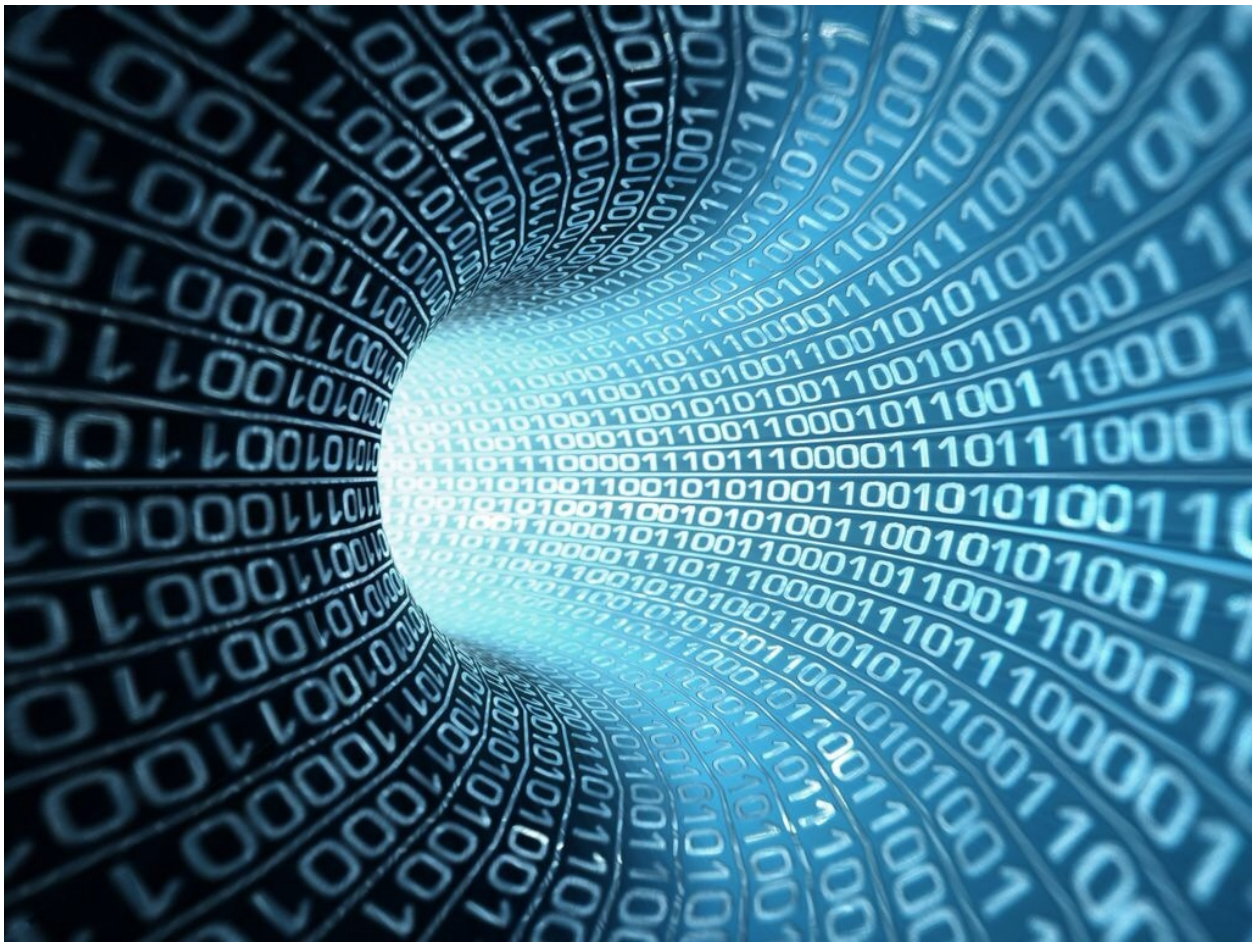
Report: DIAML Assignment 5



I, the undersigned, have read the entire contents of the syllabus for course 18-785 (Data Inference and Applied Machine Learning) and agree with the terms and conditions of participating in this course, including adherence to CMU's AIV policy.

