

CSBP320 Assignment

Data Analytics

Fall Semester, 2022

Dataset Due Date:	November 17 th
Video Presentation Due Date:	November 27 th
Interview Date:	Last lecture of class
Where to Submit:	Electronic submission

The goal of this assignment is to gain practical experience is using Weka and applying it to storytelling with data. You can work on this assignment on teams of up to three students. **The standard academic honesty rules apply.**

First Deliverable – Dataset:

You will be selecting and downloading a dataset of your choice for the purpose of this assignment. Your selected dataset should satisfy the following criteria:

- Contains at least 5 dimensions/features, including at least one categorical and one numerical dimension.
- Contains a clear class label attribute (binary or multi-label).
- Be of a simple tabular structure (i.e., no time series, multimedia, etc.).
- Be of reasonable size, and contains at least 2K tuples.

While the assignment is open ended, you are expected to select an interesting dataset, which in turn tells an interesting story. Many such datasets are available on public repositories such as: UCI, Kaggle, KDnuggets, etc. Attached, please find some suggested datasets to select from.

Second Deliverable – Video Presentation:

Data Exploration Tasks

- The name and source of dataset.
- A description of how the dataset was collected or created.
- A summary of the purpose of each column in your dataset, including the class label.
- An overview of the data .ARFF file, and how you created it (if needed).
- Explain any data quality problems you might have faced, and how it was handled.
- Provide a visual overview of all the attributes in your dataset.
- Discuss the top distinctive categorical attribute, which is highly correlated to the class label. Support your discussion with a visualization of that attribute.
- Discuss the top distinctive numerical attribute, which is highly correlated to the class label. Support your discussion with a discretized visualization of that attribute.
- Identify and discuss one attribute that clearly has no impact on the class label. Support your discussion with a visualization of that attribute.

Data Analytics Tasks:

In the following, always use K-nearest neighbor classification algorithm

Task 1. Using the default settings of K-nearest neighbor, report on the performance of your classifier (e.g., accuracy, precision, recall, etc.).

Task 2. Now try different values of K, and report on the obtained performance for those different values.

Task 3. In your opinion, what is the most suitable setting of K for your dataset?

Task 4. Given your answer to the previous task, try different settings for the split ratio and report on the obtained performance.

Task 5. Compare the performance of Task 4 to that of a cross-fold data partitioning.

In the following, always use the decision tree classification algorithm

Task 6. Using the default settings of the decision tree, report on the performance of your classifier (e.g., accuracy, precision, recall, etc.).

Task 7. Inspect the visualization of the obtained decision tree, and discuss: 1) the most distinctive features of your dataset, and 2) any interesting observations learned from the tree structure.

Task 8. Compare the observations obtained from Task 6 to your findings in the Data Exploration tasks of the assignment. That is, how the features you identified in the exploration phase are similar or different to the ones from Task 6.

Task 9. Adjust the decision tree parameters to allow overfitting, and compare to the results obtained in the previous task in terms of: 1) tree structure, and 2) classifier performance.

Task 10. In your opinion, what would be the best settings for the decision tree classifier for your dataset?

Deliverables:

Selected Dataset (Due November 17th) : Upload to blackboard a .pdf file that includes the following: 1) a list of your top three preferred datasets, 2) for each of the selected 3 datasets, include the following: a) name and link for the dataset you have selected, b) a brief summary describing the purpose of that dataset, c) a checklist showing that the dataset satisfies the requested criteria (e.g., number of features, number of instances, data structure, etc), d) the class label you have identified for the dataset, e) any special remarks you might have regarding the dataset. Based on your submission, one of your selected datasets will be assigned to you to work on for the 2nd deliverable.

Video Presentation (Due Nov. 27th): Make sure to plan your slides for a 10-minute presentation. Your slides should cover the highlights and main findings of as many tasks as possible. Your slides should also clearly specify the work done by each team member. During presentation time, each team member will be responsible for presenting their part of the work. Your mark will be based on: 1) the quality and depth of the slides, and 2) the quality of presentation.

Please note:

- . Any delays in submitting your deliverables will result in a late penalty of 25% for each additional day after the deadline.

Suggested Datasets:

Census Income Data Set

<https://www.kaggle.com/uciml/adult-census-income>

Customer Churn Prediction 2020

<https://www.kaggle.com/competitions/customer-churn-prediction-2020/overview>

Cardiovascular Disease dataset

https://www.kaggle.com/sulianova/cardiovascular-disease-dataset?select=cardio_train.csv

Stroke Prediction Dataset

<https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>

Engineering Placements Prediction

<https://www.kaggle.com/tejashvi14/engineering-placements-prediction>

Airline Passenger Satisfaction dataset

<https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction?select=train.csv>

Loan Prediction Based on Customer Behavior

<https://www.kaggle.com/subhamjain/loan-prediction-based-on-customer-behavior?select=Training+Data.csv>

Enjoy your assignment!