REPORT OF ASSIGNMENT 5

Carnegie Mellon University Africa

Submitted By: Peace Ekundayo Bakare

Course: Data, Inference, and Applied Machine Learning

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

Libraries Used:

1. pandas as peacepd
2. seaborn as peacesns
3. numpy as peacenp
4. matplotlib.pyplot as peaceplt
5. statsmodel.api as peacestats
6. mean_squared_error, confusion_matrix, accuracy_score from sklearn.metrics
7. SequentialFeatureSelector from sklearn.feature_selection
8. LogisticRegression, LinearRegression from sklearn.linear_model
9. train_test_split from sklearn.model_selection
10. LabelEncoder from sklearn.preprocessing

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

QUESTION 1 – Statistical Learning

1.1 Rule-based approach to decision-making

Rule-based approach or systems are used to make real-time informed decisions. It is often used in industries such as manufacturing, healthcare systems, finance, etc. Rule-based approach is also regarded as expert systems or decision support systems as they simulate human expertise decision making ability. The system is built by harnessing domain experts knowledge through a series of rules encoding.[1]

In addition, Rule-based systems are not complex models. Problems are defined in simple if-then statements. They are a set of facts, rules used to decide how operations are performed. The system can be trained i.e. the models can learn from experience and perform certain operations.[2]

Forward Chaining (data-driven), Backward chaining (goal-driven), and hybrid systems are types of rule-based approaches, each tailored to different applications.[3]

Steps to Implementing a rule-based approach

(a) **Data Input**: The system takes input data from the user of the system or any other source of data. The type of data that serves as input can be numerical values or complex information like financial information, health records etc.

(b) **Rule Matching**: After that the data is inputted into the system, the inference engine evaluates the input data against the rules stored in the knowledge base. It basically looks for the rules whose conditions match the data inputted.

(c) **Rule Execution**: When a rule is successfully matched, the inference engine then executes the subsequent actions which might involve working memory update, derivation of new facts, or output generation.

(d) **Conflict Resolution**: There are cases where multiple rules are simultaneously triggered. This is a conflict. To resolve the conflict, the inference engine uses some strategies such as prioritization of rules based on specificity or order of entry. The strategies are used to determine which rule should apply.

(e) **Output Generation**: Based on the rules executed, the system generates an output. Output generated can be a decision, a recommendation, or any other form of response depending on the domain.

Example: Rule-based system in Mailbox folders

Rules

Rule 1: If the Email received starts with "AHC", insert the mail in the "Assigned Tickets" folder.

Rule 2: If the Email received contains "Stash" or "Pull Request", insert the mail in the "Stash Request" folder.

Rule 3: If the Email received contains "Engagedly", insert the mail in the "Employee Appraisal" folder.

⇨ As soon as an email is received, the inference engine automatically execute the rules (stored in the knowledge base) against the email received (data input) and moves the email into the appropriate folder depending on the rule that applies. For instances where more than one rule apply, the first rule set is used.

YES! Domain Knowledge is required to establish a rule. To prevent incorrect decisions, an outcome of too broad rule or too narrow rule, an understanding of domain specific knowledge is mandatory. This is why domain knowledge is required to establish a rule.

1.2 What is over-fitting and why it is a problem in statistical learning

Over-fitting can be described as a situation whereby the model built has learned the training data too well, including the noise, such that it performs maximally on any seen data but poorly on a new or previously unseen data. An over-fitted model cannot forecast future activity. This is one of the problems of over-fitting.
According to Investopedia, "Overfitting is a modeling error in statistics that occurs when a function is too closely aligned to a limited set of data points. As a result, the model is useful in reference only to its initial data set, and not to any other data sets."[4]

Why is over-fitting a problem in statistical learning?
An over-fitted model is a problem in statistical learning because it is not useful in predictive performance – it performs poorly by failing to generalize to new data as it did by having high accuracy on training data.

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

Preference between simple model with one parameter and complex model with ten parameters.

The most appropriate type of model for a small dataset with ten data points would be **a simple model with one parameter**. A complex model with ten parameters would most likely result to over-fitting, capturing noise rather than the underlying patterns required. Limited data with several parameters result to overfitting.[4]

1.3 Two commonly approaches to avoid Over-fitting

(a) **Cross-Validation**: This involves the splitting of data into multiple subsets – train dataset and test dataset. The model is trained on the train dataset while it is validated/tested on the test dataset. This approach helps to ensure that the model performs well on different datasets.

(b) **Regularization**: This approach employs the addition of a penalty for larger coefficients in the model in order to prevent it from becoming too complex. Common regularization techniques are **Lasso (L1)** and **Ridge (L2)**.

1.4 Metrics used to evaluate Model Performance

(a) **Accuracy**: This measures the proportion of correct predictions out of all the predictions made.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ Number\ of\ predictions}$$

(b) **Mean Squared Error (MSE):** This measures the average squared difference between the predicted value and the actual value.

