Course	DS 203: Programming for Data Science
Activity	Type: Exercise Title: E06 – Logistic Regression and Classification Metrics using Python
Background	• Concepts of Logistic Regression and it's metrics have already been covered in the class (see : **DS203-2023-08-25-LogisticRegression.pdf**) • Logistic Regression is one of the methods used to create **Classification** ML models • It is very important to understand the various metrics associated with classification and to learn to interpret them correctly
Expected outcomes of this exercise	• Data visualization using Python MatplotLib • Train a Logistic Regression model using Python • Use the model to classify (i.e. predict) the observations • Learn to calculate and interpret the Classification Metrics: true positives, false positives, true negatives, false negatives, TPR, FPR, TNR, FNR, F1-Score. • Learn to create and interpret the **Confusion Matrix** at multiple probability thresholds • Learn to create and interpret the **ROC Curve** and **AUC**
Tools	• Python Notebook using either Jupyter or VSC
Effort estimate	• **5 hours**
Submission type	• **Mandatory submission**
Due date and time	• **September 5, 2023, 23:55 Hrs**
Submission instruction	• Submit to the appropriate Moodle submission point • Your final submission should be in the form of a single PDF file. • Your solutions and answers should include explanations / graphs / charts / Tables that are necessary to fully explain the solution(s). • **Complete this exercise and submit well within time. No extensions will be granted, no email submissions will be accepted.**
Marks for the exercise	• Credit will be given for **complete and timely** submission to Moodle • The exercise itself will not carry marks. • **Your understanding and skill – expected to be gained by completing this exercise – will be gauged in a quiz, test, or viva that will be conducted subsequently.**
References	• Python Notebooks uploaded to course page on Moodle. • Lecture notes • Help documentation for Python, Numpy, Matplotlib, Scikit-Learn. • Articles, blogs and other sources
NOTE	• **The PDF should have your roll number as the filename. You may additionally include your name. No credit will be given if this is not followed!** • **Please ensure that your submission does not include long dumps of vector / matrix values. Remove your 'debug dumps' before submitting!**

Prerequisite:

- Review the uploaded Python Notebook **Logistic-Regression-Basics.ipynb** and understand the steps
- Review the documentation for the **LogisticRegression** function defined in the **sklearn** package

The data set:

- This exercise refers to the data set in **Logistic-Regression-data-2-class-v0.csv**
- It contains labelled data related to 2 classes, '0' and '1'

Create a Python Notebook to program and complete the following tasks (estimated effort: 5 hours):

1. Visualize the data by creating a single scatter plot showing the two classes in different colours.
2. What are your observations about the data set and the two classes?
3. Train a Logistic Regression model using all the provided data (Note: normally only about 80% of the data is randomly selected and used for training the model; the remaining 20% is used for validating the model. We will do that later …)
4. Use the trained model to predict the class of every observation in the data set.
5. Create a scatter plot showing the two data sets, as in step number 1. But in this plot, show the wrongly classified observations in a third colour.
6. What are your observations about the quality of the model, i.e. the classifier?
7. Create the **Confusion Matrix** based on the predicted results and print it out. (***This confusion matrix corresponds to the default probability threshold value of 0.5***)
8. Calculate and print the following, and comment on their values: Precision, Recall, f1 value, TPR and FPR
9. Instead of using the 'predict' function, use the 'predict_proba' function to get the classification probabilities associated with the observations.
10. Classify the observations by varying the threshold probability from 0 to 1 in increments of 0.1. For every classification:
    a) Print out the confusion matrix
    b) Calculate the TPR and FPR values
11. Creae the ROC by plotting TPR v/s FPR.
12. What are your observations? What is your interpretation?
13. Can you calculate the AUC? Print and comment on it's value.

Convert the Notebook into PDF (your program segments and all the generated / added outputs) and upload it as your submission for this assignment. **Please ensure that your submission does not include long dumps of vector / matrix values. Remove your 'debug dumps' before submitting!**

*****