ASSIGNMENT 5

Andrew-ID: parmenin

DATA INFERENCE AND APPLIED MACHINE LEARNING (18-785)

11/7/22

Niyomwungeri Parmenide ISHIMWE

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: **Niyomwungeri Parmenide ISHIMWE**
Andrew ID: **parmenin**
Full Name: **Niyomwungeri Parmenide ISHIMWE**

The libraries used:

1. import numpy as np
2. import pandas as pd
3. import matplotlib.pyplot as plt
4. %matplotlib inline
5. import seaborn as sb
6. import statsmodels.api as sm
7. from sklearn.metrics import mean_squared_error, r2_score, confusion_matrix
8. from sklearn.preprocessing import LabelEncoder
9. from sklearn.linear_model import LinearRegression, LogisticRegression

QUESTION 1:

1.1) Some of the steps for implementing a rule-based approach to decision-making are problem identification, identifying the data or observations of the rule, setting criteria or conditions for the rule, setting thresholds or scores, identifying tests to be done, testing the rule against the data, and interpreting the results.

To implement a rule-based approach to decision-making, problem definition is needed first to get an understanding of the situation and problems that need to be addressed in a certain field. The next step is to identify the data or inputs to base on while implementing the rule, and then setting criteria or conditions that the rule is going to be built around. In addition, thresholds, or methods of measuring the scores while the rule-based approach is applied are set, then identifying tests of the rule and implementing those tests against the data while looking at thresholds and identifying trends. After testing, those rules are applied to facts or real-world problems to learn and understand the problems. After applying the rules, the results of the rule are interpreted to see how they may help in solving the known problem, and finally, a decision is made based on the rule results.

Examples of rule-based machine learning approaches are association rule learning and artificial immune systems [1]. In addition, the APGAR score is one of the applications of the rule-based approach while making decisions. It is used to decide if a newborn baby requires medical attention based on five criteria results which are Appearance, Pulse, Grimace, Activity, and Respiration.

While developing any rule, domain knowledge is required because the rule cannot be formulated in a field that people do not know about. For example, to formulate the APGAR score, Virginia Apgar had to be expert or knowledgeable in the medical or health sciences as it was to be applied to medical services. So, she was both a doctor and a researcher to mean that she based on her scientific knowledge and observations to come up with this rule.

1.2) Over-fitting is an issue that occurs when a machine learning model knows well the training data. This means that the model has fully learned all the patterns in the training data and performs the predictions better on data it knows but, it cannot predict correctly on new or testing data that it was not trained on before. Overfitting is a problem in statistical learning because when an overfitted model is fed with testing data that is different from the data the model was trained on, the results can be erroneous or inefficient. This model does not know how to generate outcomes on new input since it has only crammed the patterns in the training data and hence produces a lot of unstable results on new data.

If a small dataset is provided containing ten data points, a simple model would be preferred since it will not generalize and would prevent overfitting. In addition, complex models with many parameters are susceptible to overfitting because they will tend to explain the training data too well [2].

1.3) Two approaches commonly used to avoid over-fitting are cross-validation and training the model with more data. Cross-validation suggests using the original training data to divide the dataset into several small train-test splits, then using these splits to adjust the model. In addition, training the model using more data helps the model to enhance pattern learning or signal detection for algorithms that can be used while predicting or testing [3].

1.4) The two examples of metrics used to evaluate the performance of a model are R squared and accuracy. The R2 is determined by dividing the total sum of squares of errors from the average model (provided by SStot) by the sum of squares of residuals from the regression model (SSres) and then subtracting the result from 1 and it is always between 0 and 100% [4].

$$R^2 = 1 - \frac{SSres}{SStot}$$

In addition, the model's overall accuracy is given by simply dividing the total number of correct predictions done by the model by the total number of predictions. An accuracy score ranges from 0 to 1, with one representing the perfect model and zero otherwise [5].

$$Accuracy\ = \frac{Correct\ prections}{Total\ predictions}$$

Model accuracy is used to identify a perfect model to use while making predictions. This helps companies or anyone to make forecasts and hence better decisions based on the results of the model. Having an accurate model can reduce the errors in forecasts, hence preventing the costs that can be brought by errors in prediction. Moreover, R squared is used to measure the relationship between the model and the dependent variable and allows for determining the errors of the model, which shows that the model with fewer errors produces correct results.