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

Application and appropriate metrics

(a) **Accuracy** – For Classification problems
   (i) Spam Detection – emails can be easily classified into different folders e.g. spam, inbox, junk etc.
   (ii) Credit Scoring – in financial systems, customers or potential borrowers can be classified as either "high risk" or "low risk" based on their credit history

(b) **Mean Squared Error (MSE)** – For Regression problems
   (i) Weather Forecasting – prediction of rainfall, temperature, wind direction etc.
   (ii) Energy consumption prediction – based on past usage, a house, community or city can predict their consumption.

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

1.5 Why Benchmarks are useful in Machine learning

Benchmarks provide a standard for comparing different machine learning models and approaches, to guarantee consistency and reliability in performance evaluation. Benchmarks are used in Artificial intelligence, machine learning to be precise, to assess the **performance** (including accuracy, speed/latency for tasks like inference, and other related metrics), **resource evaluation** (CPU and memory utilization, power consumption, etc.) and **acceptance** (adherence to requirements and specifications).[5]

1. **Comparison of Model Performance:** With benchmarks, researchers and engineers can test the performance of different models on the same dataset. It provides a way to identify the model that performs best in terms of accuracy, speed, and efficiency.
2. **Progress Tracking:** Progress can be measured over time with benchmarks. Benchmarks help to identify where progress has been made in the field and where there is still a possibility for growth.
3. **Model Selection:** Benchmarks help to choose the right models for specific tasks by providing clear performance metrics.

Benchmark Examples:

(a) **ImageNet (2009)**: ImageNet is an image dataset consisting of millions of images with associated labels from the WordNet taxonomy.[5] This is a benchmark used for image classification tasks. It helps to compare the performance of different image recognition models. It has transformed the deep learning field.

(b) **GLUE (General Language Understanding Evaluation):** This is a benchmark created to test a model's broad language understanding across tasks like sentiment analysis and textual entailment, measuring how well models generalize language knowledge.[5]

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

QUESTION 2 – Machine Learning

2.1 Machine Learning, its evolution and why it is popular

Machine Learning (ML) is a branch of artificial intelligence that is focused on enabling systems to learn and make critical decisions from data without being explicitly programmed for each task rather it involves building models to handle the tasks. Using algorithms and neural network models, machine learning systems have improved their accuracy and performance over time through the training of models and using the models to identify patterns and predict outcomes.[6]

Evolution of Machine Learning

Machine learning has advanced significantly since 1950, when Alan Turing developed the Turing Test, which proposed that a machine may be considered intelligent if it could interact convincingly with humans and pass as one. This generated interest in what it meant for robots to "learn." In 1952, Arthur Samuel at IBM took a practical step with ELIZA, a program that could play checkers and improve with experience, demonstrating the ability of a machine to learn. A few years later, in 1957, Frank Rosenblatt of Cornell University expanded on this with the perceptron, the first neural network model, laying the framework for the complex neural networks we use today.

Machine learning had advanced by 1990, combining computer science and statistics to construct models capable of analyzing and learning from data, progressing from basic rule-based systems to deeper pattern recognition. Then came the Big Data explosion in 2010, when the sheer volume of data accessible enabled massively parallel training of algorithms, making them more powerful and accurate. Finally, in 2014, the development of Open Data and the Internet of Things (IoT) enabled machines to access and analyze real-time data from a number of sources. Each of these developments was critical in changing machine learning from a theoretical concept to the dynamic, far-reaching technology that we see today.

Why is Machine Learning popular today?

Machine Learning is popular today because:

1. It powers essential technologies across industries, from real-time personalization and fraud detection to self-driving cars and diseases prediction in healthcare systems.
2. Its adaptability, scalability, and ability to continuously learn make ML an attractive tool for businesses seeking efficiency and new insights, further driven by applications in other technologies such as Internet of Things (IoT), Natural Language Processing (NLP), dynamic pricing, and more.
3. Machine Learning is popular today because of its ability to handle big data and automate complex tasks. Traditional methods cannot keep up with the large

volume and complexity of data in our world today, but machine learning models are better built with large datasets.

2.2 Examples of Machine Learning techniques – Supervised or Unsupervised

Supervised learning is a machine learning algorithm that learns from labelled data i.e. data that has been paired or tagged with the correct answer or classification. Supervised learning is used for classification, regression and object detection problems.

Unsupervised learning is a machine learning algorithm that learns from data that is not labelled or categorized. The objective of unsupervised learning is to detect patterns and relationships in the data without any explicit supervision. Unsupervised learning is used for clustering and association problems.

(a) **Clustering/Segmentation**: This is an unsupervised machine learning technique that groups data into similar segments and exposes patterns in the data set.

(b) **Dimensionality Reduction**: This is also an unsupervised machine learning technique that identifies features of the data values that provide the most differentiation among data points.

(c) **Classification**: This is a supervised machine learning technique that learns a particular function that maps from the input features to a probability distribution over the output classes.[7]

(d) **Regression:** This is a supervised machine learning technique that learns a function that maps from the input features to the output value.

