Question 1

1.1 Four steps to implement rule-based approach in decision-making

1. Create a table for the documents you want to classify and populate it.
2. Create a rule table. The rule table consists of categories that you name, such as "medicine" or "finance," and the other rule that sort documents into those categories.
3. Create a CTXRULE index on the rule table.
4. Classify the documents.

1.2 Overfitting

The main goal of each machine learning model is having a well generalization. And generalization refers to the ability of a model to perform well (give nearly accurate results) when given unseen data.

The term **overfitting** in machine learning refers to a model that performs well on trained dataset but provides inaccurate results on unseen data. The overfitted model has **low bias, high variance,** and **low generalization** [1].

Why is overfitting a problem?

As mentioned above, the quality of a good model is evaluated based on the performance of a model when dealing with unseen data.

So, overfitting is a big issue in machine learning because while the evaluation results of a model on a trained dataset is different to the evaluation results of a model on the unseen data. And what people care about, is the quality of a model on new unseen data rather than on the trained dataset [2].

If is not cared on, this effect (bad statistical results from overfitted model) of overfitting may potentially lead to bad decisions on the whole population because of the wrong results from sample datasets done by an over-fitted model.

According to Occam's razor, the simpler a model is, the more accurate it would be. So, if give a simple data set, I would prefer a simple model with one parameter over a complex one, because the more you make a model complex, the more it might be likely to overfit.

1.3 Two commonly used approach to avoid overfitting

- **Cross-validation:** This approach is helpful to avoid overfitting because the approach ensure that a train data set is splitted into different portion and those different portions are used to train a model. It is clear that a model will be trained more than once (using all data set portions), and the more time it is trained, the more its accuracy will be examined. Therefore, it might me so rare to get confused whether the accuracy of the model is questionable or not. Briefly, the more times you train a model on new unseen portions, you may identify its accuracy on new unseen real datasets by examining the mean error from all tests. One of the most commonly used technique

to perform cross-validation is **k-fold,** where a train data set is divided into k different parts and each of them used to train a model. A single portion is used as testing data and the remaining k-1 portions play as training data. Normally, **k-fold** approach is used to detect overfitting, so that it can be prevented.

- **Training a model with more data and examples:** It is so important to train a model using enough data that would contains all possible cases a model would face in the future. However, an attention is needed because as dataset is enlarged, we should make sure that data are clean and relevant [3]. If we adjust a data by adding noisier data, we are furtherly confusing a model instead of avoiding overfitting. Therefore, we need to use clean and relevant data.
- **Reduce the model complexity:** According to Occam's razor, the simpler a model is, the better it might be. The complexity of a model is just number of features or terms in a given model. When there are some irrelevant features that a model has (irrelevant complexity), better to remove them, as they may lead to inaccurate model prediction.[3] This also might be helpful to reducing the probability of a model to overfit tests.

1.4 Two metrics for evaluating the performance of a model

- Confusion Matrix
- Cross validation

1.5 Why benchmarks are useful in machine learning

It is important as it is used to select the best solution (technology) in machine learning.

Examples, inference speed and model precision are examples of bencmarks

2.

2.1 Machine learning

Machine learning is an area in Artificial Intelligence that refers to the study of developing the computer algorithms that have the ability to automatically through experience and the use of data that they deal with every day they are running. This new field is very important in different fields of our daily life including medicine, business, sport, leisure, etc… One of the best important aspects of machine learning (algorithms) is that this application can predict the future situation based on the past data patterns and this allows people to take decisions based on the results of machine learning model.

History of machine learning

As found at [6] It was in 1950 when one of the most brilliant and influential mathematician and computer scientist from Britain called Alan Turing created Turing Test that has the purpose of determining whether the computer might have human intelligence. The computer needs to convince people that it is another person in order to pass the Turing test. Not later, in 1952, an American guy called **Arthur Samuels** just wrote a first machine learning algorithm

named game checkers. From these ideas, many different people continued to think about the possible ways to develop systems that may improve their performance as they keep learning from experience, data, and information. **In 1990s** scientists and data researchers created computer programs that could analyze large amounts of data and draw some conclusions, and this led the development of data Science. Later in 2006, the term deep learning came around, and one of the important years for machine learning was 2010-2011 where IBM and Google companies developed deep neural network application that could classify or categorize objects. Since around 2012-2014, big fish companies in the world got interested in machine learning and did some interesting application of machine learning such as algorithm to identify YouTube videos containing cats **by Google**, Deep Face application by **Facebook, etc… In 2016** Google developed one of the world's most complex board game and **recently in 2020** open Artificial Intelligence launched an exceptional and advanced language processing model called **GPT-3.**

