Report

Question 1.

1.1

Rule-based approaches are methods that are used to make decision where the decisions are made according to the set of rules that are set by the user. [1] Thus as it is can be understood from above definition, rule based approach involve setting the rules that guide decision making in the system. Here below, are some important decision-making steps:

1. **Set the rule:** what you do first is to set the criteria or conditions on which the decision is going to be made. These rules are typically expressed in the form of "if-then" statements. Each rule should map a specific combination of criteria to a decision or action. For instance, the rule might be if the returns on investment for a certain project after one year is 10000$, then keep with the project otherwise turn it down.

2. **Evaluate the rule and set the priority:** After creating all the rules, the next step is to evaluate the rules and rank them according to the precedence, which rules should be followed first. Prioritize some rules over other rules according to their importance, the likelihood, and the potential impact that could result from following the rule or not following rules.

3. **Apply the rule:** Apply the rule to the specific situation or data. The decision maker system then has to check whether the conditions that are met and then make the decision or action accordingly. Rules are applied sequentially, and the first rule that matches the conditions is usually the one that is executed.

4. **Decision validation and Adaptation:** After having a decision, it is of importance to validate the decision and checks it consequences. This consists of ensuring that the chosen course of action is valid and appropriate for a given context. If the decision made was unexpected or when there is a new kind of data, the rule-based system has to be updated.

   Example of the rule-based approach

   Rule-based systems are commonly used to filter spam emails. Rules can be defined to flag or move emails to the spam folder based on criteria like specific keywords, suspicious attachments, or senders with certain characteristics.
   With the above example, the developer needs to have knowledge about classification that is knowledge about which characters and features that characterize the spam emails.

   1.2

overfitting is the problems that commonly in statistical and machine learning where by the model learns to fit the training data too closely, capturing noise and random fluctuations in the data rather than the underlying patterns or relationships that is, model learns the expected output for every data point. Thus in simple words, it can be defined as undesired conditions which that happens in machine learning where a model output the accurate prediction for the training data but not for a new data

Reasons why overfitting is a problem

Overfitting is associated with the following problems [2]

1. **No generalization:** no overfitted models excels at the training data but often fails to generalize to new, unseen data. This is due to the fact that the overfit model memorize the entire training dataset, even its noise instead of learning the fundamental patterns of the data. Consequently those overfit models are unreliable and are unable to offer valuable predictions or insights beyond the data that they were trained on.

2. **Less efficient in prediction**: In addition, overfits models are highly specialized to the specific data. They may not capture the broader relationship in data, which is the main goal of statistical learning. When given a new data that were not the model were no trained on, the overfit models are likely to produce inaccurate or inconsistent predictions.

Given a small data set containing ten data points, I would prefer the simple model with one variable over the complex model with 10 variables due to the following reasons:

- With a small dataset, a complex model with more number of parameters is more prone to overfitting which would not work well with predicting new outcomes and which will lack generalization. In contrast, simpler model is less prone to overfitting because it has fewer degree of freedom to fit the noise.
- Computational efficiency: the computation in simple model is less demanding than that in complex models so this means there higher computational efficiency with simple model than that with complex models.
- Occam's Razor: Occam's Razor says, simple models are preferred over complex models given that other features are equal. So this is also one of the reasons that made me to choose the simple model over the complex one. [3]

1.3 The two main methods that are used to avoid overfitting are cross-validation and regularization, let me try to explain them:

- Cross-validation: this is a technique that is used to evaluate a model's generalization ability and performance while avoiding overfitting. It involves grouping the available date in various sections, typically a training set and a testing set (for validation). There exists several types of cross-validation techniques and K-fold cross-validation is the most common one. With K-fold, data is divided into k equally sized subsets called folds. The model is trained on k-1 folds and validated on one remaining fold.
- Regularization: This is a technique that adds a rule to the model training process to encourage simpler models. It helps prevent models from becoming overly complex and reduces the risks of overfitting. Some of techniques used in Regularization include L1 Regularization (Lasso), L2 regularization (Ridge), and Elastic net. For instance L2 regularization adds a penalty term proportional to the square of the model parameters. While it does not force parameters to become zero, it helps control the magnitude of the parameters, preventing them to become very large.