2.3 Difference between Classification and Regression

The most important difference between classification and regression is the type of data that they are used to predict. Below is a table showing a few difference between classification and regression.[7]

S/No.	Regression	Classification
1.	It is used to predict continuous data values such as house prices, stock prices, etc.	It is used to predict categorical data values, such as whether an email is spam or not, whether a medical imaging has a disease or not, etc.
2.	Regression algorithms learn a function that maps from the input features to the output value.	Classification algorithms learn a function that maps from the input features to a probability distribution over the output classes.
3.	Examples of Regression algorithm are Linear Regression, Polynomial Regression, etc.	Examples of Classification algorithm are Logistic Regression, Naïve Baye

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

2.4 Difference between Supervised and Unsupervised Machine Learning

Supervised and Unsupervised learning are both machine learning techniques that have quite different concepts and approach to learning. The main difference between supervised and unsupervised learning is labelled data.[8]

Below is a summary of their differences.

S/No.		Supervised Learning	Unsupervised Learning
1.	Goals	The goal is to predict outcomes for new data	The goal is to get insights from large volumes of new data
2.	Applications	It is ideal for spam detection, sentiment analysis, weather forecasting, pricing predictions, etc.	It is ideal for anomaly detection, recommendation systems, customer personas and medical imaging.
3.	Complexity	It requires a simple method, typically calculated by using python programs or R.	Powerful tools for working with large amounts of unclassified data is needed. They are computationally complex.
4.	Drawbacks	Training models consumes time. Also, the labels for input and output variables require expertise.	It can have wildly inaccurate results unless there is human intervention to validate the output variables.
5.	Number of Classes	Known number of classes.	Number of classes is unknown.
6.	Examples	Classification, Regression.	Clustering/Segmentation, Association, Dimensionality Reduction.

2.5 Successful Applications of Machine Learning

Below is a tabular representation of the successful applications of Machine Learning, explanation of the appropriate technique and the type of learning involved.

S/No.	Application	Technique	Type of Learning
1.	**Image & Facial recognition** used in security systems to grant access. It is also used in image tagging in social media platforms.	Convolutional Neural Networks (CNNs)	**Supervised** (labelling of images) and **Unsupervised** Learning (used in grouping similar faces)
2.	**Self-Driving Cars** – autonomous vehicles use machine learning to make decisions	Deep reinforcement learning, computer vision	**Supervised** learning for training on labelled roads and **Reinforcement**

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

		(CNN), sensor fusion	learning for continuous learning
3.	**Healthcare Diagnostics** – machine learning is used in medical imaging and predictive analytics in healthcare for diagnosis and prediction of disease outcomes	Convolutional Neural Networks, Decision Trees, Random Forest	**Supervised** Learning (training on labelled medical data), **Semi-supervised** learning (on some unlabeled data)
4.	**Speech Recognition** – Machine learning is used in virtual assistants such as Siri, Alexa, Cortana, etc.	Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks	**Supervised** Learning (on labelled audio) and **Unsupervised** learning (on new phrases or accents)
5.	**Loan Approval Prediction** – banks and financial institutions use machine learning techniques to decide whether to approve or deny loan requests based on historical data.	Decision Trees	**Supervised** Learning

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

QUESTION 3 – DIABETES DATA

3.1 Produce a correlation matrix of the explanatory variables, make an heatmap of the matrix and describe the relationships.

After loading the dataset *Diabetes_Data.xlsx* file as *df_diabetes* dataframe, I checked the first 5 rows of the dataframe using *head()* to confirm that I have the exact columns as mentioned in the question. Also, I confirmed the shape of the dataframe which is (442, 11) using the *shape* function of the pandas library.

To produce the correlation matrix of the explanatory variables, it's important to understand what it is and why do we need to produce the correlation matrix. The correlation matrix calculates the linear relationship between two variables. The matrix shows the relationship between all possible pairs in the table. It is constructed by computing the correlation coefficient for each pair of variables and inserting it into the relevant cell of the matrix. Correlation coefficients range from -1 to +1, where -1 means a perfect negative correlation, +1 means a perfect positive correlation, and 0 means there is no correlation between the pair.[9]

The formula to calculate the Correlation Coefficient between two pair is:

$$r = \left(\frac{n\sum xy - \sum x \sum y}{\sqrt{((n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2))}}\right)$$

The explanatory variables are AGE, SEX, BMI, BP, S1, S2, S3, S4, S5, S6

Correlation Matrix

	AGE	SEX	BMI	BP	S1	S2	S3
AGE	1.000000	0.173737	0.185085	0.335428	0.260061	0.219243	-0.075181
SEX	0.173737	1.000000	0.088161	0.241010	0.035277	0.142637	-0.379090
BMI	0.185085	0.088161	1.000000	0.395411	0.249777	0.261170	-0.366811
BP	0.335428	0.241010	0.395411	1.000000	0.242464	0.185548	-0.178762
S1	0.260061	0.035277	0.249777	0.242464	1.000000	0.896663	0.051519
S2	0.219243	0.142637	0.261170	0.185548	0.896663	1.000000	-0.196455
S3	-0.075181	-0.379090	-0.366811	-0.178762	0.051519	-0.196455	1.000000
S4	0.203841	0.332115	0.413807	0.257650	0.542207	0.659817	-0.738493
S5	0.270774	0.149916	0.446157	0.393480	0.515503	0.318357	-0.398577
S6	0.301731	0.208133	0.388680	0.390430	0.325717	0.290600	-0.273697

