List of Projects

Data Analytics (CS61061)
Autumn-2023

Date of Announcement: 11.11.2023

Instructions:

- There is no group in this project. One student one project.
- The particular project assigned to a student should implement that project only. If you implement any other project, it will not be considered for evaluation.
- Any plagiarism attracts rejection of the submissions.
- No credit will be given if a student solves a project which is not assigned to him/her.
- The evaluation of the project performance will be based on the report.
- The report should include all the steps involved in the project implementation with programming code, a snapshot of the output, and results as appropriate.
- **Last date of submission of the report: 01.12.2023, 22:00 hours (IST) (hard deadline).**
- CLICK HERE for the link to submit your project. You can submit only one PDF file of size less than 10MB. Don't submit any .py files, .dcox files, etc.

Projects:

Project Code: DA-01

1. Identify the independent and dependent attributes.
2. Characterize the independent attributes depending on three types of variables
   a. Nominal
   b. Categorical
   c. Continues
3. Find the correlation coefficient between individual independent attributes and dependent ones based on the nature of attributes.
4. Arrange the correlation coefficients in descending order.
5. Identify the three most highly correlated independent attributes from the set of attributes.

Dataset URL: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

Project Code: DA-02

1. Identify the independent and dependent attributes.
2. Characterize the independent attributes depending on three types of variables

a. Nominal
b. Categorical
c. Continues

3. Perform below stated two non-parametric tests between each independent and dependent attribute and calculate p-values.
    a. Mann-Whitney U test for continuous variable
    b. Chi-square test for nominal and categorical variable
4. Arrange the p values in ascending order.
5. Identify the three most significant independent attributes that have a high impact on dependent variable.

Dataset URL : https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

Project Code: DA-03

1. Identify the independent and dependent attributes.
2. Calculate the following statistical details of each attribute and represent them in tabular form
    a. Mean
    b. Median
    c. Mode
    d. Standard deviation
    e. Q1, Q3
    f. Kurtosis
3. Identify the data distribution pattern (normal, skewed) of the proper attributes and remove the outliers accordingly.
4. After removing outliers from the entire dataset, calculate the statistical details of each attribute according to point 2 and represent them in tabular form.

Dataset URL :
https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease

Project Code: DA-04

1. Find the data distribution pattern of each attribute and, according to that, remove the outliers from the attribute values and generate an updated dataset.
2. Create a new attribute, “Spending” with the following attributes.

MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds

[**Note:** ”Spending” is the sum of the amount spent on the 6 product categories]

2. Test the hypothesis, whether there is any monotonic association between income and spending amount. (Test the hypothesis in 5%, 1% level of significance)

Hint: Spearman rank correlation can be applied.

3. Test the hypothesis, whether Education and Marital_Situation are independent or not. (Test the hypothesis in 5%, 1% level of significance)

Hint: Chi-square test can be applied.

Dataset URL :
https://drive.google.com/file/d/13c2CG8hCDh6IU_wsv1mlmKiFl6ONS06o/view?usp=drive_link

Project Code: DA-05

1. Create Histogram plots for all the relevant attributes to visualize the patterns in the dataset.
2. Answer the following:
   a) Does gender affect who gets searched during a stop?
   b) How does drug activity change by time of day?
3. Use suitable statistical hypothesis testing method/methods to check the claim that ‘the average age of the white males who were stopped for speeding is less than 34’.
4. Show the variation of accident frequencies throughout the time of the day. (time of the day versus accident frequencies)

Dataset URL : https://www.kaggle.com/datasets/faressayah/stanford-open-policing-project/data

Project Code: DA-06

1. Using correlation analysis, find out which of the two attributes are mostly correlated.
2. Create a heat map and other plots to show the correlation between all pairs of attributes.
3. Use statistical testing to check the claim that the ‘radius_mean’ of the Malignant tumors is less than 14.

4. Use the Bayesian Classifier to classify between Malignant and Benign tumors. Use 10-fold cross-validation and report the classification accuracy, precision, recall, F1-score, etc., for individual folds as well as the overall average.