Signature: Tunga Tessema

Andrew ID: tchamiss

Full Name: Tunga Tessema

Librairies

- **Pandas:** to read excel and csv files, convert them to dataframes and perform operations on them
- **numpy:** to create an array, calculate square roots and calculate correlation coefficients
- **matplotlib.pyplot:** to plot graphs
- **Seaborn**: to create a heatmap
- **tabulate:** to create a table
- **statsmodels.api**: to calculate linear and logistic regression models
- **sklearn.model_selection**: to split the data into training and testing datasets
- **sklearn.metrics**: to calculate confusion matrix and display it

Question 1

1.1 Describe at least four steps to implementing a rule-based approach to decision-making and give an example. Is any domain knowledge required to establish a rule? Support your answer with an explanation.

Rule Based Approach refers to the AI modeling where the relationship or patterns in data are defined by the developer. The machine follows the rules or instructions mentioned by the developer, and performs its task accordingly.
To implement a rule-based approach, we need the following steps:

1. Knowledge base = a set of rules that are condensed into if-then statements. The rules are found by talking to an expert and trying to find patterns that the expert uses to make a decision
2. Database of facts = the facts that describe the current situation
3. Inference engine = processes the rules and the facts and comes up with one or more conclusions
4. Explanation mechanism = the system explains the rules and the facts used in coming up with the conclusion

For example, if we take the case of a doctor diagnosing a patient. The rules will correspond to the years of medical training they had. They have learned that if a patient has the following symptoms and they have it for a specified period of time then make a specific diagnosis. The next step for the doctor would be to gather facts about the patient. The doctor will do measurements and ask questions. The doctor then uses the facts and measurements he measured to make a diagnosis. He then explains his reasoning to the patient.

A domain knowledge is required to establish the rules. For example, in the above example that we gave, if someone is not trained as a medical doctor, they will not be able to come up with the rules, questions and diagnosis.

1.2 Explain over-fitting and why it is a problem in statistical learning. If you have a small dataset containing ten data points, should you prefer a simple model with one parameter or a complex model with ten parameters? Support your answer with an explanation.

Overfitting is when a model fits exactly against its training data. This is a problem because the algorithm cannot perform accurately against unseen data, defeating its purpose. Generalization of a model to new data is ultimately what allows us to use machine learning algorithms to make predictions and classify data.
If we have a small dataset with ten data points we should choose a simple model with one parameter. If we choose a complex model with ten parameters the model runs the risk of overfitting because using all the parameters available it will start to learn the noise within the dataset.

1.3 There are two commonly used approaches to avoid over-fitting; describe each one.

The first approach is to train the model with more data. Adding more data can increase the accuracy of the model by giving it a chance to see the dominant relationships between the input and the output. The second approach is to reduce the number of independent variables. When trying to build a model we will have a number of parameters to consider. But some of these parameters are irrelevant and some redundant. By removing the redundant or irrelevant parameters, we simplify or model and help it focus on the general trend.

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

The coefficient of determination, R^2, measures the proportion of variability in a data set that is accounted for by a statistical model. It is calculated by the following formula:

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \quad SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2 \quad SS_{res} = \sum_i (y_i - \hat{y}_i)^2$$

and R^2 is defined as

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 is used to compare the goodness of fit of 2 or more models. The greater the value of R^2, the more the model fits the value.