	S4	S5	S6
AGE	0.203841	0.270774	0.301731
SEX	0.332115	0.149916	0.208133
BMI	0.413807	0.446157	0.388680
BP	0.257650	0.393480	0.390430
S1	0.542207	0.515503	0.325717
S2	0.659817	0.318357	0.290600
S3	-0.738493	-0.398577	-0.273697
S4	1.000000	0.617859	0.417212
S5	0.617859	1.000000	0.464669
S6	0.417212	0.464669	1.000000

It was important to drop column Y for the dependent variable, not to include it in the correlation matrix. I used the new dataframe created, *df_diabetes_explanatory,* it excludes the dependent variable column, to create the correlation matrix.

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

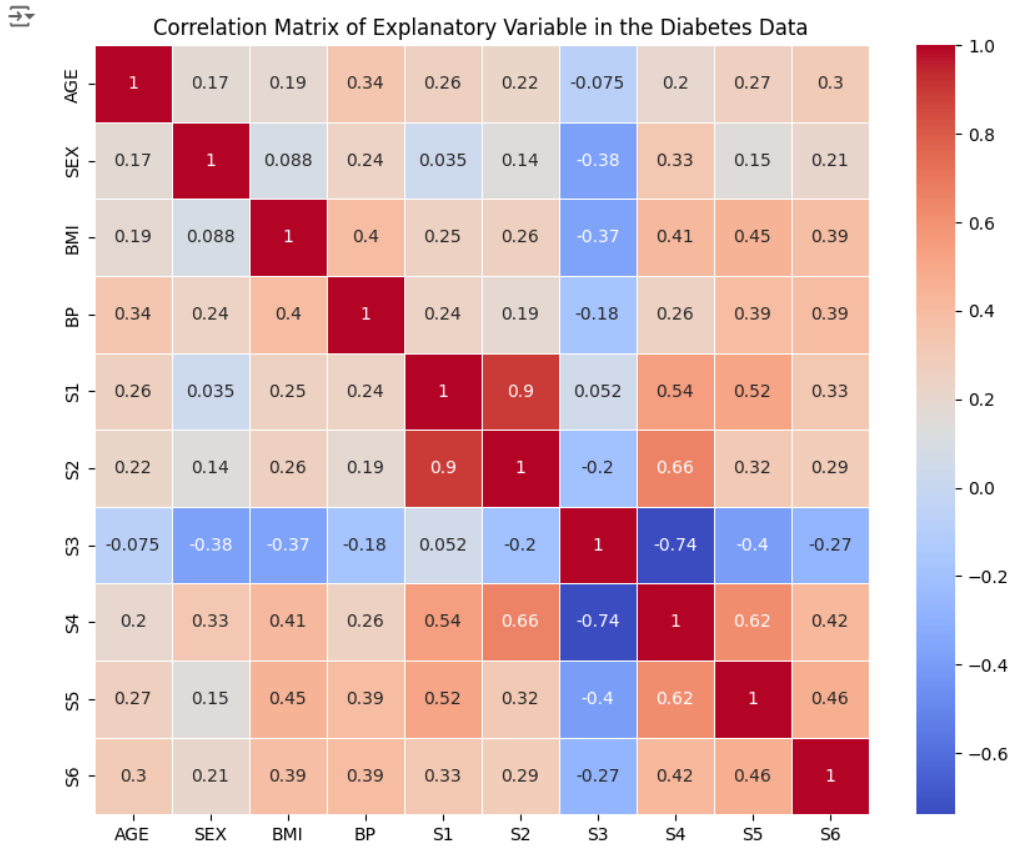


Fig: Heatmap showing the correlation matrix of the explanatory variables

Interpretation of the Heatmap, documenting insights from the observed relationships

S1 and S2 have a **very strong positive correlation** of **0.9.** This means that they likely measure similar aspects of the blood serum. This overlap could be a sign of redundancy in the model.

S2 and S4 are **moderately correlated**, with a correlation coefficient of **0.66**. So they might capture related factors in the blood, which could also lead to some overlap.

S3 and S4 show a **strong negative correlation** of **-0.74**. This indicates that an increased measurement of one, reflects the decrease in the measurement of the other. This might suggest that they are recording opposite trends in blood serum.

BMI and BP have a **moderate positive correlation** of **0.4**. This suggests that body mass may be somewhat related to blood pressure in this group.

BMI and S4 also show a **moderate positive** correlation of **0.41**. This also suggests that body mass may influence what S4 measures.

S5 and BMI have a **moderate correlation** of **0.45**. This suggests that S5 may also be tied to body mass.

