CARNEGIE MELLON UNIVERSITY - AFRICA

# DATA, INFERENCE & APPLIED MACHINE LEARNING

(COURSE 18-785)

Professor Patrick McSharry

**Assignment 5**

MUNYANEZA Kenny Roger

07th November 2023

# Introduction

This is a final report designed as a fulfillment of the requirements to complete assignment 5 DIAML.

# Question 1

## 1. Statistical Learning

### 1.1

### Steps to implement a rule-based approach

- Identify criteria associated with each decision.
- For each criterion, create decision rules or conditions that determine the outcome of the decision. These are sometimes referred to as If-Then statements based on expert or literature knowledge.
- Test and validate the decision rules using historical data to ensure that the decision rules accurately reflect the desired decision outcomes.
- Implement the rule-based decision system using a chosen technology such as a software application.
- Monitor the efficiency of the rule-based decision system and update the rules if necessary.

### Example:

### Admission to a semester at CMU Pittsburgh for CMU Africa students

1. Identify decision criteria:
    - Current GPA
    - Core units completed
    - Total units studied
    - Expected graduation year
2. Creating decision rules:
    - If current GPA is greater than or equal to 3.5, and core units completed equal to 60 and total units studied less than or equal to 108 and expected graduation year is 2024 then grant the admission.
    - If current GPA is less than 3.5, or core units completed not equal to 60 or total units studied greater than 108 or expected graduation year not 2024 then deny the admission.

3. Validate decision rules:
    - Validate the admission decision rules against the admission decision made in the past and adjust on the rules where necessary.
4. Implement the rule-based decision system:
    - Create an online application system where applicants submit their academic information and essays. The system automatically processes the data and provides admission for a semester at CMU Pittsburgh decisions based on the predefined rules.
5. Monitoring and Maintenance:
    - Regularly monitor the performance of the admissions system and add new rules governing the admission for a semester at CMU Pittsburgh as the University policies change.

## Is any domain knowledge required to establish a rule?

Yes, Domain knowledge is required because rules are established from the knowledge acquired from literatures and experts of a given domain. This knowledge is then applied in a situation to make more effective rule-based decision. [1] [2]

## 1.2

## Over-fitting

Overfitting occurs when a model is excessively complex and fits not only the underlying patterns in the data but also captures random noise and fluctuations. This fitting of noise is a problem to statistical learning because it makes the model learn from the noise and this makes the model less accurate in generalizing unseen data.

An Over-fitted model performs better on training data than the testing data. [3]

## Choice of model

Following the law of Parsimony, I would choose a simple model with one parameter rather than a complex model with 10 parameters. This is because even though complex models produce a better fit of the observations, a complex model with many parameters will not distinguish between the necessary dynamics to be extracted and fluctuations due to measurement errors, non-stationarity, and noise.[3]

## 1.3

## Approaches to avoid over-fitting

1. **Cross-Validation:** This approach ensures that the historical performance of the model can be achieved in practice. This is done by using a walk-forward approach and making sure that the model is constructed using data in the time window [t-Ttrain, t] and evaluated in the time window [t, t+Ttest].
2. **Information Criteria:** This approach penalizes the model's complexity and selects a model which is parsimonious to provide a balance between complexity and goodness of fit.

## 1.4 [4]Metrics used to evaluate the performance of a model.

1. **Mean Squared Error (MSE):** For regression models this metric calculates the average of the squared difference between the target value and the value predicted by the model.

MSE is given by the formula:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

Where:

N is the number of data points.

$y_i$ represents the actual (observed) value for the i-th data point.

$\hat{y}_i$ represents the predicted value for the i-th data point.