1.5) Benchmarks are useful in machine learning because they help to compare different observations, identify trends, measure performance, forecast, and hence plan for improvements for performing better in the future. They are generally, reference points against which other things can be measured. Referring to the course slides 6B, P.63., examples of benchmarks include persistence which is used commonly in meteorology to analyze and forecast temperatures. Averages, medians, moving averages, neural networks, and so on are also benchmarks used in machine learning.

QUESTION 2:

2.1) Machine learning is a process of learning through data and experience where systems or applications are developed with the ability to learn and apply what they learned to different situations. Machine learning is a subfield of artificial intelligence that emphasizes on using data and algorithms to mimic how people learn, gradually increasing its accuracy. This calls for the design and study of systems that can learn from data and improve with exposure to data or use. Machine learning has evolved over the years as events shown below:

Before 1950

There was the development of mathematical and statistical techniques that were used to analyze different situations.

1950-1980: Machine Learning was being studied in the lab.

Alan Turing developed the Turing Test in 1950 to see if a machine could think like a person. Two years later, IBM's Arthur Samuel coined the term "Machine Learning" for the first time. In the decades that followed, researchers developed the first machine learning (ML) applications for straightforward tasks like enhancing a computer's performance in the game checkers and identifying hazy patterns and forms. During this period, simple algorithms were used to make different research on developing machine learning and it is where the Bayesian approaches for probabilistic inference were introduced.

1980-2000 Machine Learning was integrated from the lab into reality

In 1982, interest in neural networks, a branch of machine learning that studies algorithms based on the organization and operation of the human brain, began to rise once more. Many more developments have been made in the 1990s thanks to the expansion of the internet and the growing availability of digital data. Data-driven methods have replaced knowledge-driven methods in the ML sector. 1997 saw the defeat of the chess world champion by IBM's Deep Blue computer. Since that time, ML has left the lab and has returned to reality. In this period, Machine learning research shifts its emphasis from knowledge to data. The creation of computer programs by scientists to analyze vast amounts of data and conclusions.

