[Year]

Assignment 5

DATA INFERENCE AND APPLIED MACHINE LEARNING (COURSE 18-785)

YASSER CORZO

CARNEGIE MELLON UNIVERSITY | 5000 Forbes Ave, Pittsburgh, PA 15213

LIBRARIES USED

- Matplotlib
- Numpy
- Seaborn
- Statmodel
- Stepwise_regression
- Tabulate
- Sklearn

REPORT

Q1 STATISTICAL LEARNING

1. The four steps to implementing a rule-based approach to decision-making is [9]:
   1) Problem Identification & Rule Creation
      Identifying the problem before implementing the rule-based system is vital to the effectiveness of the system. Without clearly identifying the problem, rules cannot be created and applied to solving a problem. Consulting with a team of experts with in-depth knowledge of the problem is important as they are the ones who typically create the rules for a rule-based system.
   2) Encoding Rules into System & Storage
      After creating the rules, they need to be encoded into the system in a format a computer can understand, such as "if-then" statements with boolean operations. Once the rules have been created and encoded into the system, they need to be stored into a database or knowledge base. In addition to storing the rules, the knowledge base may store other types of information, such as definitions, explanations of the terms and concepts used in the rules, etc, that can help to improve the understanding of the system.
   3) Data Collection
      Gathering the necessary data and information to the to the problem at hand is needed to help the system make decisions. This information will need to be matched to if-part of rules (a.k.a. pattern matching) to activate a decision. Data can include current environmental conditions, measurements, observations, or any other information needed for the decision process.
   4) Rule Evaluating & Applying
      Using the input data, the system can evaluate the rules in its knowledge base. This is done by implementing algorithms that search for and apply the rules (e.g. Forward & Backward Selection).
   5) Monitor & Update System through testing.
      Validating and testing the rules in a rule-based system is important to ensure accuracy and effectiveness. Incorporating the feedback from testing to monitor and update the system can aid in improving accuracy and effectiveness. As new information becomes available and conditions change, rules may be updated or revised. Implementing a process for system maintenance and updates can help ensure the system remains accurate and effective over time.

   An example of a rule-based approach to decision-making is implementing a system for diagnosing a patient [9]. The problem here is disease diagnoses of a patient, given a list of the patient's symptoms. Consulting with doctors with various specialties can help with creating rules that determine a diagnosis for a patient. Examples of rules for determining a diagnosis can include the following [9]:

   Rule #1: If the patient has a fever and a rash, then they may have measles.
   Rule #2: If the patient has a fever and joint pain, then they may have influenza.
   Rule #3: If the patient has a fever and a cough, then they may have pneumonia.

   Once the rules have been created by the panel of experts, they can be encoded and stored in a database. Next would come the data collection part, which in this example could include a set of symptoms that a patient is experiencing, such as fever, cough, and rash. The system then would search through its database and match the input data with any rules that are relevant to the patient’s symptoms. In the case of this patient with a fever, cough, and rash, the system might apply the rule: “If the patient has a fever and a rash, then they may have measles." If the system determines that this rule is applicable, it may diagnose the patient with measles.

Domain knowledge is required to establish a rule as rules are based on knowledge on the subject the rule is being applied to. For example, when creating a rule for diagnosing whether a patient has measles or not, knowledge regarding the symptoms of measles is needed for creating the rules that determine whether a patient has measles or not. Without this domain knowledge how can the system determine whether a patient has measles.

2. Over-fitting is when a model models the training data too well, i.e., instead of generalizing over the general distribution of the data, the model learns the expected output for every data point, thus fitting noise. This is a problem in statistical learning because without the ability to generalize data, a model is unable to make accurate predictions on new, unseen data, undermining the overall goal of statistical learning of representing, estimating, and evaluating data.

   If I have a small dataset containing ten data points, I will prefer a simple model with one parameter over a complex model with ten parameters because a model with too many parameters will not distinguish between the general pattern in the dataset that we wish to extract and fluctuations due to noise. Thus, the aim should always be to choose the simplest model that is compatible with the observations.