**Note:** Lower MSE indicates a better model.

**Application of MSE**

**Example 1:** A model to forecast the number of monthly traffic accidents in a year.

The performance of this model can be assessed, whereby the average of the squared difference between actual and forecast monthly accidents would be used to determine how the model performs. Lower MSE would indicate that the model has a better performance.

| Month | Actual Accidents | Forecasted Accidents | Error | Squared Error |
|---|---|---|---|---|
| 1 | 23 | 19 | 4 | 16 |
| 2 | 9 | 11 | -2 | 4 |
| 3 | 10 | 10 | 0 | 0 |
| 4 | 18 | 15 | 3 | 9 |
| 5 | 21 | 19 | 2 | 4 |
| 6 | 32 | 33 | -1 | 1 |
| 7 | 27 | 25 | 2 | 4 |
| 8 | 16 | 16 | 0 | 0 |
| 9 | 25 | 25 | 0 | 0 |
| 10 | 9 | 7 | 2 | 4 |

| | | | | |
|---|---|---|---|---|
| 11 | 10 | 10 | 0 | 0 |
| 12 | 23 | 24 | -1 | 1 |

**MSE =** 43/12 = 3.58

The MSE value above shows that the model has moderate performance in predicting monthly traffic accidents.

2. **Accuracy:** For classification models, accuracy is simple performance metric used, and it refers the number of correct predictions to the total number of predictions.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ predictions}$$

**Application of Accuracy**

A classification model to predict whether a patient is infected or not by a particular disease. It would be recommended to use Accuracy to assess the performance of the model when the target classes in data are approximately. For example, if 43% of the patients in the dataset are in the infected class and 57% of the patients in the dataset are in the non-infected class. However, It would not be recommended to use the Accuracy when the target variable is dominated by one classification. [4]

## 1.5 Benchmarks

Benchmarks are established as levels of forecast performance that can be readily attained without needing complex mathematical models. [3]

Examples:

- Median of observation is a benchmark for forecasting
- Average of observation is a benchmark for forecasting.

# Question 2

## 2. Machine Learning

### 2.1 Definition of Machine Learning and its evolution

**Machine learning** refers to the discovery of knowledge from data and experience.

**Below is the evolution of Machine Learning from 1940s to 2023:**

The evolution of machine learning can be traced through distinct phases over the years. In the early stages from 1950 to 1980, machine learning was primarily a laboratory endeavor. Notable milestones included Alan Turing's Turing Test in 1950, Arthur Samuel's coining of the term "Machine Learning" in 1952, and the development of the first machine learning programs for simple tasks. These programs aimed at improving computer performance in games and recognizing basic patterns. The 1970s introduced symbolic AI and decision trees, marking the start of modern AI and machine learning research. Decision tree algorithms, such as the ID3 algorithm, became significant. The 1980s brought the emergence of neural networks and the backpropagation algorithm. The 1990s saw the rise of support vector machines and reinforcement learning. As we entered the 2000s, big data transformed the landscape, and ensemble learning methods became prevalent. Then came the deep learning revolution in the 2010s, with the triumph of deep neural networks in various domains. The 2020s continue this progress but also present new challenges, including model interpretability, ethics, data privacy, and environmental impact, which shape the future of machine learning.[5] [6]

**Why is Machine Learning popular?**

Machine learning has gained immense popularity for several key reasons. First and foremost, it thrives on data, much like a dam relies on water. In today's world, an abundant supply of data is readily available, be it on the cloud, social media, or various online platforms. The availability of massive datasets has made it possible to train machine learning models effectively.

Moreover, the development and adoption of General-Purpose Graphics Processing Units (GPGPUs) have significantly accelerated machine learning processes. GPGPUs act as the robust steel pipes that channel data efficiently and can withstand the pressure, thanks to their parallel processing capabilities, ideally suited for the linear algebra calculations integral to machine learning.

Cloud computing has also played a pivotal role in the popularity of machine learning. It provides accessible and scalable computing power to users, allowing them to execute machine learning tasks without the need to invest in and maintain specialized hardware. Cloud computing's ease of provisioning and pay-as-you-go model has democratized machine learning, making it accessible to a broader audience.