2.2 Examples of machine learning (supervised and unsupervised)

In **supervised learning**, there is **classification technique and regression technique.**

These two techniques are both of supervised technique because they all use well labelled data while training and from those data, they predict the outcome or classify the data.

The main different between the two is that classification aims at categorizing the data based on different parameters while regression aims at predicting the outcome in continuous manner (in the future) [9].

In unsupervised learning, we may say about **clustering technique and Association technique.**

Clustering refers to purchasing like behaviours of data and discover the inherent groupings in the data. For instance, grouping the customers of a company from their behaviours.

Association refers to discovering the rules that might describe the large part of the data. For instance, this might be helpful while finding like people that by m product highly tend to buy n product again.

Some of the important algorithms used in unsupervised learning include **k-means for clustering [9].**

2.3 Difference between classification and regression

Classification is the process of calculating a function that divide the dataset into different classes depending on different parameters and criteria. In machine learning, the computer program with this algorithm is trained to take a dataset and classify the data based on different parameters. It is used for discrete data and can be used to solve classification problems such as speech recognition, identification of cancer cells, etc…

While regression is a process of finding the correlation between predictor variables and dependant variables. This machine learning technique in predicting variables continuously, where it is trained with data patterns and learn to predict the future value of dependent variable given an independent variable(s). it is used for continuous data and can be used to solve regression problems such as weather prediction, house price prediction, etc… [7]

2.4 Difference between supervised and unsupervised learning

Supervised learning is the branch of machine learning by which a model is trained with well labelled data and produce a correct outcome [8]. It is trained for the purpose of being able to predict the output in the future when given the unseen data.

While **unsupervised learning** is another machine learning technique by which a model is given unlabelled input data and try to find the pattern and the structure from those input data. It does it on its own because the data given are not labelled. It is able to find the hidden patterns in an unknown dataset.

2.5 Examples of successful applications of machine learning and technique if learning involved.

Youtube: Youtube might use unsupervised learning. For instance, when it is classifying different videos that are similar, or has the same rhythm or same beat, this method might be taken as unsupervised learning as the data are not labelled.

Smartphone auto complete: This system might use supervised learning, because it depends on the words that a user use to type more often and the sequence a user use to write them and from those data, if a user starts typing the same first word again, a system suggests them the probable next words.

3.

Libraries used:

- **Pandas** (import pandas as pd)
- **Seaborn** (import seaborn as sns)
- **Matplotlib** (import matplotlib.pyplot as plt)

Steps to solution

Step 1. I loaded data from the file with the help of pandas library.

Step 2. I have cleared data and extract data for explanatory variable.

Step 3. Printed the correlation matrix using **DataFrame.corr()** function.

Step 4. Printed heat-map using **seaborn.heatmap()** function

Here is the correlation matrix that I came with.

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6
AGE	1.000000	0.173737	0.185085	0.335428	0.260061	0.219243	-0.075181	0.203841	0.270774	0.301731
SEX	0.173737	1.000000	0.088161	0.241010	0.035277	0.142637	-0.379090	0.332115	0.149916	0.208133
BMI	0.185085	0.088161	1.000000	0.395411	0.249777	0.261170	-0.366811	0.413807	0.446157	0.388680
BP	0.335428	0.241010	0.395411	1.000000	0.242464	0.185548	-0.178762	0.257650	0.393480	0.390430
S1	0.260061	0.035277	0.249777	0.242464	1.000000	0.896663	0.051519	0.542207	0.515503	0.325717
S2	0.219243	0.142637	0.261170	0.185548	0.896663	1.000000	-0.196455	0.659817	0.318357	0.290600
S3	-0.075181	-0.379090	-0.366811	-0.178762	0.051519	-0.196455	1.000000	-0.738493	-0.398577	-0.273697
S4	0.203841	0.332115	0.413807	0.257650	0.542207	0.659817	-0.738493	1.000000	0.617859	0.417212
S5	0.270774	0.149916	0.446157	0.393480	0.515503	0.318357	-0.398577	0.617859	1.000000	0.464669
S6	0.301731	0.208133	0.388680	0.390430	0.325717	0.290600	-0.273697	0.417212	0.464669	1.000000

Figure 1: Correlation matrix

And this is the heat-map that I came up with.