3. The two commonly used approaches to avoid over-fitting are regularization and out of sample testing. Regularization aims to combat overfitting by reducing the weights or magnitude of coefficients of features. Regression techniques, such as Ridge Regression, use L2 regularization to shrink the parameters and reduces model complexity, thus reducing over-fitting, by reducing the coefficients. [8] Out of sample testing, such as cross validation, avoids over-fitting by separating the estimation and evaluation procedures where the data set is split into two: learning data set for training the model by estimating its parameter values and testing data set for evaluating the model and calculating performance metrics. [8]

4. Two examples of metrics used to evaluate the performance of a model is mean absolute error (MAE) and coefficient of determination.
   Formula of MAE: $\frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i - y_i |$
   Formula of coefficient of determination: $R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$
   Where $SS_{tot} = \sum_i (y_i - \bar{y})^2$, $SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2$, $SS_{res} = \sum_i (y_i - \hat{y}_i)^2$

   An application where coefficient of determination would be used is in models that predict house prices. If the goal is to build a regression model that predicts house prices based on various features such as number of bedrooms, etc. After collecting data and building the model, we can evaluate the model by calculating the coefficient of determination to assess how well the model fits. A value close to 1 would indicate the model is capturing the relationships between the features and house prices, while a value close to 0 would indicate the model isn’t capturing the relationship between the features and house prices, indicating a poor fit. An application where MAE would be used is in weather forecasting where a model can be created to forecast daily temperature. After collecting historical weather data, a regression model can be built that uses features such as temperature, data, and location to forecast temperature. The model can then be evaluated with MAE by taking the absolute difference between the predicted temperatures and the actual temperatures from a test data set and dividing by the number of data points in the data set. A lower MAE would indicate the predicted temperatures are close to the actual temperatures, thus showing the model is accurate in forecasting temperature, while a higher MAE would indicate the model is less accurate.

5. Benchmarks are useful in machine learning in establishing levels of performance. Two examples of forecasts are persistence and unconditional average forecast.

Q2 MACHINE LEARNING

1. Machine Learning refers to algorithms whose performance improves with exposure to data. In 1950, the development of the Alan Turing Test led to a machine that could learn and communicate like a human. In 1952 came ELIZA, a checker-playing program developed by Arthur Samuel from IBM, which had significant skill to challenge world champions. In 1957, Frank Rosenblatt invented the perceptron, paving the way for neural networks. In 1990, the computer science and statistics field combined to provide a data-driven approach to machine learning. Big Data (exponential growth in the volume, velocity and variety of data available for analysis and research) made its mark in 2010. In 2014, Open Data and IoT, the infrastructure and standards for providing open access to data via APIs, was developed. Machine learning is so popular because it allows engineers and data scientists to use algorithms to identify nonlinear relationships in large datasets, something very complex for a human to do by hand.

2. Three examples of supervised machine learning techniques are decision trees, logistic regression, and SVM. Three examples of unsupervised machine learning techniques are K-means, Hidden Markov Models, and PCA.