The mean absolute error (MAE) is the average absolute value of the difference between the forecasted value and the actual value. This measure focuses on the magnitude of the errors. It is calculated by the following formula:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} | \hat{y}_i - y_i |$$

MAE is commonly used in wind energy forecasting and may be given as a fraction of the total energy being generated

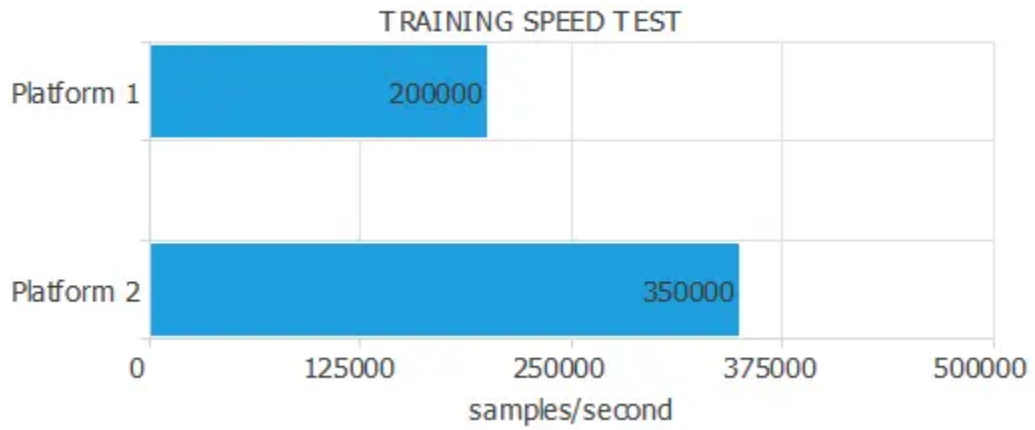
1.5 Why are benchmarks useful in machine learning and give two examples.

In machine learning, benchmarking is the practice of comparing tools to identify the best-performing technologies in the industry. The volume, variety, and velocity of information stored in organizations are increasing significantly. Therefore, for machine learning tools to be efficient, they need to process large amounts of data in the shortest time possible.

The first example of a benchmark is training speed. Training speed is usually measured as the number of samples per second that the platform processes during training.
To compare the training speed of machine learning platforms, we follow the next steps:

- Choose a reference benchmark (data set, neural network, training strategy...).
- Choose a reference computer (CPU, GPU, RAM...).
- Compare the training speed.

The following figure illustrates the result of a training speed test with two platforms.



As we can see, the training speed of Platform 1 is 200,000 samples/second, while that of platform 2 is 350,000 samples/second. So we would choose platform 2.

The second example of a benchmark is model precision. We can define precision as the mean error of a model against a testing data set. We follow the next steps to compare the precision of different machine learning platforms:

- Choose a reference computer (CPU, GPU, RAM...).
- Choose a reference dataset (variables and samples number).
- Choose a reference model (number of layers, number of neurons...).
- Choose a reference training strategy (loss index, optimization algorithm...).
- Choose a stopping criterion (loss goal, epochs number, maximum time...).

Let’s take an example of two platforms A and B and their correlation:

As we can see, Platform A can build a model with a correlation of 0.8. On the other hand, Platform B can build a model with a correlation of 0.9. So we would choose platform B.

2. Machine Learning

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

Machine learning is a branch of computer science that enables machines to learn on their own from past experiences without being explicitly programmed.

The concept of machine learning first appeared in 1950, when Alan Turing published a paper in response to the question "Can machines think?"
Frank Rosenblatt created the first neural network for computers in 1957, which is now known as the Perceptron Model.
Bernard Widrow and Marcian Hoff developed two neural network models, Adeline, which could detect binary patterns, and Madeline, which could eliminate echo on phone lines, in 1959.
The Nearest Neighbor Algorithm, developed in 1967, enabled computers to perform very basic pattern recognition.
In 1981, Gerald DeJonge proposed the concept of explanation-based learning, in which a computer analyzes data and develops a general rule to discard irrelevant information.

During the 1990s, machine learning research shifted from a knowledge-driven to a more data-driven approach. During this time, scientists began developing computer programs that could analyze large amounts of data and draw conclusions or "learn" from the results. This eventually culminated in the modern era of machine learning.