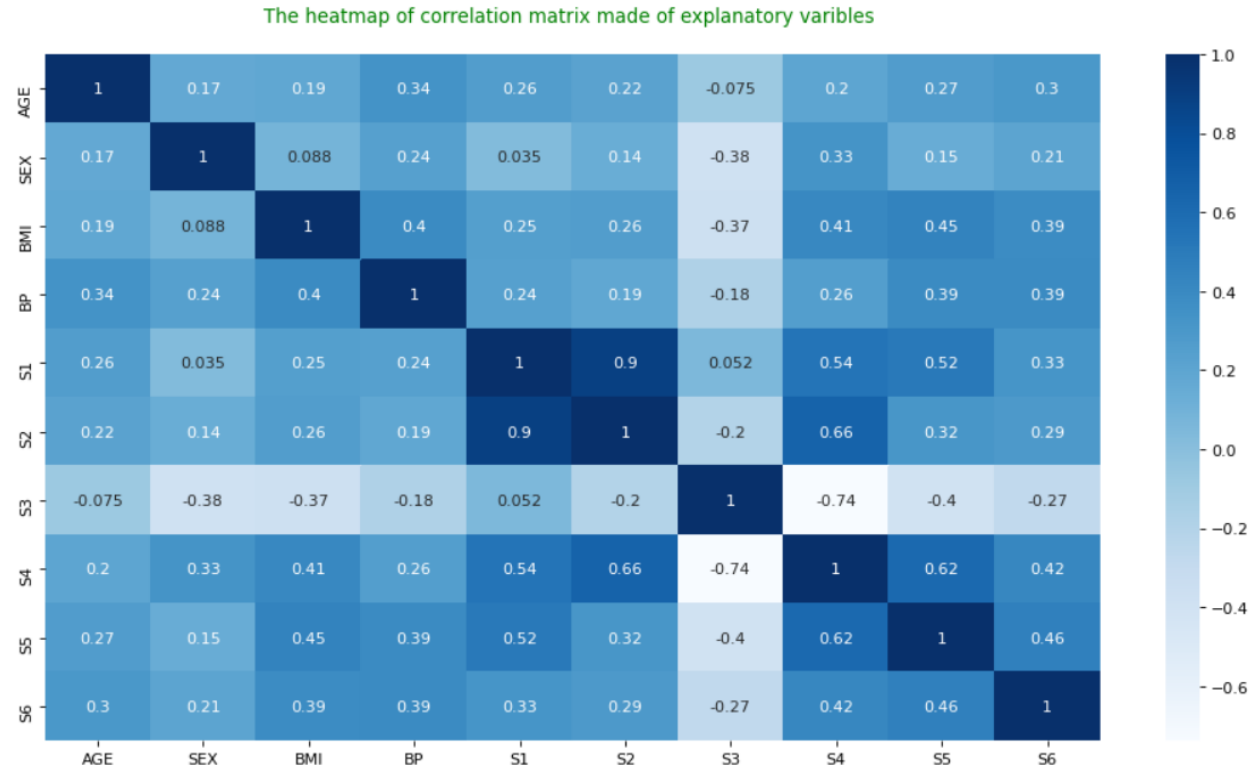

Figure 2: Correlation Heatmap

3.1 Key insights from heat-map and correlation matrix

From the above corelation heat-map image, it is obvious that most two explanatory variables which are highly correlated are (**S1 and S2).** The other pairs that seem to be strongly correlated are (**S2 and S4)**, (**S4 and S5),** (**S1and S4),** and **(S1 and S5).** These are only pairs whose correlation appear to be strong (means are above 0.5). The other remaining have weak correlation which is below 0.5. The variables with the weakest correlation are (**S3 and S4)**,

(**S3 and S6),** (**BMI and S3), (SEX, S3), etc…**

3.2. The collinearity refers to the dependence between the predictor variables. Collinear variables have an exact linear relationship between themselves. It means that the collinear predictor variables, each of them cannot independently predict the value of the dependent variable, but they need to be together to give accurate prediction.