Lastly, the continuous refinement and innovation of machine learning algorithms have been a driving force. These algorithms are the core of machine learning, enabling systems to learn and make predictions from data efficiently. The availability of open-source algorithms and the confluence of data, processing power, and improved algorithms have collectively propelled the widespread adoption of machine learning.[7]

## 2.2 Examples of Machine Learning Techniques

- **Regression:** Employed to make predictions of continuous values, such as estimating prices.
- **Classification:** Applied to make binary class distinctions, like determining whether an animal falls into the cat or dog category.
- **Clustering:** Involves assigning labels by gathering similar data into groups based on shared characteristics, for example, organizing music into genres based on their attributes.[8]

## 2.3 Difference between Classification and Regression

Regression Algorithms are utilized when dealing with continuous or real-valued output variables, aiming to map input values (x) to these continuous outputs (y). In contrast, Classification Algorithms are employed when working with discrete output variables, mapping input values (x) to these discrete outputs (y). Regression Algorithms find optimal fit lines or functions to make accurate predictions. They are typically applied to problems like Weather Prediction and House Price Estimation. On the other hand, Classification Algorithms seek decision boundaries to segregate data into different classes and are well-suited for tasks such as Identifying Spam Emails, Speech Recognition, and Detecting Cancer Cells. Regression Algorithms can be further categorized into Linear Regression for linear relationships and Non-linear Regression for non-linear patterns. Classification Algorithms can be classified as Binary Classifiers for two-class problems and Multi-class Classifiers for those involving more than two classes.[9]

## 2.4 Difference between Supervised and Unsupervised Learning

Supervised learning relies on labeled data to train algorithms and provides direct feedback to verify predictions. This approach is used in cases where both input and corresponding outputs are known, typically involving Classification and Regression problems. The goal of supervised learning is to produce accurate predictions when new data is presented, making it less aligned with true artificial intelligence.

Unsupervised learning, conversely, utilizes unlabeled data to unveil hidden patterns within datasets and gain useful insights. It operates without the need for direct feedback and is applicable when only input data is available. It encompasses algorithms such as Clustering, K-Nearest Neighbors (KNN), and the Apriori algorithm. Unsupervised learning is considered closer to true artificial intelligence, as it learns similarly to how a child learns from daily experiences.[10]

## 2.5 Successful Machine Learning applications

**Image Recognition application**

**Technique used:**

Image recognition systems, especially in applications like face recognition, involve classification learning. These systems classify images into specific categories or labels. For example, they can classify an image as “face” or “not face,” making it a binary classification task.

**Type of learning used:**

In image recognition, supervised learning is commonly used. This involves training a model to categorize images into predefined classes or labels. For instance, training a model to recognize different objects in images, such as cats, dogs, cars, and more, based on labeled training data.

**Stock Market Trading application**

In stock market trading, the goal is often to predict the future price or performance of stocks, commodities, or financial assets.

**Technique Used**

This involves regression learning because it focuses on predicting a continuous variable, such as the future price of a stock.

**Type of learning used**

The process is supervised because the models are trained using historical data where both features and the target variable (future stock prices) are known.

**Virtual Try On application**

In Virtual Try On systems, the primary goal is to provide users with a virtual experience of trying on products like glasses, sunglasses, or other accessories.

**Technique used**

Clustering technique of Machine Learning is used, whereby each cluster represents a group of users with similar facial features. These clusters can be used to personalize product recommendations or to enhance the virtual try-on experience. For instance, users within a cluster may have similar face shapes, and specific products may be recommended to them.

**Learning type used**

Unsupervised learning algorithms, such as k-means or hierarchical clustering, are applied to the facial feature data. These algorithms group users based on similarities in their facial characteristics.