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

These patterns suggest that some of the variables, especially S1, S2, and S4, might provide similar information, which could make it harder to get clear, independent effects for each one in a model. This issue, called **collinearity**, can affect the accuracy of the model's predictions.

3.2 Collinearity

Collinearity can refer either to the general situation of a linear dependence among the predictors, or, by contrast to multicollinearity, a linear relationship among just two of the predictors. In other words, collinearity occurs when two or more predictors are highly correlated.

Effect of collinearity:

(i) Simply, it can lead to unreliable coefficient estimates, making it challenging to interpret the model's predictors' contributions accurately.
(ii) It will inflate the variance and standard deviation of coefficient estimates which in turn reduce the reliability of our model.
(iii) It makes the p-value that shows that the independent variable is statistically significant unreliable.[10]

3.3 Build a Multivariate Linear Model, Calculate the Mean Squared Error and adjusted R^2

The objective of this step is to build a comprehensive model, **Model1,** that captures the combined influence of all available predictors on diabetes progression.

These are the steps I followed, and the results obtained:

Step 1: I set up a multivariate linear regression model using all the ten predictor variables. I achieved this by assigning *df_diabetes_explanatory* to a new variable X.

Step 2: I trained the model using the diabetes dataset. I used an algorithm, least-squares estimation, that iteratively adjusts the coefficients until the best fit for the data is achieved.

Step 3: I evaluated the performance of the model after training using the two metrics, MSE (Mean Squared Error) and adjusted R-squared.

The mean squared error measures the average squared difference between the predicted values and the actual values.[11] While the adjusted r-squared value informs us of the proportion of variability in the disease progression that can be explained by Model1.

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

Mean square error value is 2859.6963475867506

Adjusted R-squared value is 0.5065592904853231

Here is the summary of the model, **Model1**

```
                            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:                Sat, 02 Nov 2024   Prob (F-statistic):           3.83e-62
Time:                        18:31:29   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
==============================================================================
Omnibus:                        1.506   Durbin-Watson:                   2.029
Prob(Omnibus):                  0.471   Jarque-Bera (JB):                1.404
Skew:                           0.017   Prob(JB):                        0.496
Kurtosis:                       2.726   Cond. No.                     7.24e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.24e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Variable Significance

What helps to determine if a variable has a statistically significant relationship with an outcome is the **p-value.** Looking at the summary table, all the variables have p-values but not all the values are significant. The significant ones are SEX, BMI, BP and S5, having p-values P > |t| that is less than 0.05. Other variables such as AGE, S1, S2, S3, S4, and S6 are considered not statistically significant because their p-values are greater than 0.05.

In addition, the variables with p-values < 0.05 are considered to have a strong likelihood that they contribute meaningfully to the prediction of the dependent variable while those with p-values > 0.05 are considered not to have meaningful contribution to the prediction of the dependent variable.

Summarily, not all variables are significant! This suggests that the model could be improved by removing non-significant predictors.

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

Could this be a problem of collinearity?

From the summary statistics, the condition number of 7.24×10^3 is large which could suggest a condition of strong multicollinearity. Also, high standard errors of the predictor variables can result from strong collinearity or multicollinearity. When predictor variables are collinear, the model finds it difficult to distinguish between the effect of each of the predictors from another, leading to larger standard errors. Conclusively, the statistical insignificance of predictor variables like AGE, S2, S3, S4, and S6, suggests multicollinearity.

3.4 Difference between Forward Selection and Backward Selection

The major difference between forward selection and backward selection is the pattern of selecting their predictor variables.

Forward Selection begins with no predictors and then adds them one by one, selecting the one that improves the model the most at each step.

Backward Selection begins with all the predictors, removing the least significant one at each step, until only significant predictors remain.

3.5 Use function stepwise to interactively compose a model using forward selection

I used the *statsmodel* library to implement the stepwise selection. Here are the steps:

Step 1: I split the dataset into training and testing sets, *X_train, X_test, y_train*, and *y_test*.
Step 2: I initialized a linear regression model as *lin_reg*.

Step 3: I performed forward selection using *SequentialFeatureSelector*.

Step 4: I selected the features and stored them in *selected_features* so that I can use it to fit the model.

The Selected Features are: SEX, BMI, BP, S1, and S5.