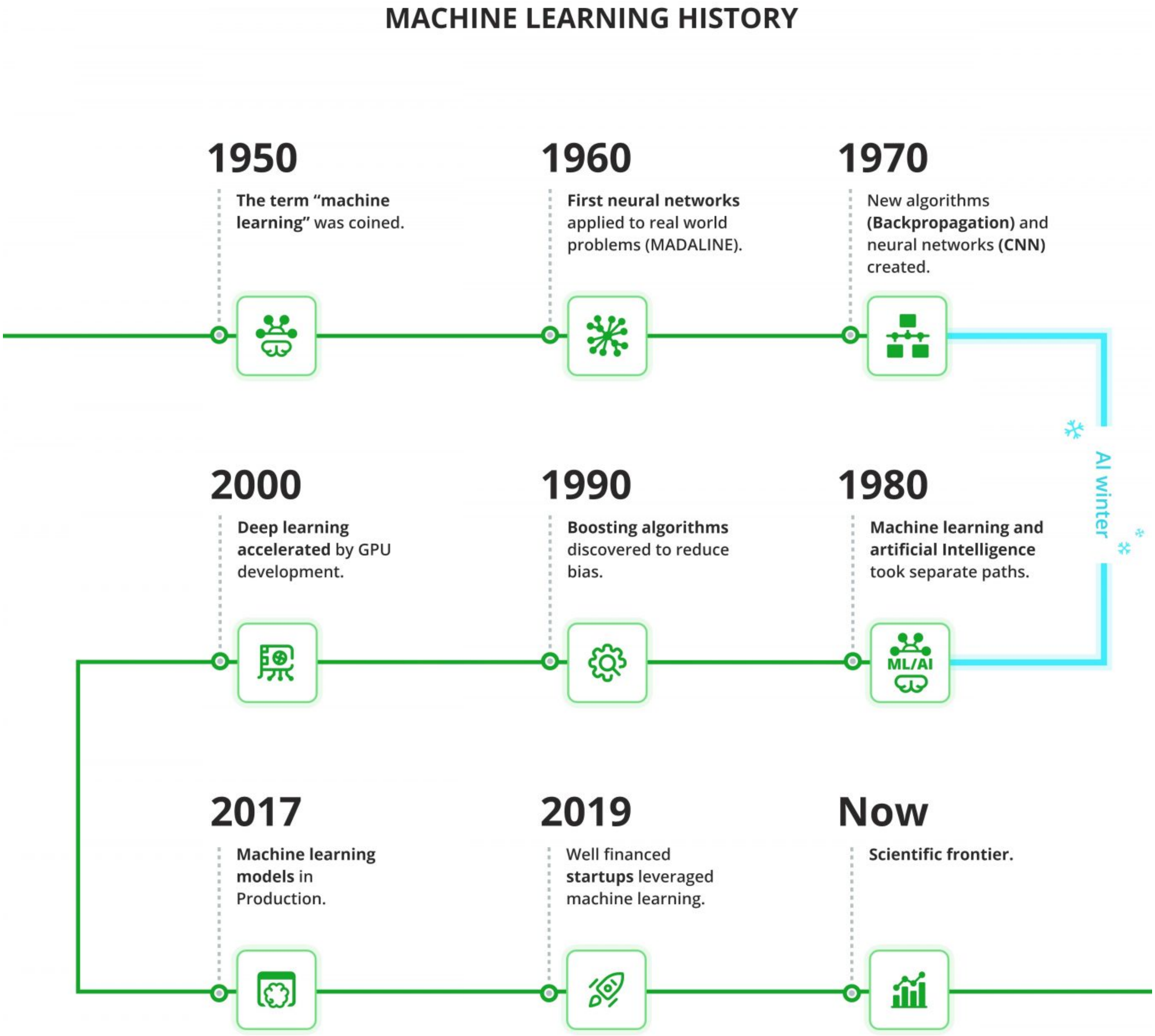

MACHINE LEARNING HISTORY
1950
The term "machine learning" was coined.
1960
First neural networks applied to real world problems (MADALINE).
1970
New algorithms (Backpropagation) and neural networks (CNN) created.
AI winter
1980
Machine learning and artificial Intelligence took separate paths.
ML/AI
1990
Boosting algorithms discovered to reduce bias.
2000
Deep learning accelerated by GPU development.
2017
Machine learning models in Production.
2019
Well financed startups leveraged machine learning.
Now
Scientific frontier.

Figure 1: Machine Learning History [6]

2000-2010: Interest in businesses

More and more companies began focusing on ML around the start of the twenty-first century. Tech firms like Google began to recognize machine learning's promise for using vast data to apply intricate mathematical computations. To keep ahead of their rivals, they are putting a lot of money into the field and conducting extensive research. The two-year-old AI startup Deep Mind was purchased by Google in 2014 for $500M, making it the company's largest European acquisition to date. During this whole period, the adoption of unsupervised machine learning techniques, support-vector clustering, and alternative kernel methods was introduced.

2010-2020: Modern Machine Learning

Due to scarce datasets and high computing costs, ML was ready to take off in the corporate world in the previous 60 years. In the 2010s, these obstacles were destroyed, and ML at last sparked significant improvements in the real world. Massive amounts of data may be stored thanks to a synergy between data collecting and less expensive memory. Machine learning applications are now feasible because of faster processing power made possible by GPU (Graphics Processing Unit) manufacturers like Nvidia. Additionally, tech behemoths like Google speech recognition and Facebook DeepFace face recognition have accelerated the growth of Machine learning into other industries. During this whole period, deep learning becomes feasible, and machine learning is used extensively in many well-known software services and applications [7].