1.4

Two examples of metrics that are used to evaluate the performance of a model alongside with their formulas are discussed below:

1. Mean Absolute Error(MAE):

FORMULA MAE= $\frac{\sum |actual\ value - predicted\ value|}{n}$ $where\ n\ is\ the\ dataset\ size.$

Mean absolutes error produces the average absolute difference between the predicted values and actual values. It is mostly used in regression model to predict how well the model aligns with the actual values. The best model are those will the smallest possible mean absolute error.

<u>Example:</u> MAE can be used in house price prediction. In the real estate context, you can use Mae to assess how close the predicted house prices are to the actual selling prices

2. Classification Accuracy:

Formula: classification Accuracy= $\frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions}$

The classification accuracy measures the proportion of correctly predicted instances out of all the instances in a classification problem. It is used to evaluate classification models.

Examples: apart from email filtering that was said above, classification is also used in image recognition to classify images into different categories. For instance once given a image of animals this can be used to help you to classify animals into their types such as dog, cat, bird, cow . the classification accuracy can be used in this case to determine the efficiency/accuracy of the predicting model by taking number of true prediction over the number of all predictions

1.5

Benchmarks in machine learning are the standard or reference points that is used to measure the performance or effectiveness of the models or algorithms mostly by providing common dataset and evaluation criteria. Example of those benchmarks includes ImageNet Large Scale Visual Recognition Challenge which is focuses on image classification tasks in computer vision.

So, after understanding benchmarks in machine learning context let ‘s be back to our question.

Benchmarks in machine learning are essential for objective performance evaluation and progress tracking. They allow for fair comparisons of different models and techniques using standardized datasets and evaluation metrics. For instance, the ImageNet challenge aids in advancing computer vision by assessing image classification models. In natural language processing, benchmarks like Common Crawl and GLUE monitor the progress of language models. Benchmarks play a crucial role in identifying the most effective solutions for various tasks, shaping the field, and solving real-world problems. [4]

Question 2.

2.1

Machine learning is a subdivision of artificial intelligence which focus on designing algorithms and model that enable computers to learn and make prediction or decision on something based on fed up data without being explicitly programmed. Briefly, machine learning deals acquiring knowledge from the data and being able to make decisions and respect to future events/ unseen data according to the training data.

Evolution of machine learning [5]

Machine learning technology has been around since 1952. It has changed a lot in the past ten years and had some important changes in the mid-90s

1. 1950's -1960's **Early Foundations :** machine learning was found and started to be used in early 1950's. at this stage, machine learning was its infancy and the algorithms were relatively simple
2. 1990's **transition to data driven learning** : people started to use vast amount of data to train computers algorithms. The data was obtained with the help of internet.
3. 1995-2005 **Natural Language, search, and information Retrieval**: in this period there was a great emphasis on applying machine learning in natural language understanding, search engines, and information retrieval systems. Machine learning helped these systems understand and respond to human language, making search engines more effective and information retrieval faster and more accurate.

4. 2000s-2010s **Neural networks revival**: the subset of machine learning called neural networks which is inspired by human brain, made a significant comeback after in 2005, thanks to advances in deep learning. It had once existed and later faded out but at that time it made it again to regain its popularity. Since then, deep learning revolutionized fields including natural language processing, computer vision, and speech recognition which in turn allowed computers to see, hear and speak.

5. 2016 onwards **Mainstrean Adoption**: in 2016, experts believed that machine learning was getting into the phase of "according to Gartner's Hype Cycle" this means there was a lot of excitement about its potential. It was predicted to become the most widely used within the following 2-5 years and this was recognized later to be accurate.

6. On-going progress: Machine learning is day to day evolving rapidly. Researching and engineers are developing new algorithms, models, techniques that are more powerful.

2.2

some machine learning algorithms can be used in either supervised or unsupervised contexts depending on how they are applied. Here are three examples of those algorithms:

- Decision Trees
- K-means clustering
- Principal component analysis(PCA)

2.3 difference between classification and regression [6]

the main difference between classification and regression is the type of the output they produce. Regression is used to foresee the continuous numerical values, whereas classification assign data to discrete categories or classes. In other words, the objective of regression is to predict continuous numerical value but the goal of the classification model is to categorize input data into predefined classes or categories.

2.4 the difference between supervised learning and unsupervised learning: [7]

Supervised learning is a type of machine learning which requires a labelled data (both input and output data) during the training phase. While unsupervised learning does not require labelled data, it is the training of a model by using raw, unlabelled data. Overall, the main distinction between supervised machine learning and unsupervised learning is that supervised learning requires labelled training data but unsupervised does not require labelled training data but it instead uses raw unlabelled data.