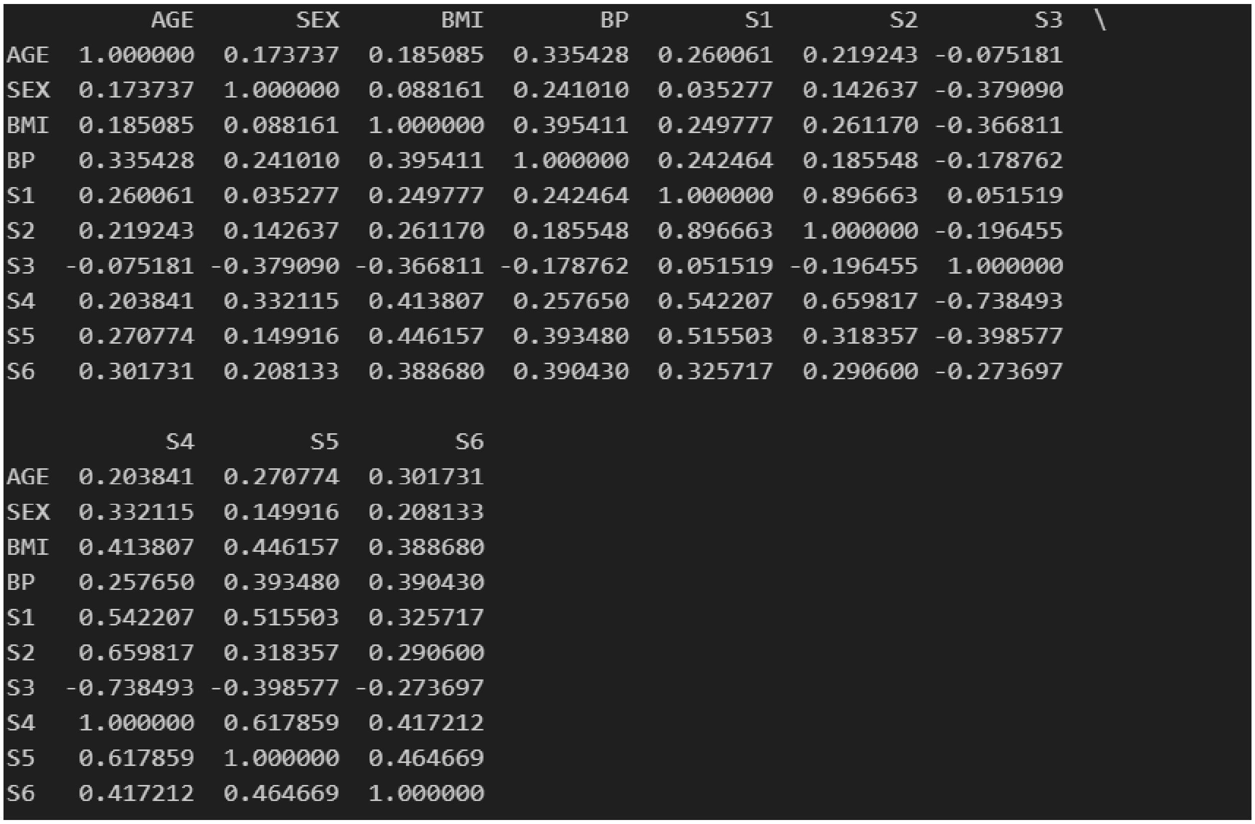
3. The difference between regression and classification is classification outputs categorial values such as True or False, Male or female, etc. Regression, on the other hand, outputs continuous values such as price, income, age, etc.

4. The difference between supervised and unsupervised learning is that in supervised learning, labeled datasets are used to train algorithms for classification. The labeled dataset also has outputs tagged to the corresponding inputs, so the machine learns how to predict unseen data. [6] In unsupervised learning, training datasets do not contain any labels. Instead of learning how to predict outcomes by the inputs and outputs in the labeled training set, algorithms in unsupervised learning need to find previously hidden patterns in the unlabeled dataset. [6]

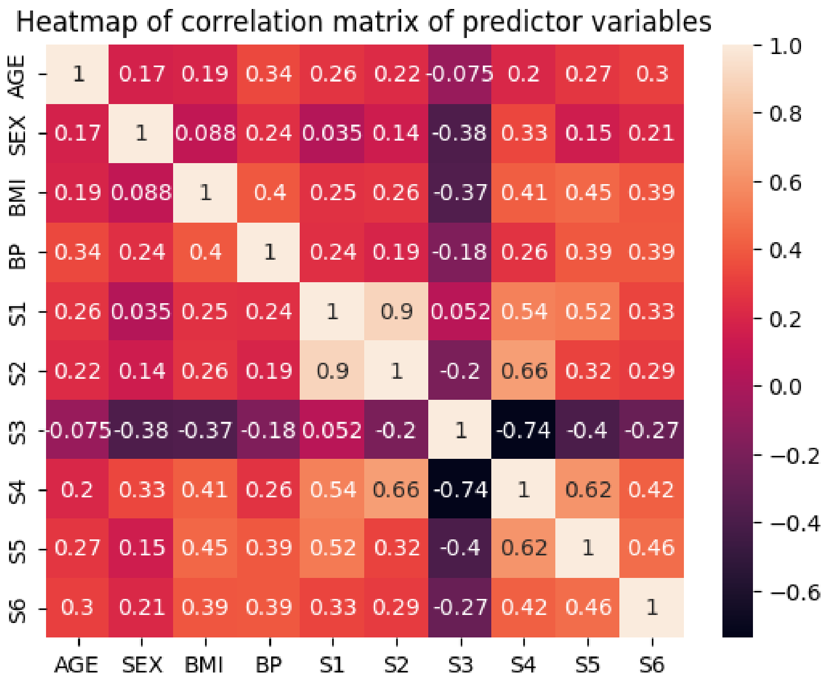
5. Examples of successful applications in machine learning include image classification, weather forecasting, and anomaly detection. [6] The technique appropriate for image classification would be CNNs (convolutional neural networks) and the type of learning involved here is supervised. The technique appropriate for weather forecasting would be logistic regression and the type of learning involved is supervised. The technique appropriate for anomaly detection would be K-means and the type of learning involved is unsupervised. [6]

Q3 DIABETES DATA

1.