Today’s problems are complex to solve with manual work. Since machine learning provides simple ways to solve them without much human work required, that made it so popular. Machine learning allows the creation of models that can be used in situations where we have big data, and it can be used to forecast future situations using the current data. Machine learning is helping companies in interacting with customers, and in marketing and it is why it is being incorporated into modern software apps. From the course slides (7A, P.12), the growth of machine learning is a result of improvements in computing power, the accessibility of vast amounts of digital information (big data, open data), and improved theoretical knowledge.

2.2) Three machine learning algorithms that can be viewed as supervised learning are regression, decision trees, and random forests. When the output is a real or continuous value, regression techniques are widely employed to generate predictions on numerical values. Regression techniques are used to forecast or explain a specific numerical outcome using a previous data set. A decision tree integrates certain decisions, but a random forest combines numerous decision trees. Consequently, it is a slow yet lengthy process. While a decision tree can quickly and efficiently process huge data sets, especially linear ones, a random forest model requires extensive training [8].

2.3) The main difference between classification and regression is that classification gets challenged when predicting with a discrete class label while regression's objective is to forecast a continuous quantity of data. An output is categorized into a certain category in a classification task, but an output in a regression problem is an actual value [9].

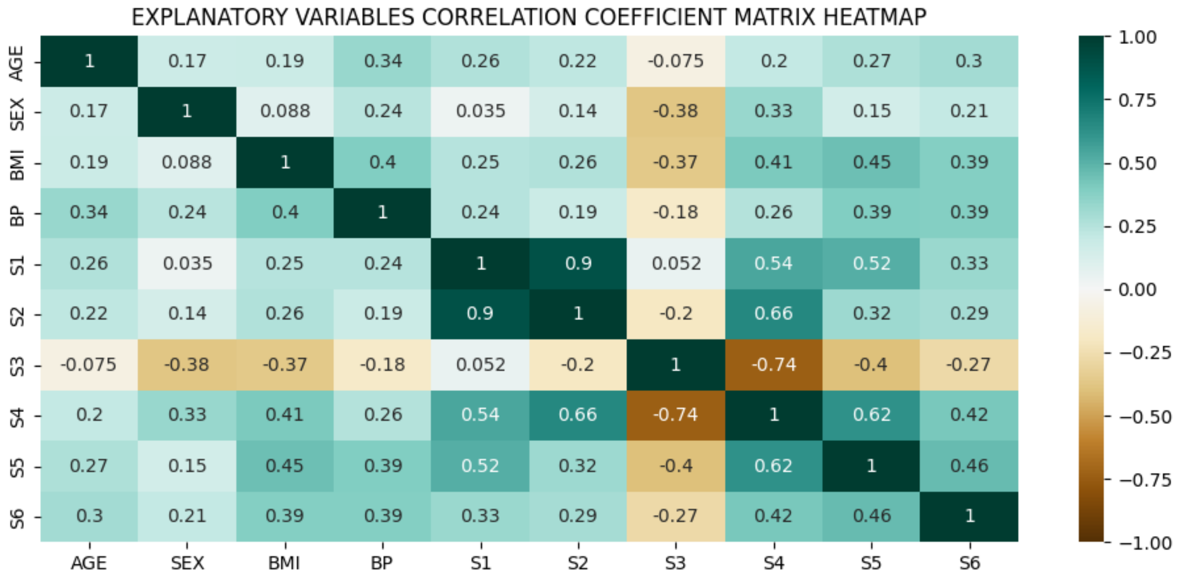
2.4) The main difference between supervised learning and unsupervised learning is that both x and y are needed to evaluate the relationship. The goal of supervised learning is to acquire a function that is pertinent to the issue. One of the applications of this is regression, where we cannot determine the dependent variable without the independent variable. On the other hand, unsupervised learning is a type of learning done when there are no labels, lots of data, and only the variables x and y. Unsupervised learning does not require the input dataset to be labeled, classified, or categorized, in contrast to supervised learning. Examples of unsupervised learning are classification, clustering, and segmentation which may all be done with x only [9].