Machine learning is popular because of its wide range of applications and its incredible ability to adapt and provide solutions to complex problems efficiently, effectively and quickly. A few applications of machine learning are virtual personal assistant, facial recognition, email spam filter, recommendation engine .. etc
In addition, there is an abundance of data to learn from and an abundance of computation to run methods (you can get powerful computers for very few dollars using the cloud).

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

Below I have listed 3 examples of machine learning techniques:

1. Classification: Classification in machine learning is where the networks will segment and separate data based on specific rules that you give them. Classifying is used in supervised training for learning algorithms. They will classify the data for you and separate it based on your specifications, so you can serve the results based on the different classes. For example, classification machine learning models can help marketers separate demographics of customers so you can serve them a unique ad based on their classification.

2. Clustering: Clustering is similar to classifying in that it separates similar elements, but it is used in unsupervised training, so the groups are not separated based on your requirements. Clustering is commonly used in machine learning models when researchers are trying to find the differences between sets of data and learn more about them. In data analytics or data science, if a researcher is trying to discover what makes certain groups different, they might try clustering to see if the computer can point out some of the subtle differences.

3. Regression: In machine learning, regression algorithms are used to plan and model, finding the likelihood of a specific variable. Machines are able to look at different variables and forecast their connection, helping leaders understand

what to expect in the future. Regression helps identify connections between data points.

2.3 What is the difference between classification and regression?

Regression algorithms predict a continuous value based on the input variables. The main goal of regression problems is to estimate a mapping function based on the input and output variables. If your target variable is a quantity like income, scores, height or weight, or the probability of a binary category (like the probability of rain in particular regions), then you should use the regression model.

Classification is a predictive model that approximates a mapping function from input variables to identify discrete output variables, which can be labels or categories. The mapping function of classification algorithms is responsible for predicting the label or category of the given input variables. A classification algorithm can have both discrete and real-valued variables, but it requires that the examples be classified into one of two or more classes.

The most significant difference between regression and classification is that while regression helps predict a continuous quantity, classification predicts discrete class labels. While a regression model can be used to predict temperature for the next day, we can use a classification algorithm to determine whether it will be cold or hot according to the given temperature values.

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

Supervised learning is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time.
Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention.
The main distinction between the two approaches is the use of labeled datasets. To put it simply, supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not.

In supervised learning, the goal is to predict outcomes for new data. You know up front the type of results to expect. With an unsupervised learning algorithm, the goal is to get insights from large volumes of new data. The machine learning itself determines what is different or interesting from the dataset.
Supervised learning is a simple method for machine learning, typically calculated through the use of programs like R or Python. In unsupervised learning, you need powerful tools for working with large amounts of unclassified data. Unsupervised learning models are computationally complex because they need a large training set to produce intended outcomes.
Supervised learning models can be time-consuming to train, and the labels for input and output variables require expertise. Meanwhile, unsupervised learning methods can have wildly inaccurate results unless you have human intervention to validate the output variables.