The main effect of collinearity is the estimation of predictor variables while estimating the relationship between a one predictor variable and the dependant variable (Y), and the estimation results are not precisely correct as it would be if predictor variables would be uncorrelated. So, one predictor variable cannot independently predict the value of dependent variable and it makes it difficult to identify a relationship between individual predictor variable and the dependent variable (Y) even though there may be a strong relationship between them.

3.3 Create multivariate model4

Libraries used:

- **Statsmodel.api** (import statsmodels.api as sm)
- **sklearn.metrics** (from sklearn.metrics import mean_squared_error)

Steps and process

To create this model, I used the diabetes data used on sub-question one.

Step 1. I created a model 1 called **model1** using sm.OLS(Y,x)

Step 2. I fitted my model using fit ()

Step 3. I predicted Y by passing x data in the model

Step 4. To find the ADJUSTED_R_SQUARE, I extracted it from the model summary.

Step 5. To find the MEAN_SQUARED_ERROR, I used **mean_squared_error(Y, predicted_Y)**

What is MSE and Adjusted R square

```
Mean Squared Error (MSE): 2876.683251787016
R_Squared:  0.515
```

Are all variable significant?

No, all variables are not significant, because from the table of mode summary that shows th e p-values of each variable, it is obvious that some variables have a p-value above 0.05. So, t hese variables with such values are not significant.

Here are the significant variables, and are sorted from the most significant one to the lowest significant variable

```
The significant variables are: ['BMI', 'BP', 'S5', 'SEX']
```

Here is the model summary.

```
                    Results: Ordinary least squares
==========================================================================
Model:                OLS                Adj. R-squared:       0.507
Dependent Variable:   Y                  AIC:                  4793.9857
Date:                 2021-11-01 19:06   BIC:                  4838.9901
No. Observations:     442                Log-Likelihood:       -2386.0
Df Model:             10                 F-statistic:          46.27
Df Residuals:         431                Prob (F-statistic):   3.83e-62
R-squared:            0.518              Scale:                2932.7
--------------------------------------------------------------------------
      coefficient standard_error    t    p_values   0.025     0.975
--------------------------------------------------------------------------
const   -334.5671        67.4546 -4.9599   0.0000 -467.1481 -201.9862
AGE       -0.0364         0.2170 -0.1675   0.8670   -0.4630    0.3902
SEX      -22.8596         5.8358 -3.9171   0.0001  -34.3299  -11.3894
BMI        5.6030         0.7171  7.8133   0.0000    4.1935    7.0124
BP         1.1168         0.2252  4.9583   0.0000    0.6741    1.5595
S1        -1.0900         0.5733 -1.9012   0.0579   -2.2169    0.0369
S2         0.7465         0.5308  1.4062   0.1604   -0.2969    1.7898
S3         0.3720         0.7825  0.4754   0.6347   -1.1659    1.9099
S4         6.5338         5.9586  1.0965   0.2735   -5.1778   18.2454
S5        68.4831        15.6697  4.3704   0.0000   37.6846   99.2817
S6         0.2801         0.2733  1.0249   0.3060   -0.2571    0.8173
--------------------------------------------------------------------------
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         Condition No.:            7236
==========================================================================
* The condition number is large (7e+03). This might indicate
strong multicollinearity or other numerical problems.
```

Result 3

Yes, this is the problem of collinearity

3.4 Forward selection VS backward selection

Both forward and backward selection are the methods used to select features of a model in machine learning, but the way each of them is implemented are different.

Forward selection is a selection technique that is works in iterative process. It starts with **empty (null) model** and at each **iteration** it **keeps adding the new features** and **evaluate** the performance of a model **to check whether the new added feature does improve the model performance or not**. This process **continues until the new added** features are reducing the model performance [4],[5].