2.5) Some examples of successful applications of machine learning are Real-time mobile personalization, recommender systems, sales data analysis, and learning management systems. Real-time mobile personalization is used to enhance the user experience of the system. Recommendation systems are used to identify pattern changes, using classification techniques, and supervised learning to make customization for each client experience. ML is also used for dynamic price generation based on different aspects like demand, location, and so on. This can be thought of clustering technique or supervised learning. In addition, Learning Management Systems use ML in decision-making tools, and it is also a classification technique or supervised learning. Other important applications of Machine learning include Natural Language processing, dynamic prices, email-spam detection, facial recognition, and bank fraud detection which use regression techniques as well [10].

QUESTION 3:

3.1)

As it was tasked to load the diabetes data and produce a correlation matrix of the explanatory variables, and then make a heat map of the matrix to show the relationships between the variables, it can be inferred that aside from the linearity between the same variable, the correlation of 0.9 which is between S1 and S2 is the strongest hence they have a strong positive relationship, the relationship between S3 and S4 is the strong negative relationship since the correlation coefficient is -0.74. Those with lower values indicate the ones with weak relationships means they have no relationship between them. Those are SEX and S1 which have a weak positive correlation coefficient of 0.035 and a weak negative would be -0.18 which is between BP and S3.

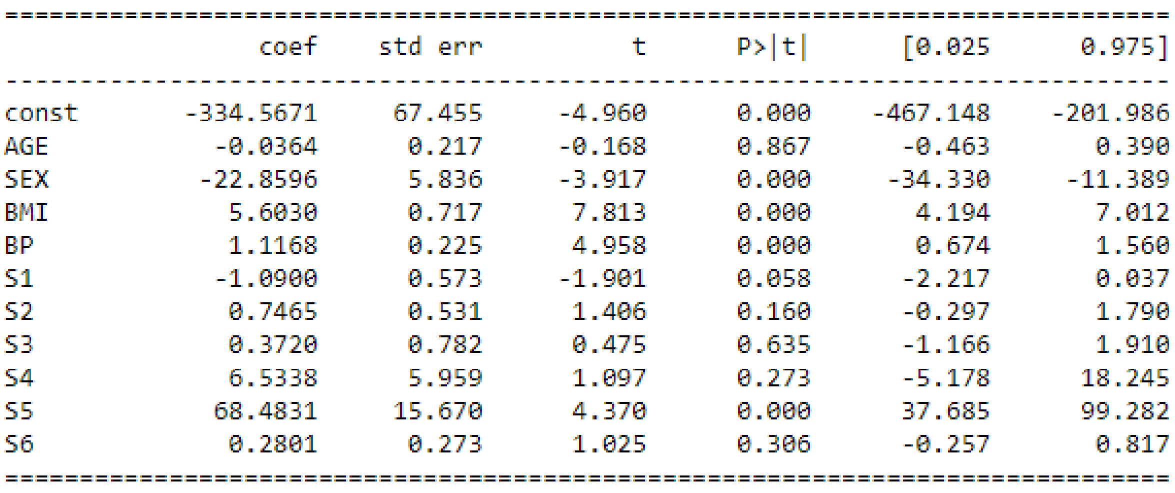


	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6
AGE	1	0.17	0.19	0.34	0.26	0.22	-0.075	0.2	0.27	0.3
SEX	0.17	1	0.088	0.24	0.035	0.14	-0.38	0.33	0.15	0.21
BMI	0.19	0.088	1	0.4	0.25	0.26	-0.37	0.41	0.45	0.39
BP	0.34	0.24	0.4	1	0.24	0.19	-0.18	0.26	0.39	0.39
S1	0.26	0.035	0.25	0.24	1	0.9	0.052	0.54	0.52	0.33
S2	0.22	0.14	0.26	0.19	0.9	1	-0.2	0.66	0.32	0.29
S3	-0.075	-0.38	-0.37	-0.18	0.052	-0.2	1	-0.74	-0.4	-0.27
S4	0.2	0.33	0.41	0.26	0.54	0.66	-0.74	1	0.62	0.42
S5	0.27	0.15	0.45	0.39	0.52	0.32	-0.4	0.62	1	0.46
S6	0.3	0.21	0.39	0.39	0.33	0.29	-0.27	0.42	0.46	1

Figure 2: Heat map for explanatory variables correlation coefficients