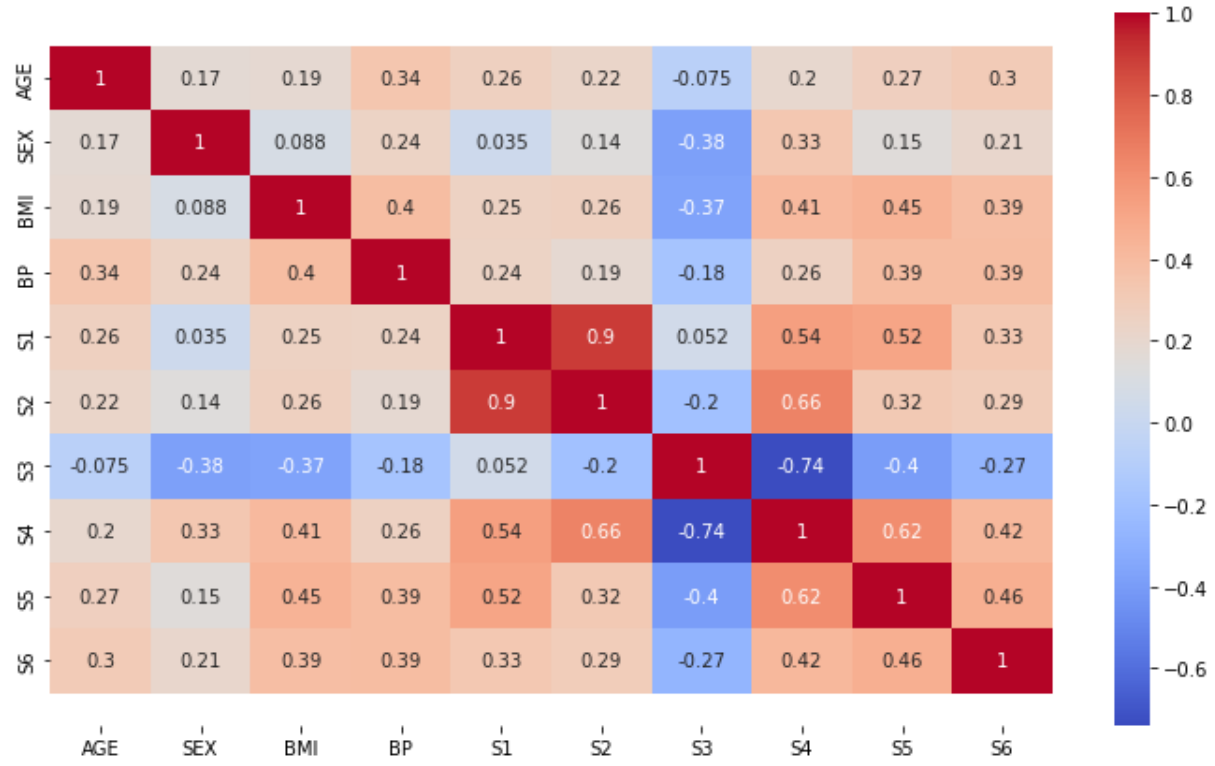
2.5 Give examples of successful applications of machine learning and explain what technique is appropriate and what type of learning is involved?

Here are 3 examples of successful machine learning applications:

- Facial recognition = Facial recognition is a method of identifying or verifying a person's identity by looking at their face. People can be identified in photographs, films, or in real-time using facial recognition technology. Facial Recognition uses deep learning. Deep learning is an artificial intelligence subset of machine learning that uses neural networks to learn unsupervised from unstructured or unlabeled data.
- Email spam filter = Spam emails are junk emails, an email sent without explicit consent from the recipient. Spam emails are used as marketing techniques to promote products and services in order to turn a profit. But nowadays, spam is progressively being viewed as a more severe messaging threat, as it is coming to be used to deliver worms, viruses, and trojans. Email spam filters use supervised learning and the classification technique.
- Recommendation engine = Recommendation engines leverage machine learning to recommend relevant products to users. They use unsupervised learning techniques such as clustering.

3. Diabetes data

3.1. Load the diabetes data into MATLAB or Python from here. Produce a correlation matrix of the explanatory variables. Make a heat-map of the matrix and describe the relationships between the variables.



I calculated the correlation of the variables by using the corr function on the data frame. I then used the heatmap function of seaborn library to create the above heatmap. As we can see on the above heatmap, S1 and S2 have a correlation of 0.9 (they are highly correlated). We can also see that S3 and S4 have a correlation of -0.74 (they are correlated highly in a negative way.) S4 and S5 are correlated moderately with 0.62 correlation. S5 and S6 are also moderately correlated with a correlation coefficient of 0.46.

3.2. What is collinearity? What effect does collinearity amongst predictor variables have on their estimated coefficient value?

Collinearity occurs because independent variables that we use to build a regression model are correlated with each other. This is problematic because as the name suggests, an independent variable should be independent. It shouldn't have any correlation with other independent variables.

If collinearity exists between independent variables, the key point of regression analysis is violated. In regression analysis, we want to isolate the influence of each independent variable to our dependent variable.