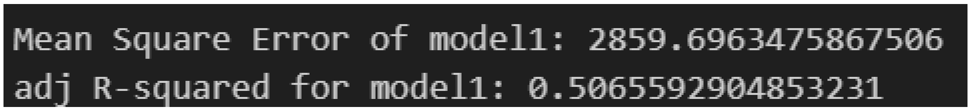
```
            AGE       SEX       BMI        BP        S1        S2        S3  \
AGE  1.000000  0.173737  0.185085  0.335428  0.260061  0.219243 -0.075181
SEX  0.173737  1.000000  0.088161  0.241010  0.035277  0.142637 -0.379090
BMI  0.185085  0.088161  1.000000  0.395411  0.249777  0.261170 -0.366811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

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



The relationship between the variables and themselves will always have a correlation of 1 (represented by the diagonal entries in the above matrix). The variables deemed to have a stronger correlation are those variables with a correlation of at least 0.5 which include S4 & S2, S5 & S4, S2 & S1, S4 & S1, and S5 & S1. Most of the variables are positively correlated between one another, except for S3 which is negatively correlated with every other explanatory variable (excluding itself).

The steps involved in producing the correlation matrix were reading the excel sheet through pandas, retrieving the columns of the explanatory variables (Columns 1-10) and using pandas corr() method [5] for calculating the correlation matrix. Creating the heatmap involved the use of Seaborn's heatmap() method. [4]

2. Collinearity is when there is a strong correlation between two variables. The effect that collinearity has on the estimated coefficient value of predictor variables is an increase in variance of the coefficient estimates, making the estimates sensitive to minor changes in the model. [3]

3. 

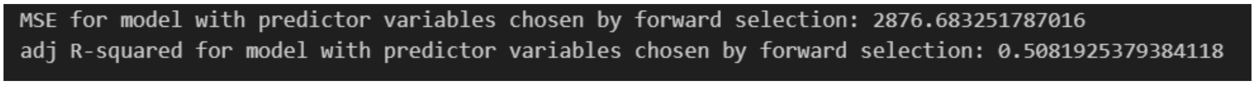
```
Mean Square Error of model1: 2859.6963475867506
adj R-squared for model1: 0.5065592904853231
```

Above are the statistics asked by the question regarding model 1. Based on the heatmap of the correlation matrix, I would say that not all variables are significant. For example, variable S3 is negatively correlated with every other variable (excluding itself). This could be a problem of collinearity as there are many variables where a strong correlation exists, which could lead to an increase in variance and a decrease in accuracy in the estimated coefficient, hence an increase in noise sensitivity.

For calculating the r-squared and MSE of the models, statsmodel.OLS() function was used to fit the data between the explanatory and prediction variables.

4. The difference between forward and backward selection is in forward selection, one starts with no variable and includes one variable at a time, testing for statistical significance [2]. On the other hand, backward selection involves one starting with all the variables, rejecting one variable at a time, testing for statistical significance. [2]

5. The stepwise approach works by, through various iterations, at each iteration variables which are made obsolete by new additions are removed. Thresholds are set for adding a variable and removing a variable (based on p-value). The algorithm stops when no more variables are added or when a variable is removed immediately after being added. The variables chosen during forward selection were BMI, S5, BP, S1, sex, and S2.