3.2) Collinearity in machine learning is a situation in which two or more predictor variables are correlated hence making it difficult to determine their regression. This means that correlated predictor variables in a single regression model cannot predict the value of the dependent variable on their own [11]. While collinearity is said when we have two independent variables, having a high correlation between two or more explanatory variables is called multicollinearity. This may present a challenge for parameter estimation since it raises the variance of the regression parameters, which could lead to the wrong identification of crucial predictors in a statistical model. Furthermore, collinearity might lessen the precision of the predicted coefficients, which can weaken the statistical power of the regression model. Using the p-values to determine if independent variables are statistically significant may not be accurate.

3.3)



```
==============================================================================
                 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
==============================================================================
```

Figure 3: Ten variables summary statistics

Looking at the table above, all variables are not significant because there some have a p-value that is greater than the significance value of 0.05. Variables like AGE, S1, S2, S3, S4, and S6 have greater p-value which implies that they are not significant. Furthermore, this can be due to a collinearity problem since most of the variables have high correlation coefficients are the heatmap above entails. While calculating the mean squared error it is found as **2859.6963475867506** and the adjusted R squared is **0.5065592904853231.**

3.4) Generally forward selection and backward selection are both used to select the best predictor variables to use for predicting the dependent variable outcomes. By using forward selection, variables are frequently added to an empty collection of variables until a halting condition is met. On the other hand, backward selection starts with a set of variables that is normally complete and keeps excluding variables from it until a stopping condition is met [12].

3.5) As the name implies, stepwise regression selects variables one at a time. The approach adds or removes each independent variable based on its statistical significance. Stepwise changes either make the most significant variable larger or make the least significant variable smaller. After the process, only one feasible regression model is considered. Typically, the specifics of the step-by-step procedure are under your control. You might specify whether it can add, remove, or both variables, for example. It is also possible to provide a significance threshold, which is often set at 0.05, for both including and omitting the independent variables [12].

Using the stepwise forward selection, ['BMI', 'S5', 'BP', 'S1', 'SEX', 'S2'] variables were selected and the MSE and R^2 for this new model were **2876.683251787016** and **0.5148837959256445** respectively.

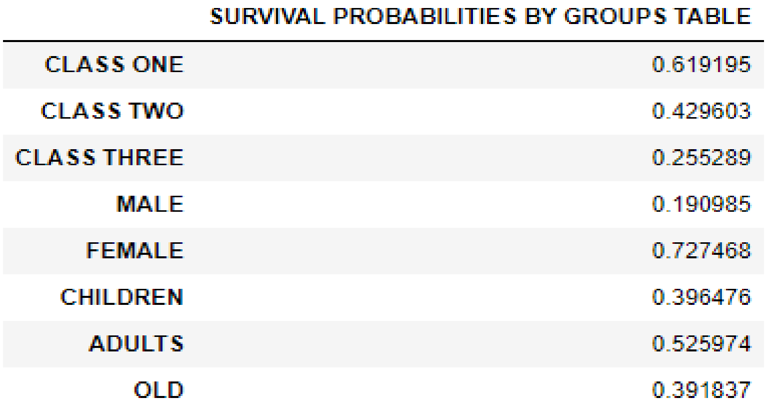
QUESTION 4:

4.1) The primary distinction between linear and logistic regression is that logistic regression produces discrete output, while linear regression produces continuous output. Once it has been fitted to recent values of the input variables, a linear regression model may be used to forecast the values of the dependent or response variables. As a result, it provides an estimate of the dependent variable if the independent factors change. For example, predicting the prices of goods depends on different factors change as weather, oil price, and so on.

Logistic regression is used with the dependent variable being discrete (often binary) and at that time, the explanatory variables may be either discrete or continuous [13]. The logistic regression function maps categorical dependent or response variables to the input variables. Unlike linear regression, logistic regression produces a probability that often ranges from 0 to 1. Instead of predicting or forecasting the outcome, logistic regression calculates the likelihood of a binary outcome. Therefore, it is used to determine the likelihood that an event will occur, such as determining if tissue is malignant or benign. In general, it puts the inputs into categories.

4.2) As calculated by taking the total number of survivors divided by the total number of passengers, the probability of survival for a passenger on the titanic is **0.3819709702062643.**

4.3) As detailed below, survival probabilities are broken down by passenger class (3 classes), gender (male and female), and age (children: 0-17, adults: 18-39, old: 40+).

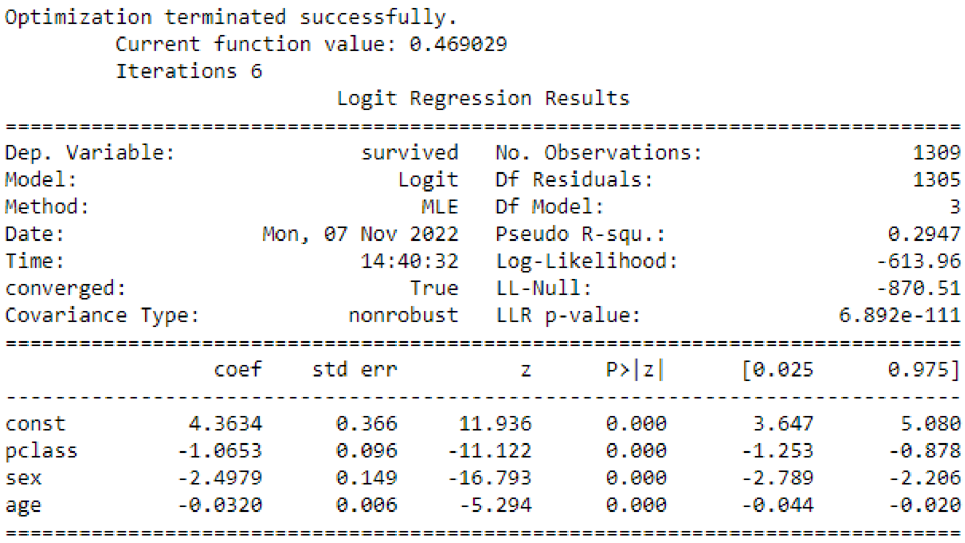


	SURVIVAL PROBABILITIES BY GROUPS TABLE
CLASS ONE	0.619195
CLASS TWO	0.429603
CLASS THREE	0.255289
MALE	0.190985
FEMALE	0.727468
CHILDREN	0.396476
ADULTS	0.525974
OLD	0.391837

Figure 4: Survival probabilities by categories

We can infer from the table that females are more likely to survive as their surviving probability is the highest (**0.727468**) and the probability of surviving for males was the least with **0.190985.**

4.4) The logistic regression was required, with passenger class, sex, and age as the independent variables and survived as the dependent variable. This was done using the LogisticRegression() function from statsmodels.api for building and fitting the model and the build model is used for predictions.



```
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, 07 Nov 2022   Pseudo R-squ.:                  0.2947
Time:                        14:40:32   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
==============================================================================
```

Figure 5: Logit regression summary results

Additionally, getting the parameter estimates required using statsmodels.api's Logit function to get the estimates and the p-values as represented in the table above. This shows that the parameter estimate for p class is -1.0653, -2.4979 for sex, and -0.0320 for age. In addition, these parameters are all significant compared to the level of significance of 0.05 since their p-values are zeros.