There are several things how collinearity would affect our model, which are:

- The coefficient estimates of independent variables would be very sensitive to the change in the model, even for a tiny change. Let's say we want to remove or add one independent variable, the coefficient estimates then will fluctuate massively. This makes it difficult for us to understand the influence of each independent variable.
- Collinearity will inflate the variance and standard error of coefficient estimates. This in turn will reduce the reliability of our model and we shouldn't trust the p-Value that our model showed us to judge whether an independent variable is statistically significant for our model or not.

3.3. Create a multivariate linear model using all ten variables and a constant. In the rest of this assignment this model will be referred to as model1. What are the Mean Squared Error and the adjusted R2 for model1? Are all variables significant? Could this be a problem of collinearity?

```
model1 adjrsqaured: 0.5065592904853231
model1 mse: 2859.6963475867506
```

	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

To create model1, I used all 10 variables. I then used the rsquared_adj property and got a value of 0.506. To calculate MSE, I used a custom function where I implemented the

formula. I got an MSE of 2859.69. As you can see on the above table, sex, BMI, BP and S5 are the only significant variables. They are significant because their p-value is less than 0.05. This could be a problem of collinearity because the variables that we saw have high correlation like s1 and s2 & s3 and s4 are all not significant.

3.4. What is the difference between forward selection and backward selection?

Forward selection is a variable selection method which:

- Begins with a model that contains no variables (called the Null Model)
- Then starts adding the most significant variables one after the other
- Until a pre-specified stopping rule is reached or until all the variables under consideration are included in the model

Backward selection is a variable selection method which:

- Begins with a model that contains all variables under consideration (called the Full Model)
- Then starts removing the least significant variables one after the other
- Until a pre-specified stopping rule is reached or until no variable is left in the model

3.5. How does the approach stepwise work in the sense of selecting variables? Use the function stepwise to interactively compose a model using forward selection. Which variables are selected? How does this function work? What is the MSE and R2 value for this new model?

```
the chosen independent variables: ['BMI', 'S5', 'BP', 'S1', 'SEX', 'S2']

model2 rsqaured: 0.5148837959256445
model2 mse: 2876.683251787016
```

In the stepwise approach, there are many possible criterias to select the variables:

- It has the smallest p-value, or
- It provides the highest increase in R2, or
- It provides the highest drop in model RSS (Residuals Sum of Squares) compared to other predictors under consideration.

The function I used uses the variable with the smallest p-value. The selected variables are: BMI, S5, BP, S1, SEX and S2. The mse of this model is 2876.68 and its r-squared is 0.514

4. Analyzing the Titanic data set
4.1 What is the difference between logistic regression and linear regression?

Here is a list of the differences between logistic and linear regression:

1. A linear regression model is used when the response variable takes on a continuous value such as price, height, age, distance. Conversely, a logistic regression model is used when the response variable takes on a categorical value such as: Yes or No, Male or Female, Win or Not Win.

2. Linear regression uses the following equation to summarize the relationship between the predictor variable(s) and the response variable:

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

where:

- Y: The response variable
- X_j: The j^{th} predictor variable
- β_j: The average effect on Y of a one unit increase in X_j, holding all other predictors fixed

In contrary, logistic regression uses the following equation:

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

This equation is used to predict the probability that an individual observation falls into a certain category.

3. Linear regression uses a method known as ordinary least squares to find the best fitting regression equation. Conversely, logistic regression uses a method known as maximum likelihood estimation to find the best fitting regression equation.

4.2 Load in the titanic dataset and calculate the probability of survival for a passenger on the titanic.

```
the probability of survival of a passenger: 0.3819709702062643
```

I loaded the titanic dataset using pandas dataframe. I took the sum of the “survived" column and divided it by the total number of passengers. I found that 38.2% of the time, a passenger survived from the titanic.