```
MSE for model with predictor variables chosen by forward selection: 2876.683251787016
adj R-squared for model with predictor variables chosen by forward selection: 0.5081925379384118
```

Above is the MSE and adjusted r-squared value for the model with variables chosen by forward selection.

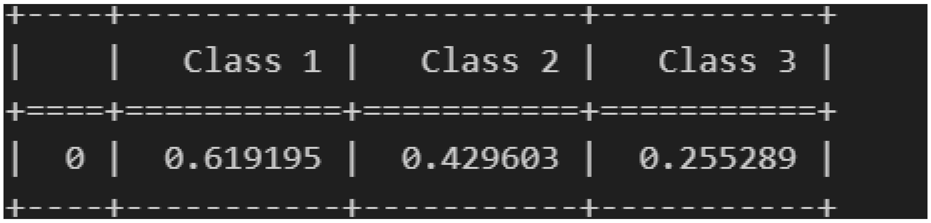
Using the forward_regression() function from stepwise_regression helped chose which variables would be chosen by forward selection.

Q4 ANALYZING THE TITANIC DATA SET

1. The difference between logistic and linear regression is linear regression is used to predict a dependent variable based on independent variables, the best fit is a straight line, and the output is a continuous value. On the other hand, logistic regression is used to classify a dependent variable based on independent variables, the best fit is given by a curve, and the output is binary. [7]

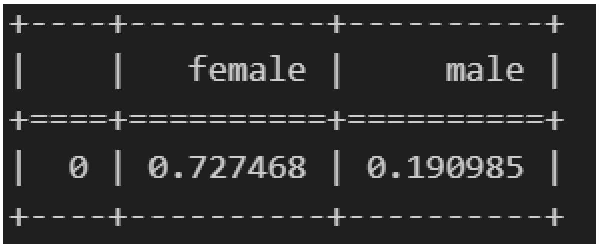
2. Probability of survival for a passenger on the titanic: 0.3819709702062643. This value was calculated by calculating the total number of passengers aboard the titanic and calculating the number of passengers that survived.

3.



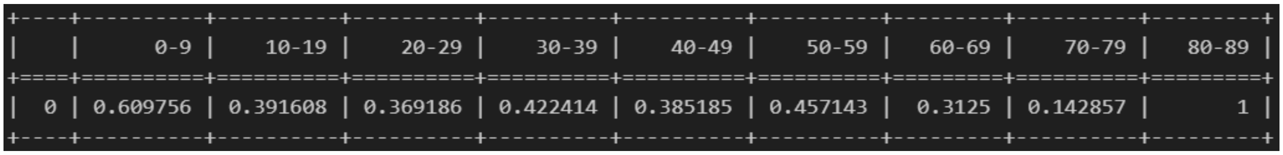
```
+----+-----------+-----------+-----------+
|    |   Class 1 |   Class 2 |   Class 3 |
+====+===========+===========+===========+
|  0 |  0.619195 |  0.429603 |  0.255289 |
+----+-----------+-----------+-----------+
```

Survival probability broken down by passenger class



```
+----+----------+----------+
|    |   female |     male |
+====+==========+==========+
|  0 | 0.727468 | 0.190985 |
+----+----------+----------+
```

Survival probability broken down by gender



```
+----+----------+----------+----------+----------+----------+----------+---------+----------+---------+
|    |      0-9 |    10-19 |    20-29 |    30-39 |    40-49 |    50-59 |   60-69 |    70-79 |   80-89 |
+====+==========+==========+==========+==========+==========+==========+=========+==========+=========+
|  0 | 0.609756 | 0.391608 | 0.369186 | 0.422414 | 0.385185 | 0.457143 |  0.3125 | 0.142857 |       1 |
+----+----------+----------+----------+----------+----------+----------+---------+----------+---------+
```

Survival probability broken down by age group [start age, end age]

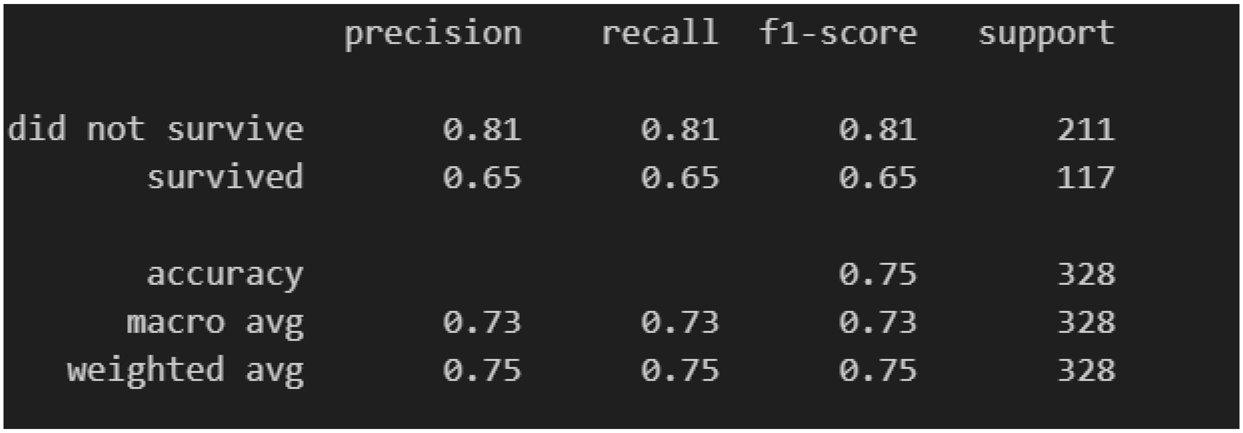
These tables were created using pandas. The probabilities were calculated by counting how many passengers were in each category and within each category counting the number of passengers that survived.

4. The parameter estimates are -0.88300359, -2.53261944, and -0.01218418. Based on the p-values calculated for class, gender, and age (3.997459e-19, 4.336158e-51, and 2.972144e-04 respectively), since all p-values are less than alpha, they are all statistically insignificant.

   Parameter estimates were retrieved using sklearn's LogisticRegression.fit().coef_ attribute and the p-values were calculated using statsmodel's pvalue attribute.

5.

	P	N
P	170	41
N	41	76



