**Data Analytics (CS40003/ CS61061)**

**Project statements 06.11.2019**

**Instructions:**

There are altogether 15 project descriptions stated in the following. You have to implement one project which has been assigned to you (see the course web page after 06.11.2019). You should not work on a project which has not allocated to you. For the submission of the project and other “to dos” you are advised to read the following carefully.

1. A project will be evaluated on 100 marks and it has 40% weightage toward the final marks.
2. **Evaluation plan:** 
   1. Understanding the theory to solve the project problem (20%): You should write a brief account of the theory which you should use to solve your project problem.
   2. Implementation of the project (60%): You are encouraged to write code of your own. In that case for the correct coding you will get full marks. In case, you use some tool/ library (for example, scikit in Python or R, etc.) you will get not more than 30% for that.
   3. Experimental results (20%): For different projects, the experimental results should be different. You should decide, for your project, what should be the experiment and result you would like to report. You can take help from us, if you have doubt in this regard. For each project, we have provided a data set to be used. You should use the recommended data set. In addition, you can use other data set, if you think a better reporting.
   4. Report submission (mandatory): There is no mark allotted for this, but a nice report should attract special rewarding toward the final score in the project.
3. Prepare a report which should include 2(a)-(c) as mentioned above. Please add a front page, including the team details, title of the project, and date of submission. All team members put their signatures on the front page.
4. Plagiarism, if found should be taken seriously for all the plagiarised submissions. At the extreme end, the teams(s) may be awarded “F” grade.
5. Submit hardcopy to Arpita (9775605784) within the deadline and between 10:00-12:00 or 15:00-18:00 in any working days. While submit your report, sign the “register” maintained by Arpita. Important: *No soft copy/ email submission (of the report) will be allowed under any circumstance.* However, ***softcopy of the code is MUST***. You can mail the softcopy (a zipped file with name as the “ProjectXX”, where XX is the project ID to [arpitachaudhuri.it@gmail.com](mailto:arpitachaudhuri.it@gmail.com) with a copy at [dsamanta@iitkgp.ac.in](mailto:dsamanta@iitkgp.ac.in) .
6. **Last date of submission is: 06.12.2019, 18:00 hours (hard deadline).**

**Topic 1 : Hypothesis testing to infer a population mean**

**Reference: MOVIE data**

1. Calculate population mean from all the movies up to 2015 on imdb\_score.
2. Collect a sample of all the movies in the year 2016.
3. Test the hypothesis that “Popularity of films increases”.

To test the hypothesis consider following:

1. Population standard deviation is known.
2. Population standard deviation is unknown

**Topic 2: Pearson’s Product-Moment Correlation analysis for paired data**

**Reference: NUTRITION data**

1. Decide whether “rating” is correlated with “sugar content” in the product. Calculate the degree of correlation.
2. If correlation exists, then consult the statistical table to report the significance test. Make any reasonable assumption, if you need in your experiment.
3. Calculate the “Coefficient of determination” and comment on your result of it.

**Topic 3 : Regression analysis for establishing a relation between response and regressor variables**

**Reference: SALARY data**

1. Database contains salary information of different employees in different organisations. It is required to test whether Overtime Pay, Other Pay and benefits altogether increases with Basic Pay for the year 2014.
2. Report your relation analysis with the following approaches
   1. Simple linear regression analysis.
   2. Simple non-linear regression analysis with degree 2.
   3. Simple non-linear regression analysis with degree 3.
3. Calculate R2 values in each of the above-mentioned case and finally conclude your results precisely.

**Topic 4: Spearman’s correlation analysis for paired data**

**Reference: SNACKS data**

1. Find the Spearman correlation matrix of all the ordinal attributes.
2. Determine the coefficient of determination.
3. Interpret the result from the two tables.
4. In each case, perform the significance test with 95% confidence level.

**Topic 5: Chi-square analysis to examine the correlation between two attributes**

**Reference: GAMES data**

1. It is test the hypothesis “Action video game is highly rated among teens (marked as T)”. Draw the relevant contingency table that is required for the aforesaid hypothesis testing.
2. Test the hypothesis with 95% confidence level.
3. Decide a confidence level, if any so that the hypothesis result will be just reversed.