# Question 3

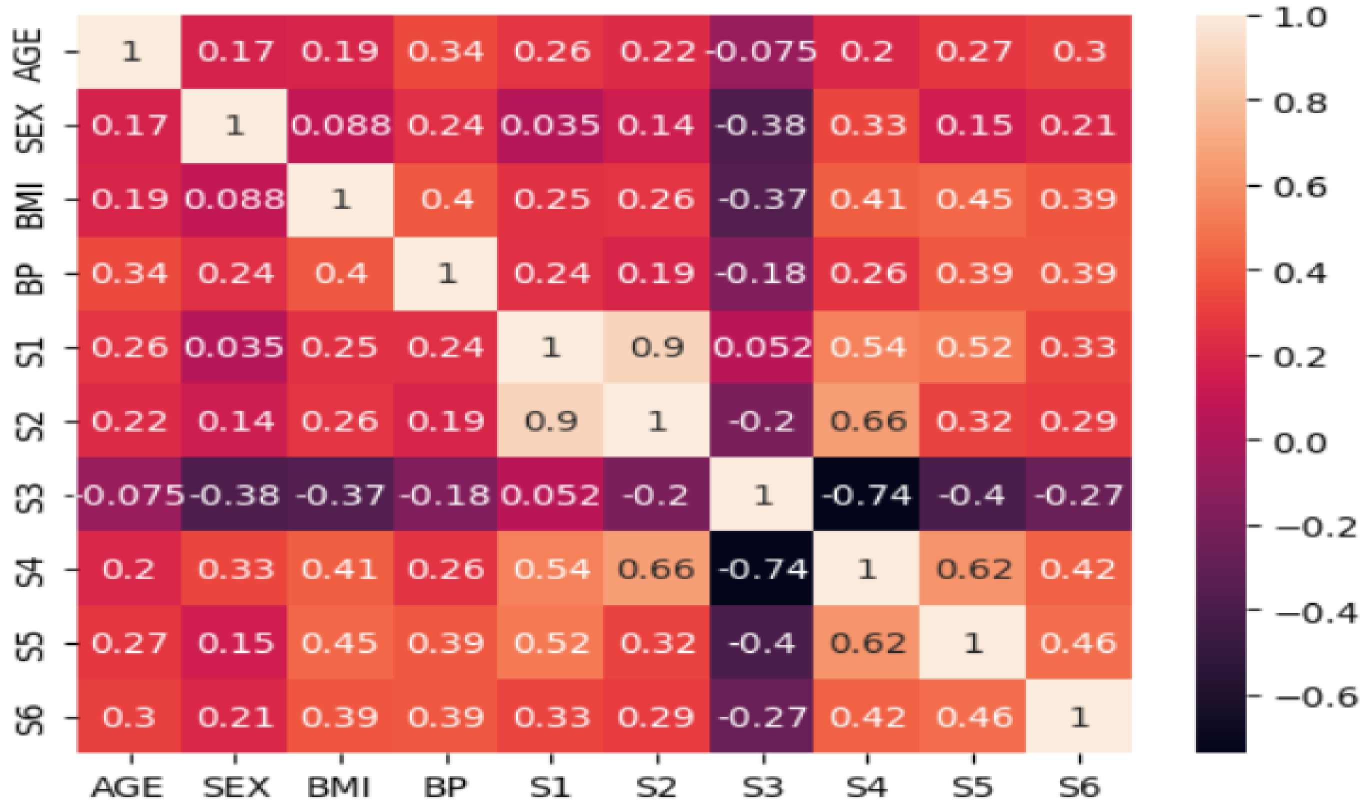
## Technologies Used

- **pandas**: It was used to read the diabetes dataset.
- **Seaborn:** Used display statistical graphics.
- **Statsmodels:** Used to calculate a logistic regression model.
- **Stepwise_regression:** Used to perform forward regression in this question.

## Implementation

I started by reading the diabetes dataset using pandas, and I separated the predictor variables from the target variable into two separate dataframes. Then I applied the "corr()" function of pandas to compute the correlation heat map matrix of predictor variable.

## 3.1 Correlation Matrix obtained




### 3.1.1 The relationship between predictor variables

The relationship between variables tends to be small for some variables like BMI and Sex or S1 and S3, which is a good indicator of isolation of the variables from each other. This makes the variables good predictors of the target variable. However, other variables like S1 and S2 have a very strong correlation coefficient, which indicates a high risk of collinearity between the variables. This would make each of the two predictors a bad predictor in the presence of the other.

## 3.2 Collinearity

### 3.2.1 Definition

Collinearity arises when the independent variables used in constructing a regression model exhibit correlation with each other. As the term implies, an independent variable is ideally meant to be isolated from other independent variables. Collinearity makes it challenging to identify the individual contributions of each variable to the dependent variable and can affect the model's performance. [11]

### 3.2.2 Effect of collinearity on the estimated coefficient value

- Coefficient estimates for independent variables become highly responsive to even small model changes. For instance, adding or removing a single independent variable can lead to significant fluctuations in these estimates, making it challenging to discern the precise impact of each independent variable.

- Collinearity increases the variance and standard error of coefficient estimates. Consequently, it diminishes the model's reliability, rendering the p-values less trustworthy as indicators of the statistical significance of independent variables for the model.

