability by multiplying categories together.

4.4 Logistic Regression Model

To arrive at the logistic model the following was done:

- Data was imported
- Rows with missing values were removed using .dropna pandas
- The Sex column was mapped numerical values (0 for male and 1 for female).
- Also, the 'age_group' variable was transformed into numerical format using one-hot encoding to create binary columns for each age group.
- The important features were selected: 'pclass', 'sex', and the one-hot encoded age group columns
- The target variable was set to ‘Survived’ (1 indicates for survived and 0 not survived)
- The data was split (75% for training and 25% for testing)
- The model was trained using LogisticRegression() and evaluated using the confusion matrix and the accuracy was calculated.

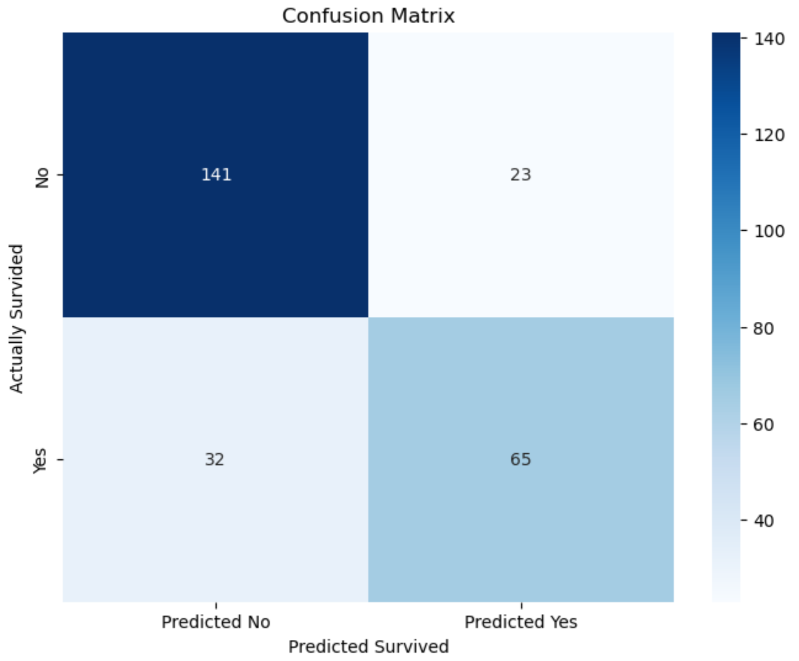
The statistically significant parameters include the Age group, Sex, Passenger class and the Constant. A summary of the parameters is here bellow.



```
=====================================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                         1.9710      0.437      4.511      0.000       1.115       2.827
pclass                       -1.0151      0.123     -8.271      0.000      -1.256      -0.775
sex                           2.5578      0.194     13.186      0.000       2.178       2.938
age_group_Adolescents        -1.2975      0.406     -3.195      0.001      -2.094      -0.501
age_group_Young Adults       -0.9516      0.330     -2.886      0.004      -1.598      -0.305
age_group_Middle-Aged Adults -1.5561      0.361     -4.311      0.000      -2.264      -0.849
age_group_Seniors            -2.4301      0.968     -2.510      0.012      -4.328      -0.532
=====================================================================================================
```

4.5 What is the performance of the model, measured by classification accuracy (number of correct classifications divided by total number of classifications) based on confusion matrix?

Using Sklearn metrics following confusion matrix for the system was obtained, and the heat map was plotted using seaborn.

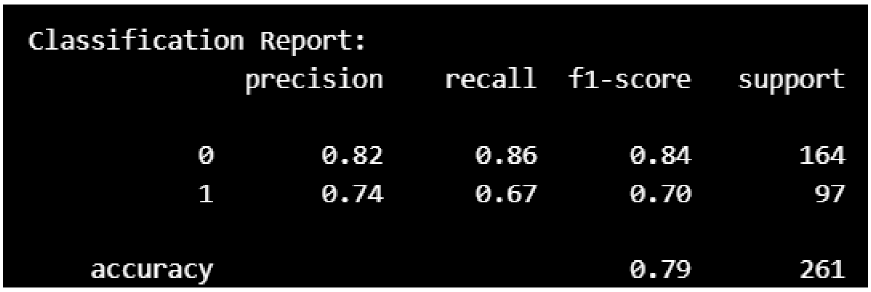


The accuracy of the model is given by:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{65 + 141}{65 + 23 + 141 + 32} = 0.79$$

The Model is therefore 79% accurate at predicting the fate of a passenger based on Gender, Passenger class and Age group.

The other qualitative attributes of our model were obtained as seen bellow, where 1 corresponds to survived (Positive) and 0 corresponds to Negative:



```
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.86      0.84       164
           1       0.74      0.67      0.70        97

    accuracy                           0.79       261
```

Therefore, or model is 82% good at predicting deaths and 74% good at predicting survival cases.

Conclusion:

In this exercise, we were able to apply logistic regression to determine the fate of passengers who boarded the Titanic ship, based on their Gender, Passenger class, and Age group. We use 75% of the dataset in training our model using SkLearn Linear Logistic model. The resulting model had an acceptable accuracy of 79% when tested on the 25% of data set. The model was better off in predicting victims that did not survive to victims that survived. Though no bench marking was done, the results are acceptable.

[1] “pandas documentation — pandas 2.2.2 documentation.” Accessed: Sep. 02, 2024. [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/index.html

[2] “NumPy -.” Accessed: Sep. 02, 2024. [Online]. Available: https://numpy.org/

[3] “Matplotlib documentation — Matplotlib 3.9.2 documentation.” Accessed: Sep. 02, 2024. [Online]. Available: https://matplotlib.org/stable/

[4] A. R. V, *AakkashVijayakumar/stepwise-regression*. (Mar. 27, 2024). Python. Accessed: Oct. 15, 2024. [Online]. Available: https://github.com/AakkashVijayakumar/stepwise-regression