Dataset URL : https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

Project Code: DA-07

1. Study and find out which of the ML classifiers are suitable for the classification in this case. Write in the report in detail about the same. What are the adjustments/modifications required for the ML classifiers or for the dataset to perform classification for categorical attributes?
2. Apply all the ML classifiers (that are covered in the theory class) for this classification task with 10-fold cross-validation.
3. Report the classification accuracy, precision, recall, F1-score, etc., for individual folds as well as the overall average.
4. Utilize the lightGBM and CatBoost classifiers, that are claimed to be specialized for categorical data classification. Compare their results with the traditional models.

Dataset URL : https://www.kaggle.com/datasets/uciml/mushroom-classification/data

Project Code: DA-08

We want to predict the amount of precipitation(rain) given the weather attributes for a particular day using regression analysis. Split the dataset into training and testing sets and report the RMSE and R-2 score for each regression problem mentioned below.

1. Are there any missing values, NaN present in this dataset? What are steps that you have taken to handle the same? Write about the same in the report in detail.
2. Use simple linear regression to predict ‘Precip’ utilizing individual weather attributes as mentioned before.
3. Use multiple linear regression to predict ‘Precip’ utilizing all the weather attributes.
4. Use nonlinear regression with order 2, 3, 4, 5, and 6 to predict the ‘Precip’ and compare all the results through scatter plots or other plots.

Dataset URL : **https://www.kaggle.com/datasets/smid80/weatherww2**

Project Code: DA-09

We want to create an ML classification model for this dataset. (CNN strictly prohibited)

1. Perform necessary preprocessing steps. What preprocessing steps that you need to perform? Discuss in detail in the report.
2. Extract meaningful features from the images. The evaluation will be mainly based on the number and quality of features extracted, feature extraction methods, etc. (*Plagiarism strictly prohibited*)
3. Utilize multiple ML algorithms you covered in the theory class for classification, following 10-fold cross-validation.
4. Report the classification accuracy, precision, recall, F1-score, etc., for individual folds as well as the overall average.

Dataset URL :
https://www.kaggle.com/datasets/ganeshmundra/classification-of-images

Project Code: DA-10

1. Identify the independent and dependent attributes.
2. Use K-Fold Cross validation method for Test and Train split of the data.
3. Use the following techniques to understand the relation between experience and salary:
   a. Linear Regression
   b. Polynomial Regression (*should test with multiple polynomial degrees*)
4. Use at least three different evaluation metrics for all the experiments.
5. Find the best method for the given dataset stating proper reasons.

Dataset URL :
https://www.kaggle.com/datasets/saquib7hussain/experience-salary-dataset

Project Code: DA-11

1. Analyze the dataset statistics and provide some fist level insights.
2. Use K-Fold Cross validation method for Test and Train split of the data.
3. Use the following algorithms for classification:
   a. SVM (*OVO*)

    b. SVM (*OVA)*
    c. Decision Tree

4. Explain the confusion matrix for each of the classification algorithms with proper insights.
5. Chose the best model using proper evaluation metric like Precision, Recall, F1-Score and Accuracy.

Dataset URL : https://archive.ics.uci.edu/ml/datasets/Wine

Project Code: DA-12

1. Analyze the dataset statistics and provide some fist level insights.
2. Use feature selection techniques to remove at least two features from the dataset.
3. Use feature extraction to generate a new feature.
4. Use K-Fold Cross validation method for Test and Train split of the data.
5. Use polynomial regression (*should test with multiple polynomial degrees*) for Housing price prediction.
6. Chose the best model using proper evaluation metric and state which dataset gives the best results.

Dataset URL : https://www.kaggle.com/datasets/ashydv/housing-dataset

Project Code: DA-13

1. Analyze the dataset statistics and provide some fist level insights.
2. Use K-Fold Cross validation method for Test and Train split of the data.
3. Use the following algorithms for classification:
    a. SVM (*Use proper technique to fit categorical data*)
    b. Decision Tree
4. Explain the confusion matrix for each of the classification algorithms with proper insights.
5. Chose the best model using proper evaluation metric like Precision, Recall, F1-Score and Accuracy.