Collinearity's effects on the model include the instability of coefficient estimates and a reduction in the model's reliability, affecting our ability to gauge the significance of independent variables accurately.[11]

## 3.3 Creating a multivariate linear model using all ten variables and a constant

Using Ordinary Least Squares (OLS) is a statistical method of regression analysis, I fitted a multivariate model with the ten predictor variables and a constant. Then, I computed the Mean Squared Error using the function **mse_resid()** and adjusted $R^2$ of the model using **rsquared_adj()** function of statsmodel module.

```
Model1 Mean Squared Error (MSE): 2932.6816372003336
Model1 R-squared (R²): 0.5177484222203498
Model1 adjusted R-squared (R²): 0.5065592904853231
```

Inferences on the results:

- Model1 has some predictive power (R-squared value of 0.5177), indicating that the independent variables in the model provide a moderate level of explanation for the variance in the dependent variable.
- The adjusted R-squared value (0.5066) suggests that Model1 may benefit from some refinement, possibly by removing or optimizing certain independent variables to avoid overfitting.
- The relatively high MSE indicates that the model's predictions are, on average, relatively far from the actual data points by 2932.68. This implies that there may be room for improving the model's accuracy in predicting the dependent variable.

### 3.3.1 Significance of the predictor variables

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.518
Model:                            OLS   Adj. R-squared:                  0.507
Method:                 Least Squares   F-statistic:                     46.27
Date:                Sun, 05 Nov 2023   Prob (F-statistic):           3.83e-62
Time:                        12:48:10   Log-Likelihood:                -2386.0
No. Observations:                 442   AIC:                             4794.
Df Residuals:                     431   BIC:                             4839.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -334.5671     67.455     -4.960      0.000    -467.148    -201.986
AGE           -0.0364      0.217     -0.168      0.867      -0.463       0.390
SEX          -22.8596      5.836     -3.917      0.000     -34.330     -11.389
BMI            5.6030      0.717      7.813      0.000       4.194       7.012
BP             1.1168      0.225      4.958      0.000       0.674       1.560
S1            -1.0900      0.573     -1.901      0.058      -2.217       0.037
S2             0.7465      0.531      1.406      0.160      -0.297       1.790
S3             0.3720      0.782      0.475      0.635      -1.166       1.910
S4             6.5338      5.959      1.097      0.273      -5.178      18.245
S5            68.4831     15.670      4.370      0.000      37.685      99.282
S6             0.2801      0.273      1.025      0.306      -0.257       0.817
```

Not all predictor variables for the target variables because variables like AGE, S1, S2, S3, S4, and S6 have a p-value greater than 0.05, which makes them insignificant.

The presence of insignificant variables is due to the collinearity between some of the predictors such as S1 and S2.

## 3.4 Difference between forward selection and backward selection

**Forward Selection:**

Definition: Forward selection is a feature selection method used in regression and machine learning, where the process begins with an empty set of features, and features are incrementally added to the model based on their contribution to model performance.

Process: It starts with no features in the model and iteratively evaluates and selects the feature that results in the best improvement in model performance (e.g., reduced error or increased R-squared).

Iteration: The process continues until a stopping criterion is met, such as reaching a predetermined number of features or when no further improvement can be achieved.[12]

**Backward Selection:**

Definition: Backward selection is another feature selection method in regression and machine learning, where the process begins with all available features included in the model, and features are incrementally removed from the model based on their contribution to model performance.

Process: It starts with all features in the model and iteratively evaluates and removes the feature that contributes the least to the model's performance (e.g., has the highest p-value).

Iteration: The process continues until a stopping criterion is met, such as retaining a specific number of features or when further removals do not significantly affect the model's performance.[12]

### 3.5.1 How the approach stepwise work in the sense of selecting variables

Stepwise variable selection offers several approaches for choosing variables to include in a model:
**General to Specific (Backward Selection)**: This method begins with all predictor variables and eliminates them one by one.

**Specific to General (Forward Selection)**: It starts with no variables and adds predictor variables one by one.

In both cases, the process stops when the optimal model fit is achieved. The algorithm stops when no new variables are added, or when a variable is removed immediately after being added. Threshold p-values are required for adding variables.[3]

### 3.5.2 Fitting the model using forward selection

I started by determining the most useful variables using the forward_selection() function of the stepwise_regression module. Then, with the resultant useful variables and a constant, I fitted a multivariate linear model.

**Useful variables:**

