Statistical Machine Learning **Assignment 2 (Max Marks:120)**

Instructions: You are allowed to discuss but the final answer should be your own. Any instance of cheating will be considered as academic dishonesty and penalty will be applied.

1. Restrict to using only Python for coding assignments.
2. You are free to use math libraries like *Numpy, Pandas*; and use *Matplotlib, Seaborn* library for plotting.
3. Add all the analysis on the question in the written format, anything not in the report is not marked.
4. Use of inbuilt function for any evaluation metric is not allowed. Each of the metrics needs to be implemented from scratch.

Link to book scans: https://drive.google.com/open?id=15sNCsQrzZo2FGgtwZlEtjg5Dc9fXcotZ

Theory: **20 Marks**

Theory Questions from Duda Hart Second Edition. They need to be submitted in the dropbox placed:

1. Section 2.3 Question 3
2. Section 2.3 Question 9
3. Section 2.4 Question 14(a-c)
4. Section 2.8 Question 36 (a-b)

Programming: **100 Marks**

A. From the book: computer Exercise

1. Section 2.5, question 2
2. Section 2.8, question 7
3. Section 2.8, question 8

B.

Dataset Details:-

Training Set - train.txt(csv)

Test Set - test_all.txt(csv)

Test Set with Missing Features - test_missing.txt(csv)

Risk Matrices - risk.txt(json)

Given the dataset consider 2 cases:

a) Data after Class-Wise Decorrelation, ie. perform decorrelation separately on the two classes on the training set. Use the same decorrelation matrix on the test set. This gives the new training and testing data.

b) Use the train and the test set as given.

For each of (a) and (b) do the following:

1. Visualize the datasets.
2. Train a Bayes classifier on the training set. Plot the decision boundary for both train and test datasets.

3. Incorporate risk matrices given in the file 'risk.txt' to calculate and plot the decision boundaries on both train and test sets.
4. Train classifier on the train set and use test set as given in 'test_missing.txt', where feature values for 25% data points are missing. Handle the missing features using Marginalisation as taught in the class. Incorporate all the risk matrices and compare results to those in part 3.

Note:

- Make sure you plot Decision boundary, ROC for each part and subpart. Also, report Accuracy and Confusion Matrix.
- While plotting the decision boundary make sure the actual class label can be inferred from the graph.
- Report all the necessary analysis and inferences for each part.

Submission format: Please submit a report for all your analysis and observations only and only in the PDF format. Other formats will not be evaluated. All the graphs should have labels (on the axis), legends, title. You should also try to combine the graphs and plots for comparison and better representation. All the python code files need to be submitted with the following naming format: "<Roll_Number>_file.py" in a zipped folder with source folder named as "A2_<Roll_Number>.zip"

The assessment will be done on basis of the following components:
1. Working codes **(60)**
2. Analysis, inferences and clarity of results (drawing comparisons across different parts) & clarity of the report **(40)**