2.5 example of a successful applications of machine learning

here below there are three examples of successful applications of machine learning alongside with the their learning type, and technique involved.

1. Autonomous vehicles:
   Applications: self-driving cars and drones that navigate through environments an dmake a real-time decisions to drive themselves.
   Technique: Reinforcement learning, this is a type of machine learning that uses trial and error, where the agents learn to take actions to maximize a cumulative reward.
   Learning type: Reinforcement learning

2. Image recognition:
   Applications: detecting objects and patterns in images such as facial recognition and object detection in photos.
   Learning type: Supervised learning
   Technique: Convolutional Neural Networks (CNNs) , these are deep learning models that are trained on labelled image datasets, making it a supervised learning task.

3. Medical diagnosis:
   Application: here is where machine learning is applied to help medical professionals to diagnose diseases and conditions according to the patient's data and medical imaging.

   Learning type: supervised learning
   Technique: supervised learning such as deep neural networks is applied to learn from a labelled patient data and medical images.

Question 3

3.1 Here, the task is to load the diabetes data and produce a correlation matrix between explanatory variables.

Procedures:

1. Load the data in python by using appropriate read function in python according to the type of document that you want to load. In this case, it is excel file.

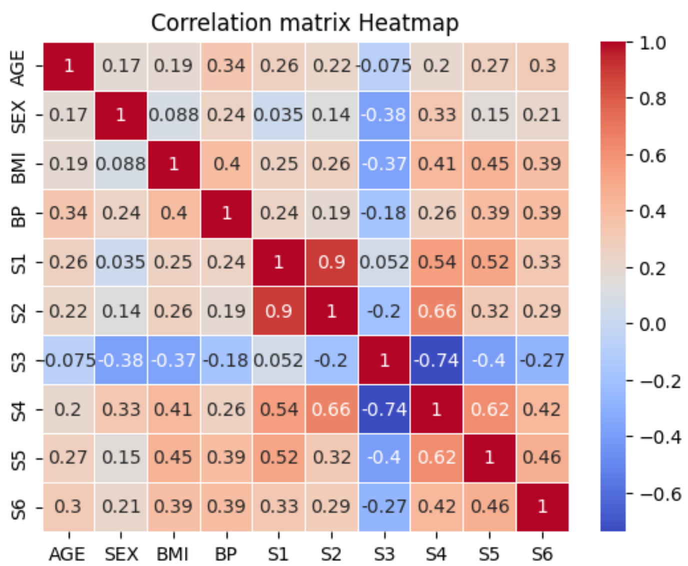
2. Select the explanatory variables. The explanatory variables also named independent or predictor variables. All the column titles except the last row of the dataset which is independent variables `'AGE','SEX','BMI','BP','S1','S2','S3', 'S4','S5'`,and `'S6'` and the dependent variable is ‘Y’

3. Find the correlation matrix between all the 10 predictor variables.

I used corr() method to obtained the correlation matrix and used heatmap to show clearly the range of correlation matrix between variables using range of colours.

The output

This is a resulting correlation matrix



interpretation of the correlation matrix

on the left, this table shows the correlation between different matrix shown by colours red colours shows possible highest positive correlation between variables and the more the correlation approaches to 1 the more the colour approaches to red colour and blue colour shows the highest negative correlation between variables and the more the correlation of variables approaches to -1 the more the colour approaches to blue colour . the correlation coefficient is neutral colour between red and blue colour.

From the above correlation matrix note that the correlation between any variable and itself is 1, it is a full correlation since all the data set is unchanged.

There is a great positive correlation between S1 and S2, a correlation of 0.9 This is to mean that S2 and S3 are highly related, this can be interpreted that 90% of the S2 and S3 data are the same.

Another thing which we can note is that, there is the highest negative correlation between S3 and S4, the correlation of -0.74, this is to mean that the 74% of data of S3 and S4 are opposite to one another.

The least related/ correlated variables S3 and AGE which is -0.075 , the least possible negative correlation between the variables. That is the relationship between AGE and S3 is least and very close to none.

3.2

Collinearity is a condition where two or more independent variables in a statistical model exhibit strong relationships or close associations with each other.

Collinearity occurs when predictor variables are strongly related in a regression model.