```
Using stepwise function, the best predictors are:  ['BMI', 'S5', 'BP', 'S1', 'SEX', 'S2']
```

The result above mentions the useful variables:

- BMI
- S5
- BP
- S1
- SEX
- S2

### 3.5.3 How the forward_regression function works

Forward regression is an iterative feature selection function that begins with an empty set of features and gradually adds them to the regression model one by one. In this process, the initial feature selected for inclusion is the one with the highest correlation to the target variable. This selection is based on the F-statistic, where the feature producing the largest F-statistic, indicating its significance in the model, is chosen. The F-statistic, or F-test, is a statistical measure used in regression analysis to assess the overall fit and significance of a regression model. It is often used to compare the fit of different models and determine whether the inclusion of certain predictor variables significantly improves the model's performance. Subsequent features are added using the same criteria, i.e., based on their correlation and F-statistic. This process continues until the F-statistic surpasses a predefined F-value threshold known as "F-to-enter." If this threshold is met, the procedure stops; otherwise, it continues. The result is a model that includes a subset of features deemed significant in explaining the target variable's variation.[13]

### 3.5.4 Accuracy of the forward selection model

```
Stepwise model's Mean Squared Error (MSE): 2922.9747064134735
Stepwise model's R-squared (R²): 0.5148837959256445
```

- The new model's MSE of 2922.97 is very close to the MSE of Model1, which was 2932.68. Both models have similar levels of error in their predictions.
- The new model's R-squared value of 0.5149 is also similar to the R-squared value of Model1, which was 0.5177. Both models explain approximately the same amount of variance in the dependent variable.

# Question 4

## Technologies Used

- **pandas**: It was used to read the titanic dataset.
- **Statsmodels:** Used to create a logistic regression model.
- **Stepwise_regression:** Used to perform forward regression in this question
- **Scikit-Learn:** It provides functions to determine confusion matrix and accuracy score.

## 4.1 Difference between Linear regression and Logistic regression

Linear regression and logistic regression are two distinct techniques used in the field of predictive modeling. The primary difference between them lies in the nature of the dependent variable they aim to predict and the type of problems they are suited for. Linear regression is employed when the task involves predicting a continuous dependent variable, such as price or age, by utilizing a set of independent variables. It is typically applied to regression problems, seeking to find the best-fitting line that allows for accurate value prediction. In contrast, logistic regression is tailored for classification problems, where the goal is to predict categorical outcomes, like binary responses (e.g., 0 or 1, Yes or No). Instead of estimating a linear relationship, logistic regression identifies an S-curve that helps classify samples. These two methods diverge further in their estimation techniques, with linear regression using the least square estimation for accuracy, while logistic regression employs the maximum likelihood estimation method. Additionally, linear regression assumes a linear relationship between variables, allowing for collinearity between independent variables, whereas logistic regression does not require linearity and should not have collinearity among the independent variables.[14]

<u>Difference in formulas used for the regression models</u>

To summarize the relationship between predictor variables and the target variable,

Linear regression uses the formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$ [15]

Whereby,

- Y: The response variable
- Xj: The jth predictor variable
- βj: The average effect on Y of a one unit increase in Xj, holding all other predictors fixed

Whereas Logistic regression uses the formula:

$$p(X) = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p} / (1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p})$$

[15]

## 4.2 Probability of survival of a passenger on the titanic

The probability of a person to survive on the titanic is given by the ratio of the number of people who survived to the total number of people on the titanic. I computed that probability by reading the dataframe of titanic using pandas and dividing the length of the dataframe of people that survived by the length of the original dataframe.

### Result obtained:

```
Probability of a person to survive on the titanic is: 0.38
```

## 4.3 Table showing the probability of survival, broken down by passenger class, gender and age

The probability table required in this question will be showing the conditional probability of survival for passengers in the titanic under various conditions. Conditional probability table was derived from the contingency table summarizing the raw data provided produced using the “crosstab()” function from pandas module. A contingency table is a type of table that summarizes the relationship between two categorical variables.

From the contingency tables I calculated the different probabilities as a ratio of the total of a sub-group to the total passengers in a group. [16]

<u>Contingency table of passengers’ survival grouped by passenger class</u>

```
pclass      1    2    3   All
survived
0         123  158  528   809
1         200  119  181   500
All       323  277  709  1309
```

## Survival probabilities broken down by passenger class

| Probability | Value |
|---|---|
| Probability of survival for a passenger from passenger class 1 | 0.62 |
| Probability of survival for a passenger from passenger class 2 | 0.43 |
| Probability of survival for a passenger from passenger class 3 | 0.26 |

Table showing survival probability by passenger class

<u>Contingency table of passengers' survival grouped by gender</u>

| sex<br>survived | female | male | All |
|---|---|---|---|
| 0 | 127 | 682 | 809 |
| 1 | 339 | 161 | 500 |
| All | 466 | 843 | 1309 |

## Survival probabilities broken down by gender

| Probability | Value |
|---|---|
| Probability of survival for a male passenger | 0.73 |
| Probability of survival for a female passenger | 0.19 |

Table showing survival probability by gender

<u>Contingency table of passengers' survival grouped by age range</u>

| ageRange<br>survived | 0-25 | 26-50 | 51-75 | +75 | All |
|---|---|---|---|---|---|
| 0 | 255 | 307 | 57 | 0 | 619 |
| 1 | 188 | 201 | 36 | 2 | 427 |
| All | 443 | 508 | 93 | 2 | 1046 |

### Survival probabilities broken down by age range

```
+---------------------------------------------------------------------------+
|              Table showing survival probability by age range              |
+---------------------------------------------------------------------------+
|                            Probability                          | Value |
+-----------------------------------------------------------------+-------+
| Probability of survival for a passenger with age in range 0-25  |  0.42 |
| Probability of survival for a passenger with age in range 26-50 |  0.4  |
| Probability of survival for a passenger with age in range 51-75 |  0.39 |
| Probability of survival for a passenger with age in range +75   |  1.0  |
+-----------------------------------------------------------------+-------+
```

## 4.4 Logistic Regression model for survival rates based on passenger class, sex, and age

I used the “logit()” function of the statsmodel module to fit a logistic regression model with passenger class, sex and age as your explanatory variables and survived as the dependent variable.

```
                           Logit Regression Results
==============================================================================
Dep. Variable:               survived   No. Observations:                 1046
Model:                          Logit   Df Residuals:                     1042
Method:                           MLE   Df Model:                            3
Date:                Sun, 05 Nov 2023   Pseudo R-squ.:                  0.3051
Time:                        18:28:36   Log-Likelihood:                -491.51
converged:                       True   LL-Null:                       -707.31
Covariance Type:            nonrobust   LLR p-value:                 3.167e-93
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.0919      0.371      5.633      0.000       1.364       2.820
pclass        -1.1332      0.112    -10.142      0.000      -1.352      -0.914
sex            2.4974      0.166     15.033      0.000       2.172       2.823
age           -0.0339      0.006     -5.395      0.000      -0.046      -0.022
==============================================================================
```

### 4.4.1 Parameter estimates

Parameter estimates are passenger class, sex, and age, and a constant, and all of them are significant since their p value is less than the alpha level (0.05).

## 4.5 Performance of the model

To gauge the performance of the model, I started by predicting the dependent variable y for each provided instance in the dataset using the predict() and I passed this dataframe of predicted y variable and a dataframe of the actual y variable to the function “confusion_matrix()” to measure the classification accuracy of the model. The confusion_matrix() is provided by the “sklearn.metrics” module. [17]

## Result Obtained:

```
Confusion Matrix :
 [[523  96]
 [126 301]]
```

Interpretation of the confusion matrix :

- True Positives (TP) = 523. These are the cases where the model correctly predicted the positive class.
- True Negatives (TN) = 301. These are the cases where the model correctly predicted the negative class.
- False Positives (FP) = 96. These are the cases where the model incorrectly predicted the positive class when it was the negative class.
- False Negatives (FN) = 126. These are the cases where the model incorrectly predicted the negative class when it was the positive class.[18]

[1] 'Autoblocks: Full-stack monitoring, debugging, and testing'. Accessed: Oct. 30, 2023. [Online]. Available: https://www.autoblocks.ai
[2] 'What are rule-based decision tools and where are they used?', decision-making-confidence.com. Accessed: Oct. 30, 2023. [Online]. Available: https://www.decision-making-confidence.com/rule-based-decision-tools.html
[3] 'CMUdiaml6.pdf: Data, Inference, and Applied Machine Learning - DIAML'. Accessed: Oct. 30, 2023. [Online]. Available: https://canvas.cmu.edu/courses/36029/files/10095509?module_item_id=5605823
[4] 'Performance Metrics in Machine Learning - Javatpoint', www.javatpoint.com. Accessed: Oct. 31, 2023. [Online]. Available: https://www.javatpoint.com/performance-metrics-in-machine-learning
[5] 'History and Evolution of Machine Learning: A Timeline', WhatIs.com. Accessed: Oct. 31, 2023. [Online]. Available: https://www.techtarget.com/whatis/A-Timeline-of-Machine-Learning-History
[6] adjhbv67thgfj, 'The Evolution of Machine Learning in Business', TurinTech AI. Accessed: Oct. 31, 2023. [Online]. Available: https://www.turintech.ai/the-evolution-of-machine-learning-in-business/
[7] 'What is Machine Learning and Why is it Popular?', INE. Accessed: Nov. 01, 2023. [Online]. Available: https://www.ine.com/
[8] 'Getting started with using Visual Machine Learning Tools for building your Machine Learning Models', TECHCOMMUNITY.MICROSOFT.COM. Accessed: Nov. 01, 2023. [Online]. Available: https://techcommunity.microsoft.com/t5/educator-developer-blog/getting-started-with-using-visual-machine-learning-tools-for/ba-p/3578397
[9] 'Regression vs Classification in Machine Learning - Javatpoint', www.javatpoint.com. Accessed: Nov. 01, 2023. [Online]. Available: https://www.javatpoint.com/regression-vs-classification-in-machine-learning
[10] 'Supervised vs Unsupervised Learning - Javatpoint', www.javatpoint.com. Accessed: Nov. 01, 2023. [Online]. Available: https://www.javatpoint.com/difference-between-supervised-and-unsupervised-learning
[11] N. Rosidi, 'A Beginner's Guide to Collinearity: What it is and How it affects our regression model', Medium. Accessed: Nov. 03, 2023. [Online]. Available: https://towardsdatascience.com/a-beginners-guide-to-collinearity-what-it-is-and-how-it-affects-our-regression-model-d442b421ff95
[12] G. Choueiry, 'Understand Forward and Backward Stepwise Regression – QUANTIFYING HEALTH'. Accessed: Nov. 03, 2023. [Online]. Available: https://quantifyinghealth.com/stepwise-selection/
[13] 'Feature selection methods with Python', DataSklr. Accessed: Nov. 05, 2023. [Online]. Available: https://www.datasklr.com/ols-least-squares-regression/variable-selection
[14] 'Linear Regression vs Logistic Regression - Javatpoint', www.javatpoint.com. Accessed: Nov. 05, 2023. [Online]. Available: https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning
[15] Zach, 'Logistic Regression vs. Linear Regression: The Key Differences', Statology. Accessed: Nov. 05, 2023. [Online]. Available: https://www.statology.org/logistic-regression-vs-linear-regression/
[16] Zach, 'How to Calculate Conditional Probability in Python', Statology. Accessed: Nov. 05, 2023. [Online]. Available: https://www.statology.org/conditional-probability-in-python/

[17] 'Logistic Regression using Statsmodels', GeeksforGeeks. Accessed: Nov. 05, 2023. [Online]. Available: https://www.geeksforgeeks.org/logistic-regression-using-statsmodels/
[18] 'CMUdiaml8.pdf: Data, Inference, and Applied Machine Learning - DIAML'. Accessed: Nov. 05, 2023. [Online]. Available: https://canvas.cmu.edu/courses/36029/files/10254849?module_item_id=5631119