```
                precision    recall  f1-score   support

did not survive       0.81      0.81      0.81       211
       survived       0.65      0.65      0.65       117

       accuracy                           0.75       328
      macro avg       0.73      0.73      0.73       328
   weighted avg       0.75      0.75      0.75       328
```

Classification accuracy is at 0.75.

Confusion matrix was created using sklearn’s confusion_matrix() function [1] and the classification accuracy was calculated using sklearn’s classification_report() function [1].

REFERENCES

[1] A. Navlani, “Python logistic regression tutorial with Sklearn &amp; Scikit,” DataCamp, https://www.datacamp.com/tutorial/understanding-logistic-regression-python (accessed Nov. 5, 2023).

[2] G. Choueiry, “Quantifying health,” QUANTIFYING HEALTH, https://quantifyinghealth.com/stepwise-selection/ (accessed Nov. 5, 2023).

[3] M. B. Editor, “What are the effects of multicollinearity and when can I ignore them?,” Minitab Blog, https://blog.minitab.com/en/adventures-in-statistics-2/what-are-the-effects-of-multicollinearity-and-when-can-i-ignore-them#:~:text=Moderate%20multicollinearity%20may%20not%20be,unstable%20and%20difficult%20to%20interpret. (accessed Nov. 5, 2023).

[4] “How to create a Seaborn Correlation heatmap in python?,” Online Tutorials, Courses, and eBooks Library, https://www.tutorialspoint.com/how-to-create-a-seaborn-correlation-heatmap-in-python (accessed Nov. 5, 2023).

[5] “Create a correlation matrix using Python,” GeeksforGeeks, https://www.geeksforgeeks.org/create-a-correlation-matrix-using-python/ (accessed Nov. 5, 2023).

[6] “Supervised vs. unsupervised learning [differences &amp; examples],” Supervised vs. Unsupervised Learning [Differences &amp; Examples], https://www.v7labs.com/blog/supervised-vs-unsupervised-learning#:~:text=The%20most%20commonly%20used%20Supervised,hierarchical%20clustering%2C%20and%20apriori%20algorithm. (accessed Nov. 5, 2023).

[7] Simplilearn, “Understanding the difference between linear vs logistic regression,” Simplilearn.com, https://www.simplilearn.com/tutorials/machine-learning-tutorial/linear-regression-vs-logistic-regression (accessed Nov. 5, 2023).

[8] Shubham. jain Jain, “Lasso &amp; Ridge Regression: A comprehensive guide in python &amp; R (updated 2023),” Analytics Vidhya, https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/ (accessed Nov. 5, 2023).

[9] “Rule-based systems explained: Understanding their functioning and advantages,” Mastering Rule-Based Systems: Implementation, Benefits, and Best Practices, https://www.xaqt.com/blog/mastering-rule-based-systems/ (accessed Nov. 5, 2023).