Effects of collinearity amongst predictor variables on estimated coefficient value

1. **Increased variability:** Collinearity makes it difficult to assess the individual effects of correlated variables on the dependent variable. When the predictor variables are highly correlated, it becomes difficult to distinguish their unique contribution to the model. Thus, the estimated coefficient values of the correlated variables may become unstable and highly variable across different samples.
2. Collinearity leads to increased standard errors for the coefficient estimates of the correlated variables. This increased standard errors in turn means that the model become less precise and the confidence intervals for the coefficient values become wider.
3. **Unreliable interpretations:** it becomes difficult to interpret the estimated coefficient due to collinearity between the variables. Strong correlations between predictor variables can result in coefficient values that seem surprising and not in line with expectations.
4. **Instability in coefficient signs:** collinearity can cause the instability of coefficient signs

3.3 the task here is to create a multivariate linear model using all ten variables and a constant.

Steps followed:

1. define explanatory variables and dependent variables
2. add constant to the predictor variables.
3. Define the model.
4. Use the model to predict the dependent variable values.
5. Find the coefficient of determination R2
6. Calculate mean squared error
7. Display coefficient of determination and mean squared error.

Results:

```
mean squared error:  2859.6963475867506
coefficient of determination: 0.5177484222203498
```

```
with this model there is a strong multicollinearity thus the condition number
of the model is very large. Some of the variables are not even significant
according to the p-values of the variables which can be obtained in the model
summary, the significant ones are SEX, BMI, BP, and S
```

The significant variables are: ['SEX', 'BMI', 'BP', 'S5'] these are the variables which have the significant contribution to the dependent variables

5

3.4 difference between forward selection and backward elimination:

These are feature selection techniques used in machine learning and statistical modelling. The difference lies mainly in how features are selected.

Forward elimination is a method used in machine learning and statistical modelling which assist building a predictive model by adding features one by one. The goal of this the technique is to select the most important features (dominant features) while avoiding overfitting.

On other hand, Backward elimination is a feature selection technique used in machine learning and statistical modelling by iteratively removing features from a model, starting with a model that includes all available features, the objective of this backward elimination is to simplify the model and find the most crucial features while makes sures that there is no overfitting in the model.

So, this table below shows clearly the distinction between the two features selection techniques.

criteria	Backward elimination	Forward elimination
Direction of selection	Starts with model with all features and remove a feature at time from the model of all variables.	Begins with an empty list features and adds one feature at time to the model
complexity	Starts with the most complex model and reduces the complexity over the iterations	Seems to start with a simpler model and gradually increases complexity as the selection is done.
Risk to overfitting	Initially, it has high risk of overfitting but it diminishes the risk as some features are eliminated.	Initially, it is less prone to overfitting, but as the more features are added the risk increases.
Suitability	Suitable when you begin with a full set of features and want to simplify the model without sacrificing predictive accuracy	It is suitable when there is a large set of features and when one want to build a model progressively while monitoring the performance
Computational efficiency	It can be less efficient when dealing with large number of features initially.	Can be computationally more efficient as the there is initially empty features, the search space starts small.

Search space	Starts with all features and narrows down the search space	Scans through the available features to add them to the model.

Forward regression is a stepwise feature selection technique used in linear regression. It begins with an empty model and iteratively adds the most significant predictor variable that improves the model's performance, as measured by a chosen criterion (e.g., R-squared or AIC). This process continues until a stopping criterion is met, such as a predefined number of features or a drop in performance. Forward regression simplifies the model and selects the most relevant features for predicting the target variable, helping to mitigate issues like multicollinearity and overfitting. It is a systematic approach to build a parsimonious linear regression model from a larger set of potential predictors.

3.5

stepwise regression is a special case of hierarchical regression which determines significant variables to end up in the simplified model by removing or adding variables in the model according to statistical algorithms. In forward selection method the model begins with no predictor variable and add significant variable one at a time until the stopping conditions is satisfied. In backward elimination the model starts with the all variables and gradually eliminates the non-significant variables until the stopping conditions is satisfied.

For this question, the task was to compose a model using forward selection.

Steps followed to perform forward selection, with results obtained.

