|  |
| --- |
| **National University of Computer and Emerging Sciences** |
| Lab Manual 11  “**Classification and Regression**” |
|  |
| Data Warehousing and Data Mining |
|  |
|  |
|  |

|  |  |
| --- | --- |
| Course Instructor | Sir Ishaq Raza |
| Lab Instructor(s) | Farhan Azhar Nouman Ali |
| Section | CS |
| Semester | Fall 2019 |

**Classification**

1. Read the titanic dataset and print the summary of the data
2. Make a count plot of the ‘Survived’ column
3. Make a count plot of the ‘Survived’ column using ‘Sex’ column as secondary
4. Make a count plot of the ‘Survived’ column using ‘Pclass’ column as secondary
5. Visualize the outliers in ‘Age’ column for every ‘Pclass’
6. Fill empty values for Age column using a suitable method
7. Drop ‘Cabin’, ‘PassengerId’ , ‘Name’ and ‘Ticket’ column
8. Covert every categorical column/alphabetic to numeric representation
9. Draw a correlation matrix for the data
10. Perform test train split using sklearn. Note the ‘Survived’ column is your target/y variable. The rest are to be used as input/x variables.
11. Use Logistic Regression model from sklearn and perform training and testing.
12. Report the accuracy score and confusion matric for 11.
13. Use the DecisionTreeClassifier from sklearn and perform training and testing
14. Report the accuracy score and confusion matric for 13

**Regression**

1. Read the housing dataset and print the summary of the data
2. Make a distribution plot of the ‘Price’ column
3. Draw a correlation matrix for the data
4. Select the ‘Avg. Area Income’ and ‘Avg. Area House Age’ columns. Perform normalization on these columns.
5. Perform test train split using sklearn. Note the ‘Price’ column is your target/y variable. The columns from 4. are to be used as input/x variables.
6. Use Linear Regression model from sklearn and perform training and testing. Visualize your prediction results using a scatter plot.
7. Report mean squared error and r square error of your predictions.