Dataset URL : https://archive.ics.uci.edu/dataset/19/car+evaluation

Project Code: DA-14

1. Preprocess the data features and Report.
2. Split the data into 80-10-10% train/validation/test data.
3. Run Bayesian Classifier, Decision Tree Classifier and SVM on the data.

4. Report Accuracy, Precision, Recall, F1-Score, AUC-ROC and Confusion Matrix for each model.

Dataset URL : https://data.world/data-society/cambridge-crime-data-2009-2016

Project Code: DA-15

1. Extracts the features from the images. You can use any popular tools to extract the features. Report the name of the tools and feature extraction procedure.
2. Split the data into 80-10-10% train/validation/test data.
3. Run Bayesian Classifier, Decision Tree Classifier and SVM on the data.
4. Report Accuracy, Precision, Recall, F1-Score, AUC-ROC and Confusion Matrix for each model.

Dataset URL :
https://figshare.com/articles/dataset/brain_tumor_dataset/1512427

Project Code: DA-16

1. Extracts the features from the images. You can use any popular tools to extract the features. Report the name of the tools and feature extraction procedure.
2. Split the train dataset into 90-10% train/validation data and use test dataset for evaluation.
3. Run Bayesian Classifier, Decision Tree Classifier and SVM on the data.
4. Report Accuracy, Precision, Recall, F1-Score, AUC-ROC and Confusion Matrix for each model

Dataset URL :
https://drive.google.com/file/d/1drmV1adl5B8_msbJNAJMIgqvki30qdC3/view?usp=sharing

Project Code: DA-17

1. Extracts the features from the images. You can use any popular tools to extract the features. Report the name of the tools and feature extraction procedure.

2. Split the data into 80-10-10% train/validation/test data.
3. Run Bayesian Classifier, Decision Tree Classifier and SVM on the data.
4. Report Accuracy, Precision, Recall, F1-Score, AUC-ROC and Confusion Matrix for each model.

Dataset URL : https://www.kaggle.com/datasets/jehanbhathena/weather-dataset

Allocation:

Serial No	Roll No	Project Code
1	16CS10055	DA-01
2	18CS30035	DA-02
3	19AE30013	DA-03
4	19AE3AI02	DA-04
5	19CH30030	DA-05
6	19CS30001	DA-06
7	19CS30003	DA-07
8	19CS30009	DA-08
9	19CS30011	DA-09
10	19CS30012	DA-10
11	19CS30016	DA-11
12	19CS30018	DA-12
13	19CS30021	DA-13
14	19CS30027	DA-14
15	19CS30028	DA-15
16	19CS30029	DA-16
17	19CS30031	DA-17
18	19CS30034	DA-01
19	19CS30035	DA-02
20	19CS30036	DA-03
21	19CS30039	DA-04
22	19CS30041	DA-05
23	19CS30043	DA-06
24	19CS30044	DA-07
25	19CS30047	DA-08
26	19CS30051	DA-09
27	19CS30052	DA-10
28	19CS30053	DA-11
29	19CS30055	DA-12
30	19EC39002	DA-13
31	19EC39019	DA-14
32	19EC39023	DA-15
33	19EC39032	DA-16
34	19EC39045	DA-17
35	19EE38009	DA-01
36	19EE38010	DA-02
37	19EE38015	DA-03
38	19EE38018	DA-04
39	19EE38019	DA-05
40	19EE38022	DA-06
41	19EE38023	DA-07
42	19ME31042	DA-08
43	20CE10065	DA-09
44	20CS10004	DA-10
45	20CS10014	DA-11
46	20CS10021	DA-12
47	20CS10022	DA-13
48	20CS10032	DA-14
49	20CS10034	DA-15
50	20CS10035	DA-16
51	20CS10038	DA-17
52	20CS10040	DA-01
53	20CS10043	DA-02
54	20CS10044	DA-03
55	20CS10048	DA-04
56	20CS10051	DA-05
57	20CS10059	DA-06
58	20CS10073	DA-07
59	20CS10075	DA-08
60	20CS10078	DA-09
61	20CS10086	DA-10
62	20CS30009	DA-11
63	20CS30010	DA-12
64	20CS30014	DA-13
65	20CS30025	DA-14
66	20CS30032	DA-15
67	20CS30035	DA-16
68	20CS30036	DA-17
69	20CS30038	DA-01
70	20CS30047	DA-02
71	20CS30049	DA-03
72	20CS30055	DA-04
73	20CS30067	DA-05