1. define the values for explanatory variables and the values for predictor variables
2. select the variables by forward selection: here we use forward_regression function that is found from step_reg package. While setting the selection conditions to be 0.05 which means that we are only selecting variables whose p_values are less than 0.05 that is we are only selecting the significant variables that have a significant contribution to dependent variable.
3. Forward selection selected only 6 variables among 10 variables which were present. **the selected variables are:** `'BMI', 'S5', 'BP', 'S1', 'SEX', 'S2'.` **Out of curiosity I tried to check which variables would be selected by using backward elimination and I found that the selected variables with backward elimination was same as the one selected with forward regression.**

4. After applying forward selection, we used the resulting variables to build a model by first adding a constant and then building and training the model.
5. The next step, was to predict the dependent variable Y using the model

6. For the sake of evaluating the performance of the model, we calculated the mean square error and the coefficient of determination.

The model's performance evaluation;

Mean squared error = 2876.68 which is actually higher than that obtained (`2859.69`**) in the model with all the variables but also not a big difference. The lower the mean squared error the better the model if other factors remains unchanged.**

```
mean squared error (forward selection):  2876.683251787016
coefficient of determination (forward selection): 0.5148837959256445
```

The coefficient of determination was found to be `0.514 which is a bit less than the one of the model with all` variables which was 05177. The higher the coefficient of determination, the better the model if other factor remains unchanged.

It requires to make a trade-off carefully because we don't have to forget that Occam's Raczor Principle says the simple the model the better the model since it will requires less computational power and it will use less computational time and less resources this could be a fact that can be used to say that the the model after forward elimination is actually the best model but on an other side remember that both coefficients of determination and mean squared error was showing that the full model using all the variables was better. According to all of the above factors, since the difference is mean squared error of the models is not significant and the difference in coefficients of determination is also not significant but the model is simplified , I conclude that the simplified model with only 6 variables which was achieved after performing regression is a better model, because it makes sense to sacrifice that non- significant performance benefits for the sake of having a simplified model which will requires less computational resources.

4.1 the difference between logistic and linear regression

Linear regression is used to predict variables which have real number values while logistic regression is used to predict binary-valued variables.

This table can clarify the distinction between logistic regression and logistic regression

	Linear regression	Logistic regression
purpose	Is used for estimating the continuous real valued variable.	It is used for predicting discrete binary valued outcome.
Output type	The output is continuous And the model fit a straight line to the dataset	The probabilistic value between 0 and 1 - Probabilistic score is used to categorize the data in 2 categories
Algorithm	Use ordinary least squares (OLS) to find the best-fitting linear equation that minimizes the sum of errors.	Uses logistic function (sigmoid function)to model the probability of binary outcome best fitting parameters
Model assumption	Assume linear relationship between independent and dependent variable, it assumes normal distribution of errors.	Assume that there is linear relationship between independent variable and log-odds of the dependent variable

Performance metric (error metrics)	It uses mean squared error (MSE) , RMSE, MAPE (mean absolute percentage error)	It uses classification error, cross entropy, confusion matrix, among others.

4.2 the task here was to load the titanic data and find the probability of surviving from the titanic ship crush.

1. The data was loaded from the provided link and it was read in python using appropriate read function found in pandas module.
2. Calculation of probability
   The survival probability the of the number of all people who survived to the number of all people who were in the ship. Thus we found and summed up the number of people who survived the accident and divide it with the total number of people who were in the ship.

   the survival probability was obtained as 0.382
   this means that about 38.2% of the people who were in the ship survived the accident.

4.3

Here we were asked to find the provide a table giving survival probabilities broken down by passenger class, gender and age. For this we were able to find the three tables, one for passenger classes and their survival probabilities , another for passenger's gender and their survival probability and another of age length and their survival probability.

1. **Survival probability by passenger class**
   There were three classes of passengers, class 1, class 2 and class 3. To find the probability of survived passenger in each class. We first filter the data of the certain class and we find the number of survived passenger in the class and then we divide it with the total number of people in the class.

   Results
   The results are shown below in the table:

	passenger_classes	survival probability
0	1	0.619195
1	2	0.429603
2	3	0.255289

<u>**Interpretation of result**</u>

As shown in the table the survival probability for people in class 1 is 0.619 and that of people in class 2 is 0.429 and that of people in class 3 is 0.255. this shows that people in class 1 had higher survival probability that the rest, this might be due to the facts such as region where first class people were staying did not sink deeply, class 1 people might have given the priority to rescued quickly more than others, or due to the fact that of sitting in the safe place. The class 3 has the

the lowest survival probability, this might be also have associated with the region where they were sitting and when and how the class 3 people were rescued.