```
Selected Features by Forward Selection are:  Index(['SEX', 'BMI', 'BP', 'S1', 'S5'], dtype='object')
```

Step 5: I fitted the model and predicted the dependent variable. Also, I calculated the mean square error and the R squared.

Mean Square Error value is 3527.7641198416377

R-Squared value is 0.31204868953765896

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

Explanation/Insight: Values of MSE and R-squared are better in the Multivariate Linear model than in the Stepwise forward selection approach because the linear model used all the predictor variables, irrespective of some of the predictors being redundant or weak. This affords it to make more accurate prediction and explaining the variability in the data (lower MSE and higher R-squared, when compared to the Forward selection). In the case of the forward selection, it builds the model by adding predictors that improve the model fit the most at each step, one at a time.

Here is a rule:

1. The lesser the MSE, the closer the model's predictions to the actual data points. MSE is the difference between the model's predicted value and the actual value.
2. The higher the R-squared or adjusted R-squared, the more the variability in the data. R-squared measures the amount of the variance in the dependent variable that is explained by the independent variables (predictor variables).

How Stepwise work in the sense of selecting variables?

The stepwise approach is an iterative method used in statistical modeling to select variables for a model. It combines **forward selection** and **backward elimination** to build a model that balances simplicity and explanatory power. The model starts with an initial set of variables, and at each step, a predictor variable can either be added or dropped, depending on which of the actions would improve the model fit the most.

The Forward selection method begins with an empty model, then it adds variables one at a time, the variables that best improves the model. It uses criterion such as p-value or BIC. This is done until the addition of more variables no longer improves the model significantly.

As opposed to forward selection process, backward selection or elimination begins with all predictor variables and then removes the least significant variable at each step until all the remaining predictor variables are statistically significant.

How Selected variable and Function work?

The stepwise function (here implemented as *SequentialFeatureSelector* with forward selection) selects variables by assessing each one's contribution to reducing Mean Square Error. Only the variables that improve the model the most are kept. This results in an array of variables that are most relevant to predicting the target variable.

SequentialFeatureSelector tests each feature iteratively, adding one feature at a time. It calculates a performance metric, using cross-validation to find the best subset of features.

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

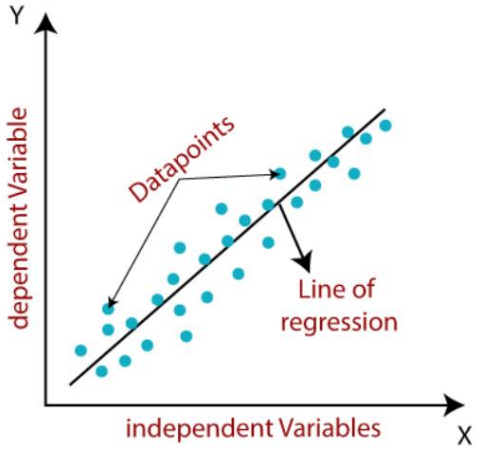
QUESTION 4 – ANALYZING TITANIC DATASET

4.1 Difference between Logistic Regression and Linear Regression

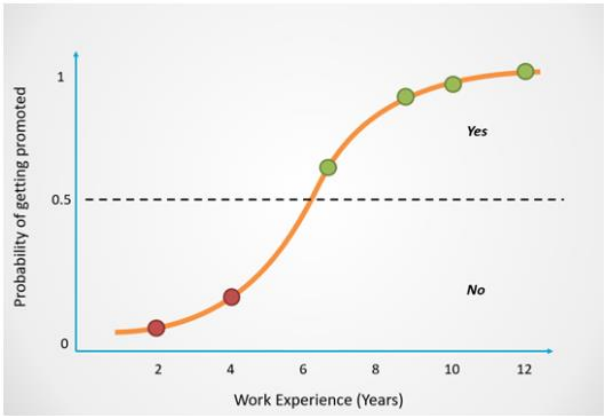
Logistic Regression and Linear Regression are both Supervised machine learning algorithms, but they have a few differences. I will classify their differences in the following categories: Application, Loss Function, Output type, Assumptions, and Equation.

S/No.	Category	Logistic Regression	Linear Regression
1.	**Application**	It is used for classification tasks, i.e. prediction of **categorical** values.	It is used for regression tasks, i.e. prediction of **continuous** values.
2.	**Loss Function**	It minimizes the logistic loss (log loss)	It minimizes the sum of squared differences.
3.	**Output Type**	It predicts probabilities for binary classification.	It predicts continuous values.
4.	**Assumptions**	It assumes a binomial distribution of the dependent variable.	It assumes a linear relationship between variables and normal distribution of the dependent variable.
5.	**Equation**	It uses a logistic function to transform the output into probabilities.	It uses a simple linear equation $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$
6.	**Example**	Outcome of the roll of a six-sided die.	Prediction of house price based on the number of rooms, neighborhood, and age.

Linear Regression representation [Source: Link]



Logistic Regression Representation [Source: Link]



There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

4.2 Calculate the Probability of Survival for a passenger on the Titanic

After loading the dataset *titanic3.csv,* I used the pandas *head()* function to have a clue of what the dataset looks like. I then calculated the probability for survival for a passenger on the titanic, using the 'survived' column in the dataframe, *df_titanic,* by calculating the mean of the data.

The Probability of survival of a passenger is **0.3819709702062643**, approximately 0.38.

4.3 Table of survival probabilities by passenger class, gender and age

Step 1: I defined the age groups using the *cut()* function of the pandas library to categorize the 'age' column into specified bins and assigned labels to each group namely, Child, Teenager, Young Adult, Adult, Senior.

Step 2: I grouped the data by passenger class (pclass), gender (sex), and the newly created group (AgeGroup), and then calculated the mean survival rate.

Step 3: I displayed the survival probabilities by passenger class, gender and age group as seen below.

AgeGroups		Child	Teenager	Young Adult	Adult	Senior
pclass	sex					
1	female	0.000000	1.000000	0.971429	0.977273	0.833333
	male	1.000000	0.500000	0.406250	0.312500	0.066667
2	female	1.000000	0.875000	0.893939	0.812500	NaN
	male	1.000000	0.000000	0.097087	0.035714	0.166667
3	female	0.466667	0.607143	0.451220	0.272727	1.000000
	male	0.342857	0.081081	0.174274	0.064516	0.000000

4.4 Build a Logistic Regression Model for survival rates based on passenger class, sex and age.

I used the statsmodel for logistic regression so that I can have a summary of the model, determine the parameter estimates and their statistical significance. Here are the steps I used:

Step 1: I dropped rows with missing values in the age column to ensure that the data is complete for the selected predictors.

Step 2: I used *LabelEncoder* to convert the column 'sex' into a numeric format i.e. 1 for female and 0 for male, to make it compatible with the logistic regression model.

Step 3: I defined the predictor variables X and target variable y. I set the Predictor variable to include the variables pclass, sex, and age. Also, I set y as the target variable 'survived', which represents whether a passenger survived or not.

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

Step 4: I split the data into Training and Test sets. 80% training set and 20% testing set, to help test the model's accuracy on data yet unseen.

Step 5: I built and fitted the logistic regression model and displayed the summary of the model.

Statistic summary table:

```
Optimization terminated successfully.
         Current function value: 0.460459
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                survived   No. Observations:                  836
Model:                           Logit   Df Residuals:                      832
Method:                            MLE   Df Model:                            3
Date:                 Mon, 04 Nov 2024   Pseudo R-squ.:                  0.3131
Time:                         19:47:03   Log-Likelihood:                -384.94
converged:                       True   LL-Null:                       -560.38
Covariance Type:            nonrobust   LLR p-value:                 9.701e-76
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.7932      0.462     10.378      0.000       3.888       5.698
pclass        -1.2778      0.129     -9.927      0.000      -1.530      -1.025
sex           -2.4804      0.190    -13.076      0.000      -2.852      -2.109
age           -0.0341      0.007     -4.817      0.000      -0.048      -0.020
==============================================================================
```

From the statistics summary, the coefficient estimated values of the predictors are displayed as well as the standard errors, z-values and p-values P > |z| which determines the statistical significance of the model.

Here are the parameter estimates (coefficients of pclass, age and sex)

```
Parameter Estimates [[-1.24275382 -2.38830536 -0.03334353]]
```

Image showing Parameter Estimates

Passenger Class (pclass) – (-1.24275382)

Age (age) – (-0.03334353)

Sex (sex) – (-2.38830536)

Based on the parameter estimates, these parameters are statistically significant as their p-values are all less than (< 0.05). The implication is that:

1. The passenger class has a statistically significant effect on survival probability.
2. Gender has a statistically significant effect on survival probability, where 1 represents the female gender and 0 represents the male gender.
3. Age has a statistically significant effect on the survival probability.

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

4.5 Model Performance

To calculate the classification accuracy based on the confusion matrix, here are the steps:

Step 1: I used the logistic regression model to predict the probability of survival for each passenger and converted the probabilities to binary predictions of 0 or 1.

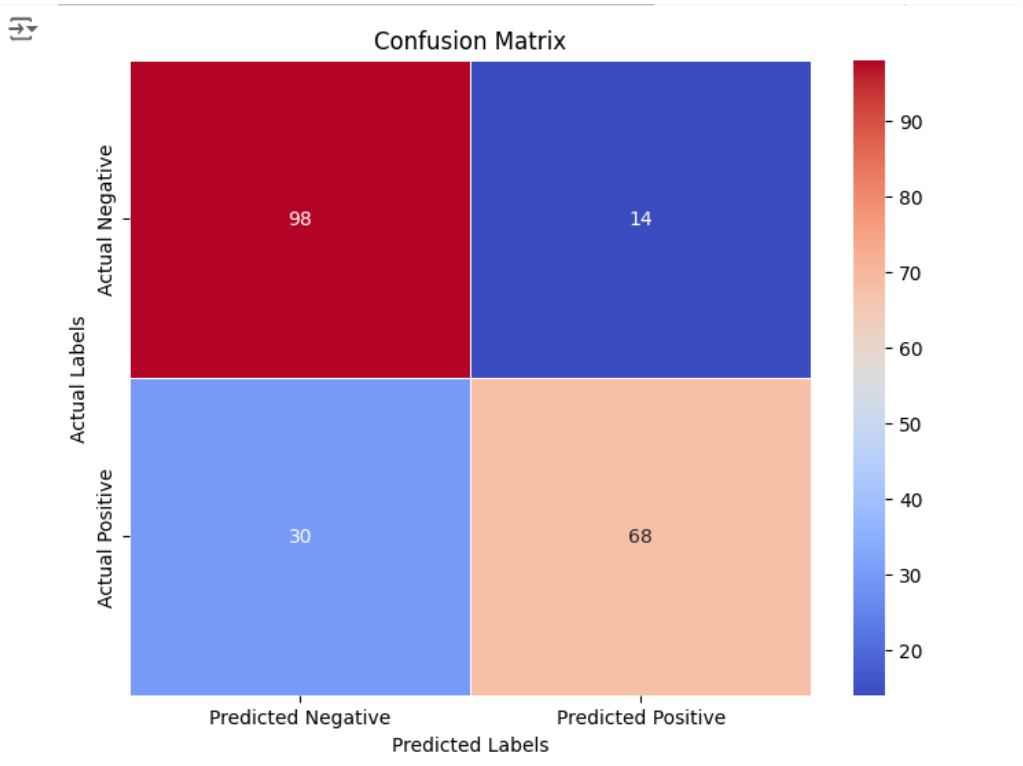
Step 2: I compared the binary predictions with the actual survival data to create a confusion matrix. The confusion matrix shows the counts of true positives, true negatives, false positives and false negatives.

Step 3: I calculated the classification accuracy of i.e. the number of correct predictions (number of correct classifications divided by the total number of classifications).

$$Accuracy = \frac{Number\ of\ correct\ classifications\ (TP + TN)}{Total\ Number\ of\ Predictions}$$

The confusion matrix has four values, representing True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN).