While backward selection is also an iterative approach that works in the opposite direction of forward selection approach. This is how it works.

The backward selection starts with a full featured model and at each iteration it removes the least significant feature (variables with the largest p-value) one at time and keeps evaluating the performance of the model. It continues the process until removing the features is not improving the model performance [4],[5].

3.5

How stepwise approach work while selecting variables?

In order to select variable, stepwise approach works in an iterative manner. It adds or remove one variable at time, and at each iteration it measures whether an action of adding or removing a variable is improving the model performance or not. The process continues until the adding or removing is not improving the model variable or there are no remaining variables to select from.

Using forward_regression function, compose a model in interactive way.

I have used the forward_regression method as given in previous assignment, but as I needed to view the steps, I made verbos=True, and I left some seconds between loop iteration.

By using the threshold= 0.05, I have interactively composed the model and show steps by steps how a model selected variable at time. The following image shows all steps after all running, and by running the codes you will see how step by step the model added new variables.

```
Steps to select significant variable

Iteration 1:     Add  const          with p-value 2.40471e-154
Remaining Variables: ['BMI', 'S4', 'S1', 'AGE', 'const', 'SEX', 'S2', 'BP', 'S6', 'S5', 'S3']
Selected Variables: ['const']


Iteration 2:     Add  BMI            with p-value 3.46601e-42
Remaining Variables: ['BMI', 'S4', 'S1', 'AGE', 'SEX', 'S2', 'BP', 'S6', 'S5', 'S3']
Selected Variables: ['const', 'BMI']


Iteration 3:     Add  S5             with p-value 3.03963e-20
Remaining Variables: ['S4', 'S1', 'AGE', 'S2', 'SEX', 'BP', 'S6', 'S5', 'S3']
Selected Variables: ['const', 'BMI', 'S5']


Iteration 4:     Add  BP             with p-value 3.74262e-05
Remaining Variables: ['S4', 'S1', 'AGE', 'S2', 'SEX', 'BP', 'S6', 'S3']
Selected Variables: ['const', 'BMI', 'S5', 'BP']


Iteration 5:     Add  S1             with p-value 0.00145443
Remaining Variables: ['S4', 'S1', 'AGE', 'S2', 'SEX', 'S6', 'S3']
Selected Variables: ['const', 'BMI', 'S5', 'BP', 'S1']


Iteration 6:     Add  SEX            with p-value 0.00923056
Remaining Variables: ['S4', 'AGE', 'S2', 'SEX', 'S6', 'S3']
Selected Variables: ['const', 'BMI', 'S5', 'BP', 'S1', 'SEX']


Iteration 7:     Add  S2             with p-value 0.000272302
Remaining Variables: ['S4', 'AGE', 'S2', 'S6', 'S3']
Selected Variables: ['const', 'BMI', 'S5', 'BP', 'S1', 'SEX', 'S2']


The final selected variable are: ['const', 'BMI', 'S5', 'BP', 'S1', 'SEX', 'S2']
```

Figure 3: Composing a model

Which variables are selected?

As shown on an image above, the selected variables are **BMI, S5, BP, S1, SEX, S2** and adding **const**.

How does this function work?

Forward selection function starts with an empty set that will hold the selected variables. A function works in an iterative manner and at each iteration, it adds a most significant variable in a selected variable set and keeps checking the performance of the model. The process continues until adding a new variable does not improve the model performance.

What is the MSE and R^2

```
Mean Squared Error (MSE): 3887.0284449370897
Adjusted R_Square:    0.866
```

4.

4.1 Difference between Linear Regression and Logistic regression

Both regressions are most famous machine algorithms that are used in supervised learning for solving regression problems.

Although they are all classified in supervising learning, they are different from each other.

Linear Regression: It is a technique used to predict continuous dependent variables with the help of predictor variables known as independent variables. If one independent variable is used to predict, we call it Simple Linear regression and when more than one independent variable is used, we call it multiple Regression. What this regression tells gives is a best fit line that can predict the future behaviour of dependent variable as the data continues to grow. To do that we just find the best fit line that in form of **y=a+bx** that can help us to easily predict the output. Briefly, Linear Regression is used to solve regression problems

On the other hand, logistic regression is a prediction method in Machine learning that is mainly used for classification problem. The main intention of a logistic regression to categorize dependent variables using (based on) independent variable. It is more often goes with the probability of a value to be in some categories or not. To do the estimation and/or the classification of samples, Sigmoid Curve (S-curve) function is used is used in Logistic regression.