2. **Survival probability by age:**

	age_classes	survival probability
0	adult	0.350126
1	child	0.492228
2	middle-aged	0.411765
3	old	0.242424

I grouped the passenger in age classes as the follows

category	**Age range**
Child	**0-18**
Adult	**18-35**
Middle aged	**35-60**
old	**60-90**

The table above shows the survival probability of the passenger in the titanic ship by their ages. The survival probability for children is 0.49 and that for adult is 0.35 , that for middle aged is 0.4117 and that or old people is 0.24

The probabilities above shows that the children had the highest probability of surviving. This is because children and female were priotized to join the life boats this in turn lead to a large number of surviving children and female as compared to others. As it can also be noted that older people had the lowest surviving probabilities this is because they were not priotized and they are most of the time very vulnerable with low energy and with other kind of disease this lead to low number of surviving old people.

3. Survival probability by gender:

	passenger_classes	survival probability
0	female	0.727468
1	male	0.190985

The above table shows that about 72.7% of females who were in the titanic ship has survived the crush but on another hand only 19% survived the accident. This is because ladies and

children were given a priority to enter the life boats which offered a great contribution in rescuing people's life.

4.4 The task here, is to build a logistic regression model based on class, sex and ages and then find the parameter estimates check whether the estimates are statistically significant.

As it was requested, these are the steps followed to perform logistic regression

1. Classify the gender, the female as 0 and male as 1.
2. Select the features to be used as explanatory variables and handle missing values
   The explanatory variables were passenger class, sex and age as it was instructed.
   Missing values were handled by replacing the every missing value with the average of the ages of the all people who were in the ship.
3. Add a constant to explanatory variables
4. Create the model and train it.
5. Predict the survived people by using the build model. Note that the model prediction is in 0 for not survived people and 1 for survived people.

6 display the model summary for more understanding of the model and for checking p-values of the explanatory variables to evaluate their contribution to the dependent variable.

<u>Results</u>

The following is the summary of the obtained model

```
Optimization terminated successfully.
         Current function value: 0.469029
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:               survived   No. Observations:                 1309
Model:                          Logit   Df Residuals:                     1305
Method:                           MLE   Df Model:                            3
Date:                Mon, 06 Nov 2023   Pseudo R-squ.:                  0.2947
Time:                        18:36:10   Log-Likelihood:                -613.96
converged:                       True   LL-Null:                       -870.51
Covariance Type:            nonrobust   LLR p-value:                 6.892e-111
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.3634      0.366     11.936      0.000       3.647       5.080
pclass        -1.0653      0.096    -11.122      0.000      -1.253      -0.878
sex           -2.4979      0.149    -16.793      0.000      -2.789      -2.206
age           -0.0320      0.006     -5.294      0.000      -0.044      -0.020
==============================================================================
```

The model summary shows the parameter estimates as:

	coef	std err	z	P>|z|	[0.025	0.975]
const	4.3634	0.366	11.936	0.000	3.647	5.080
pclass	-1.0653	0.096	-11.122	0.000	-1.253	-0.878
sex	-2.4979	0.149	-16.793	0.000	-2.789	-2.206
age	-0.0320	0.006	-5.294	0.000	-0.044	-0.020

Where coef stands for coefficient and std err is standard error, z for z -score P>|z| p-value and then [0.025 and 0.975 which shows the confidence interval.

Apart from the constant, the coefficient of all variables are negative value and this indicate that as the predictor variable increases, the response variable tends to decrease.

About the standard error the age has the lowest standard error which signifies precise and consistent measurements, suggesting a strong, reliable association with the dependent variable in statistical analysis.
The p_values of all values are far less than the significance level (0.05 which means) they are all significant.
And the confidence interval was provided by providing 2.5% percentile and 97.5% of the model prediction.

The P_values of the variables the following:

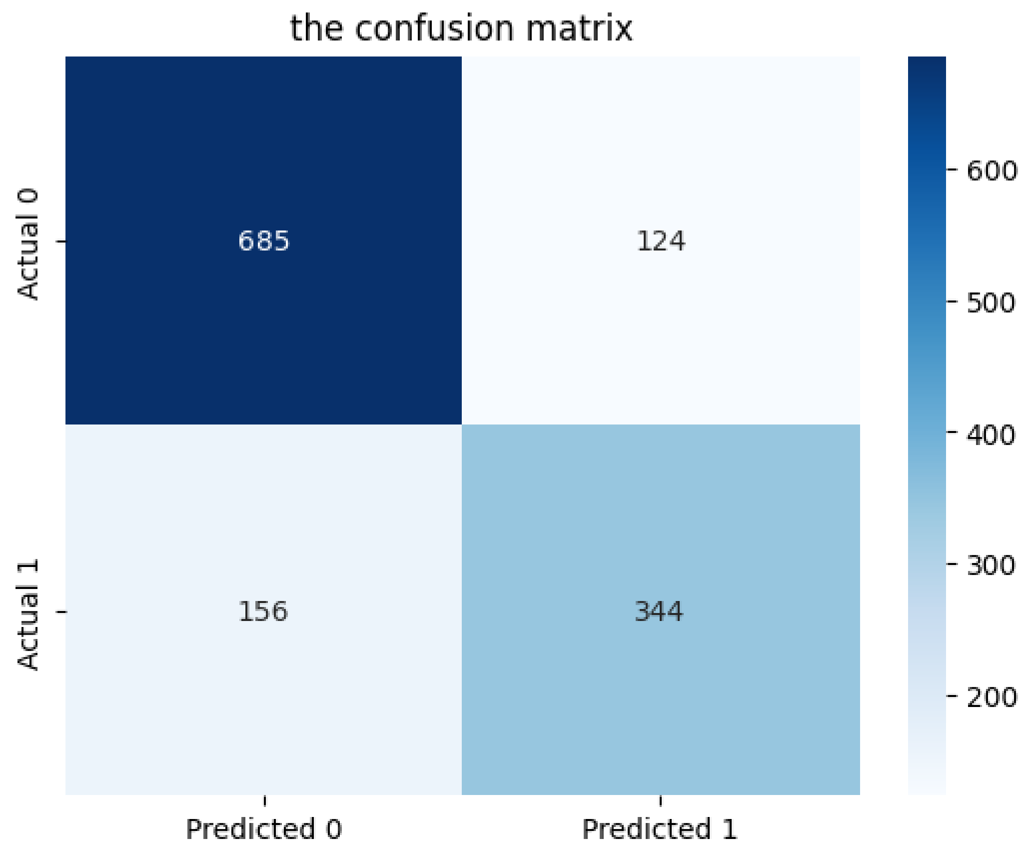
```
pclass    9.834551e-29
sex       2.747720e-63
age       1.193900e-07
```

As it can be noted the p-values for all the variables is far less than 0.05, the significance level, this is to mean all the variables **are statistically significant**

4.5 here we were asked to find the performance of the model by using classification accuracy based on confusion matrix
steps followed:
build on the read data of titanic, create a confusion matrix , by first importing the appropriate sklearn library. Find the classification accuracy of the model

the result

the following is the confusion matrix that was obtained:



The classification accuracy of the model was obtained as 0.786 or 78.6% this implies that the model correctly predicted the class label for 78.6 % of the instances in the dataset, indicating a moderate level of predictive performance.

References

[1] I. Timmer and R. Rietveld, 'Rule-Based Systems for Decision Support and Decision-Making in Dutch Legal Practice. A Brief Overview of Applications and Implications', *Droit Société*, vol. 103, no. 3, pp. 517–534, 2019, doi: 10.3917/drs1.103.0517.

[2] 'What is Overfitting? - Overfitting in Machine Learning Explained - AWS', Amazon Web Services, Inc. Accessed: Nov. 05, 2023. [Online]. Available: https://aws.amazon.com/what-is/overfitting/

[3] 'What is Occam's razor? | New Scientist'. Accessed: Nov. 06, 2023. [Online]. Available: https://www.newscientist.com/definition/occams-razor/

[4] J. Thiyagalingam, M. Shankar, G. Fox, and T. Hey, 'Scientific machine learning benchmarks', *Nat. Rev. Phys.*, vol. 4, no. 6, pp. 413–420, Apr. 2022, doi: 10.1038/s42254-022-00441-7.

[5] K. D. Foote, 'A Brief History of Machine Learning', DATAVERSITY. Accessed: Nov. 05, 2023. [Online]. Available: https://www.dataversity.net/a-brief-history-of-machine-learning/

[6] 'Regression vs. Classification in Machine Learning: What's the Difference?' Accessed: Nov. 06, 2023. [Online]. Available: https://www.springboard.com/blog/data-science/regression-vs-classification/

[7] Seldon, 'Supervised vs Unsupervised Learning Explained', Seldon. Accessed: Nov. 05, 2023. [Online]. Available: https://www.seldon.io/supervised-vs-unsupervised-learning-explained