4.5) Now it's time to determine the classification accuracy using the number of correct classifications divided by the total number of classifications based on the confusion matrix. This is done by first determining the confusion matrix done by sklearn.metrics's confusion_matrix() function which returns a binary classification matrix where the count of true negatives is 0, 0, false negatives is 1, 0, true positives 1,1 and false positives is 0, 1. To calculate the accuracy, the sum of true negatives and true positives is divided by the total outcomes. Doing that gave the accuracy of this logistic regression model as **0.7853323147440795.** This shows that the model provides frequent right outcomes at **78 percent.**

REFERENCES

[1] Wikipedia, "Rule-based machine learning," 14 July 2021. [Online]. Available: https://en.wikipedia.org/wiki/Rule-based_machine_learning. [Accessed 5 November 2022].

[2] HackerNoon, "7 Effective Ways to Deal With a Small Dataset," [Online]. Available: https://hackernoon.com/7-effective-ways-to-deal-with-a-small-dataset-2gyl407s. [Accessed 5 November 2022].

[3] O. i. M. L. W. I. I. a. H. t. P. It, "EliteDataScience," [Online]. Available: https://elitedatascience.com/overfitting-in-machine-learning#how-to-prevent. [Accessed 7 November 2022].

[4] net-informations, "Squared: Coefficient of Determination," [Online]. Available: http://net-informations.com/ds/psa/r-squared.htm#:~:text=The%20R%C2%B2%20is%20calculated%20by,then%20subtract%20it%20from%201..

[5] towardsdatascience, "8 Metrics to Measure Classification Performance," [Online]. Available: https://towardsdatascience.com/8-metrics-to-measure-classification-performance-984d9d7fd7aa.

[6] MobiDev, "TOP 9 Machine Learning Technology Trends To Impact Business in 2022," [Online]. Available: https://mobidev.biz/blog/future-machine-learning-trends-impact-business. [Accessed 7 November 2022].

[7] T. AI, "The Evolution of Machine Learning in Business," [Online]. Available: https://www.turintech.ai/the-evolution-of-machine-learning-in-business/. [Accessed 7 November 2022].

[8] u. blog, "Random Forest Vs Decision Tree: Difference Between Random Forest and Decision Tree," [Online]. Available: https://www.upgrad.com/blog/random-forest-vs-decision-tree/#:~:text=A%20decision%20tree%20combines%20some,forest%20model%20needs%20rigorous%20training. [Accessed 7 November 2022].

[9] Medium, "Machine Learning Approaches and Its Applications," [Online]. Available: https://medium.datadriveninvestor.com/machine-learning-approaches-and-its-applications-7bfbe782f4a8. [Accessed 7 November 2022].

[10] Javatpoint, "Applications of Machine learning," [Online]. Available: https://www.javatpoint.com/applications-of-machine-learning. [Accessed 7 November 2022].

[11] britannica, "collinearity," [Online]. Available: https://www.britannica.com/topic/collinearity-statistics. [Accessed 7 November 2022].

[12] datacadamia, "Statistics - Forward and Backward Stepwis," [Online]. Available: https://datacadamia.com/data_mining/stepwise_regression. [Accessed 7 November 2022].

[13] GitHub, "bio304-class/bio304-book: Bookdown based course notes for Bio 304 at Duke University, taught by Paul Magwene," [Online]. Available: https://github.com/bio304-class/bio304-book. [Accessed 7 November 2022].