4.2 I loaded titanic data using pandas library and found the probability of survival passengers.

```
The probability is 0.3819709702062643
```

4.3 To provide a table that gives the data about survival probabilities, I have found probabilities by grouping data based on different attributes. One wa sex, pclass and ages.

In ages I have divided the data into 4 categories (children, youth, adult and old).

Then after finding the probabilities, I have combined all of them in one data frame

The dataframe is

	survived
adult	0.355383
children	0.573913
old	0.400000
youth	0.371951
1	0.619195
2	0.429603
3	0.255289
female	0.727468
male	0.190985

4.4 I predicted using Logistic Regression and provided the summary of the model

Sumary

```
==============================================================================
Dep. Variable:                survived   No. Observations:                 1309
Model:                           Logit   Df Residuals:                     1306
Method:                            MLE   Df Model:                            2
Date:                 Mon, 01 Nov 2021   Pseudo R-squ.:                  0.1939
Time:                         23:40:51   Log-Likelihood:                -701.76
converged:                        True   LL-Null:                       -870.51
Covariance Type:            nonrobust   LLR p-value:                5.142e-74
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
age            0.0255      0.004      6.986      0.000       0.018       0.033
sex           -2.0736      0.136    -15.272      0.000      -2.340      -1.807
pclass        -0.0768      0.048     -1.592      0.111      -0.171       0.018
==============================================================================
```

4.5 I calculated the accuracy and the confusion matrix

```
accuracy: 0.7853323147440795
```

confusion matrix

```
686  123
158  342
```

THIS MARKS THE NED OF THE REPORT

Reference

[1] "Overfitting and Underfitting in Machine Learning - Javatpoint", *www.javatpoint.com*, 2021. [Online]. Available: https://www.javatpoint.com/overfitting-and-underfitting-in-machine-learning. [Accessed: 25- Oct- 2021].

[2] H. Schneider, D. Schmidt, T. Elders, T. Saxena and R. Chumley, "Master Machine Learning Algorithms", *Machine Learning Mastery*, 2021. [Online]. Available: https://machinelearningmastery.com/master-machine-learning-algorithms/. [Accessed: 25- Oct- 2021].

[3]"Overfitting in Machine Learning: What It Is and How to Prevent It", *EliteDataScience*, 2021. [Online]. Available: https://elitedatascience.com/overfitting-in-machine-learning. [Accessed: 29- Oct- 2021].

[4]"Feature Selection Techniques in Machine Learning - Javatpoint", *www.javatpoint.com*, 2021. [Online]. Available: https://www.javatpoint.com/feature-selection-techniques-in-machine-learning. [Accessed: 01- Nov- 2021].

[5]"Statistics - Forward and Backward Stepwise (Selection|Regression)", *Datacadamia - Data and Co*, 2021. [Online]. Available: https://datacadamia.com/data_mining/stepwise_regression. [Accessed: 01- Nov- 2021].

[6] A. Transformation and t. Agile way of working, "A brief history of machine learning - Concise Software", *Concise Software*, 2021. [Online]. Available: https://concisesoftware.com/history-of-machine-learning/. [Accessed: 01- Nov- 2021].

[7] "Regression vs Classification in Machine Learning - Javatpoint", *www.javatpoint.com*, 2021. [Online]. Available: https://www.javatpoint.com/regression-vs-classification-in-machine-learning. [Accessed: 01- Nov- 2021].

[8] "Supervised and Unsupervised learning - GeeksforGeeks", *GeeksforGeeks*, 2021. [Online]. Available: https://www.geeksforgeeks.org/supervised-unsupervised-learning/. [Accessed: 01- Nov- 2021].

[9]. Brownlee, "Supervised and Unsupervised Machine Learning Algorithms", *Machine Learning Mastery*, 2021. [Online]. Available: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/. [Accessed: 01- Nov- 2021].