**Topic 6: Auto-regression analysis with time-series data for future event prediction**

**Reference: STOCK data for the year 2016-2017**

1. Let’s consider the case of *p*-th order auto-regression analysis. Obtain the Covariance matrix suitable for *p*-th order auto-regression correlation analysis. *p* can be any value.
2. For the given data from stock exchange predict the stock value in the month 1/10/2017.
3. Report your prediction with different values of *p*.

Hint: To solve the above problem, you are free to choose any method reported in the literature.

**Topic 7: Probabilistic classification using statistical Naïve Bayes’ classifier**

**Reference: CMC data for the year 2016-2017**

1. Obtain the appropriate contingency table from an training data set comprising the prior and posterior probabilities. You should split data into two sets in an appropriate manner.
2. Test the classifiers using k-fold cross validation technique. Run with different value of k and then choose the optimum result.
3. Furnish the accuracy using an appropriate confusion matrix and report the performance evaluation with different matrix (e.g., Precision, Recall, F1 score, etc.).

**Topic 8: Decision tree based classifier model building using ID3 algorithm**

**Reference: NURSERY data**

1. Built a classifier model based on ID3 algorithm. You should divide the data set randomly in 2:1 ratio using Bootstrap sampling method and then learn the model using the training data set.
2. Verify the classifier’s performance on the test set. Use *k*-fold cross validation with different values of k. Decide the best k-value.
3. Show the performance measure in terms of Accuracy, F1-score, Precision, Recall and Confusion matrix.

**Topic 9: Information measurement based decision tree building using C4.5 algorithm**

**Reference: AUDIOLOGY data**

1. Built a classifier model based on C4.5 algorithm. You should divide the data set randomly in 2:1 ratio using any random sampling method and then learn the model using the training data set.
2. Verify the classifier’s performance on the test set. Report the performance measure in terms of Confusion matrix, Predictive accuracy, F1-score, Precision and Recall.
3. Use *k*-fold cross validation with different values of *k*. Obtain an ROC curve with different values of *k*.

**Topic 10: Decision tree based classifier model building using CART algorithm**

**Reference: SOYABEAN data**

1. Built a classifier model based on CART algorithm.
2. Verify your classifier using both the k-fold and Bootstrap estimation strategies.
3. Report the performance measure in terms of Confusion matrix, Predictive accuracy, F1-score, Precision and Recall in each case of your verification. Obtain an ROC curve comparing the different classifiers you have built during your model validation.

**Topic 11: Error-based classification using Linear Support Vector Machine**

**Reference: BCW data**

1. Built a classifier model based on Linear SVM strategy.
2. Verify your classifier using both the k-fold and Bootstrap estimation strategies.
3. Report the performance measure in terms of Confusion matrix, Predictive accuracy, F1-score, Precision and Recall in each case of your verification. Obtain an ROC curve comparing the different classifiers you have built during your model validation.

**Topic 12: Error-based classification using Non-Linear Support Vector Machine**

**Reference: GLASS data**

1. Built a classifier model based on Non-Linear SVM strategy.
2. Verify your classifier using a number of kernel function and 10-fold cross validation technique.
3. Report the performance measure in terms of Confusion matrix, Predictive accuracy, F1-score, Precision and Recall in each case of your verification. Obtain an ROC curve comparing the different classifiers you have built during your model validation.

**Topic 13: Unsupervised learning techniques with agglomerative clustering using AGNES algorithm**

**Reference: Mushroom data**

1. Implement the AGNES algorithm.
2. You should decide at least two different metrics for similarity measures in your implementation.
3. Calculate cluster quality in each of the clustering techniques you have followed.

**Topic 14: Unsupervised learning techniques with hierarchical clustering using DIANA algorithm**

**Reference: HAYES data**

1. Implement the DIANA algorithm.
2. You should decide at least two different metrics for similarity measures in your implementation.
3. Calculate cluster quality in each of the clustering techniques you have followed.

**Topic 15: Unsupervised learning techniques with partition-based k-Means algorithm**

**Reference: Cancer data**

1. Implement the k-Means algorithm.
2. Calculate cluster quality in each of the clustering techniques with different values of *k*.
3. Analyse the performance of k-Means clustering technique using i) Elbow method and Elbow method and Silhouette analysis

---\*---