74	20CS30068	DA-06
75	20EE3FP59	DA-07
76	20IE10015	DA-08
77	20IE10017	DA-09
78	20IE10046	DA-10
79	20IM10019	DA-11
80	20IM30010	DA-12
81	20MF3IM10	DA-13
82	20MI31013	DA-14
83	20NA30021	DA-15
84	20QE30002	DA-16
85	20QM30005	DA-17
86	21BT10032	DA-01
87	21BT30021	DA-02
88	21BT30030	DA-03
89	21EC10061	DA-04
90	21PH10021	DA-05
91	21PH10040	DA-06
92	21PH10048	DA-07
93	22AE60R11	DA-08
94	22AR60R21	DA-09
95	22EC63R10	DA-10
96	23CD71P01	DA-11
97	23CS60A01	DA-12
98	23CS60D01	DA-13
99	23CS60D02	DA-14
100	23CS60D03	DA-15
101	23CS60R01	DA-16
102	23CS60R02	DA-17
103	23CS60R03	DA-01
104	23CS60R05	DA-02
105	23CS60R07	DA-03
106	23CS60R08	DA-04
107	23CS60R10	DA-05
108	23CS60R12	DA-06
109	23CS60R15	DA-07
110	23CS60R16	DA-08
111	23CS60R18	DA-09
112	23CS60R19	DA-10
113	23CS60R20	DA-11
114	23CS60R23	DA-12
115	23CS60R24	DA-13
116	23CS60R26	DA-14
117	23CS60R27	DA-15
118	23CS60R28	DA-16
119	23CS60R29	DA-17
120	23CS60R30	DA-01
121	23CS60R33	DA-02
122	23CS60R34	DA-03
123	23CS60R35	DA-04
124	23CS60R37	DA-05
125	23CS60R39	DA-06
126	23CS60R40	DA-07
127	23CS60R41	DA-08
128	23CS60R42	DA-09
129	23CS60R44	DA-10
130	23CS60R45	DA-11
131	23CS60R46	DA-12
132	23CS60R48	DA-13
133	23CS60R49	DA-14
134	23CS60R50	DA-15
135	23CS60R51	DA-16
136	23CS60R52	DA-17
137	23CS60R54	DA-01
138	23CS60R56	DA-02
139	23CS60R57	DA-03
140	23CS60R59	DA-04
141	23CS60R62	DA-05
142	23CS60R63	DA-06
143	23CS60R64	DA-07
144	23CS60R65	DA-08
145	23CS60R66	DA-09
146	23CS60R67	DA-10
147	23CS60R68	DA-11
148	23CS60R69	DA-12
149	23CS60R70	DA-13
150	23CS60R71	DA-14
151	23CS60R73	DA-15
152	23CS60R74	DA-16
153	23CS60R75	DA-17
154	23CS60R76	DA-01
155	23CS60R78	DA-02
156	23CS60R79	DA-03
157	23CS60R81	DA-04
158	23CS60R82	DA-05
159	23RE91R01	DA-06

Link to submit your project report:
https://docs.google.com/forms/d/e/1FAIpQLSdK1OiVpT-5hZfOdqQLHgUbD3bzLNpjX4jn_suFla31ynUp8g/viewform?usp=sf_link
(Only one submission is allowed.)