```
Confusion Matrix:
[[98 14]
 [30 68]]
Classification Accuracy: 0.7904761904761904
```



Heatmap display of the Confusion Matrix

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

False Positives (FP) – these are passengers who were predicted to survive, but did not survive unfortunately. There are **14** False Positives in total.

False Negatives (FN) – these are the passengers that were predicted to not survive, but amazingly survived! There are **30** False Negatives in total.

True Positives (TP) – these are the passengers that were correctly predicted to survive, and they did survive! There are **68** True Positives in total.

True Negatives (TN) – these are the passengers that were correctly predicted not to survive, and they did not survive. There are **98** True Negatives in total.

The Classification accuracy score is **0.7904761904761904**

The accuracy score of **79.0%** shows that the logistic regression model has a reasonable predictive power for the survival rates on the titanic dataset.

The model is effective at detecting both survivors (68 true positives) and non-survivors (98 true negatives), indicating that it has some predictive power for both groups. In 14 cases, the model inaccurately predicted survival for passengers who did not survive. While this figure is low, it suggests that the model is excessively optimistic about survival.

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.

References

[1] "Mastering Rule-Based Systems: Implementation, Benefits, and Best Practices," Mastering Rule-Based Systems: Implementation, Benefits, and Best Practices. Accessed: Nov. 02, 2024. [Online]. Available: https://www.xaqt.com/blog/mastering-rule-based-systems/

[2] web developer, "Rule-Based System Approach for Process Automation," Cflow. Accessed: Nov. 02, 2024. [Online]. Available: https://www.cflowapps.com/rule-based-system-for-process-automation/

[3] "Forward vs Backward Chaining in Artificial Intelligence," Built In. Accessed: Nov. 02, 2024. [Online]. Available: https://builtin.com/artificial-intelligence/forward-chaining-vs-backward-chaining

[4] "Understanding Overfitting and How to Prevent It," Investopedia. Accessed: Nov. 02, 2024. [Online]. Available: https://www.investopedia.com/terms/o/overfitting.asp

[5] "coe379L-sp24/docs/unit04/ml_benchmarks.rst at master · joestubbs/coe379L-sp24," GitHub. Accessed: Nov. 02, 2024. [Online]. Available: https://github.com/joestubbs/coe379L-sp24/blob/master/docs/unit04/ml_benchmarks.rst

[6] K. D. Foote, "A Brief History of Machine Learning," DATAVERSITY. Accessed: Nov. 02, 2024. [Online]. Available: https://www.dataversity.net/a-brief-history-of-machine-learning/

[7] "Supervised and Unsupervised learning," GeeksforGeeks. Accessed: Nov. 02, 2024. [Online]. Available: https://www.geeksforgeeks.org/supervised-unsupervised-learning/

[8] "Supervised vs. Unsupervised Learning: What's the Difference? | IBM." Accessed: Nov. 02, 2024. [Online]. Available: https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning

[9] A. A. Masud, "Correlation Matrix: What is it, How It Works with Examples," QuestionPro. Accessed: Nov. 02, 2024. [Online]. Available: https://www.questionpro.com/blog/correlation-matrix/

[10] "A Beginner's Guide to Collinearity: What it is and How it affects our regression model." Accessed: Nov. 02, 2024. [Online]. Available: https://www.stratascratch.com/blog/a-beginner-s-guide-to-collinearity-what-it-is-and-how-it-affects-our-regression-model/

[11] "Mean squared error," *Wikipedia*. Jun. 11, 2024. Accessed: Nov. 02, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Mean_squared_error&oldid=1228454019

There are 4 Questions that are documented in total. Each Question is started on a new page. Please, scroll down even when you see a blank page.