4.3 Provide a table giving survival probabilities broken down by passenger class, gender and age.

Pclass1	Pclass2	Pclass3	male	female	Age12orless	Age13to19	Age20to35	Age35to50	Age50above
0.619195	0.429603	0.255289	0.190985	0.727468	0.574468	0.396947	0.388778	0.39207	0.4

To calculate the probabilities, I used the respective columns and filtered them using a criteria. I then divided their sum by the total number of rows in that dataframe. For the age, I used the following categories:

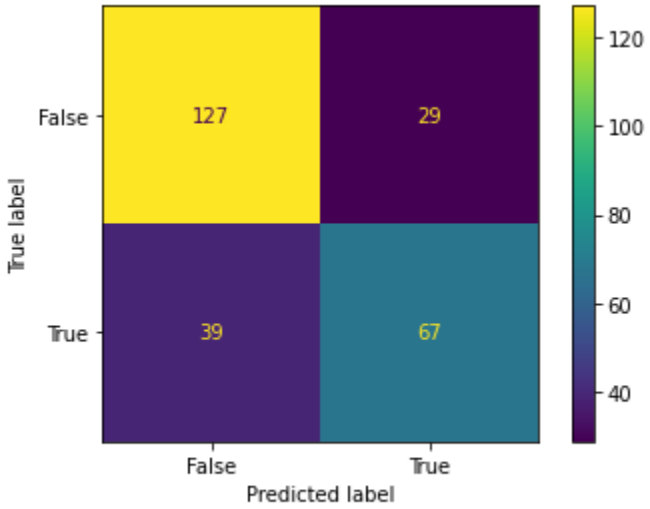
- 12 years or less
- Between 13 and 19
- Between 20 and 35
- 36 to 50 and
- Above 50

4.4 Build a logistic regression model for survival rates based on passenger class, sex and age. What are the parameter estimates and are these parameters statistically significant?

	coef	std err	z	P>|z|	[0.025	0.975]
const	4.9890	0.487	10.238	0.000	4.034	5.944
sex_male	-2.6693	0.198	-13.511	0.000	-3.057	-2.282
age	-0.0397	0.007	-5.371	0.000	-0.054	-0.025
pclass	-1.1977	0.135	-8.898	0.000	-1.462	-0.934

I first selected the wanted columns: survived, sex, age, pclass and dropped NA values, I then changed the sex value to 0s and 1s using the get_dummies function of pandas.I then used the train_test_split function to split the dataset into training and splitting. I used 75% for training and 25% for testing. I then used the sm.Logit function to create the logistic regression model. I found the parameter estimates you see on the table above (under the coef column). The parameters are statistically significant because their p-value is less than 0.05.

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




```
the model's accuracy is: 0.7404580152671756
```

I first created the confusion matrix of the model by using the confusion_matrix function. To make the confusion matrix's display more readable, I used the ConfusionMatrixDisplay function. The previous function lets us label and color the matrix. We can see that our model has 127 true negatives, 67 true positives, 29 false

positives and 39 false negatives. Using the above values we can calculate the accuracy. I found an accuracy of 74%.

References

What is Overfitting? | IBM
Logistic Regression vs. Linear Regression: Key Differences | Indeed.com
Logistic Regression vs. Linear Regression: The Key Differences - Statology
What Is The Difference Between Forward Selection And Backward Selection? | Knologist
A Beginner's Guide to Collinearity: What it is and How it affects our regression model | by Nathan Rosidi | Towards Data Science
Supervised vs. Unsupervised Learning: What's the Difference? | IBM
Regression vs. Classification in Machine Learning: What's the Difference? (springboard.com)
Supervised vs. Unsupervised Machine Learning | Deepchecks
Machine Learning: Definition, Explanation, and Examples (wgu.edu)
What is Machine learning and Why is it Important? | HackerNoon
How to benchmark the performance of machine learning platforms: data capacity, training speed, inference speed and model precision | Neural Designer

(15) Expert System Components - YouTube
Rule Based Systems In Artificial Intelligence -ProfessionalAI.com (professional-ai.com)