CSE 60647: Data Science **Fall 2025**

Coding Assignment (60 points in total)

Handed Out: October 6, 2025 *Due: November 17, 2025 11:59pm*

In the coding assignments, students are allowed to use any tools such as Google Search and ChatGPT to help them solve the problems. However, they are NOT allowed to copy other students' solutions. Students are NOT suggested to directly copy and paste the AI-generated code for homework. Multiple professors find that since ChatGPT becomes available, they have been not able to replicate the results of submitted codes significantly more often. If a submission is not executable *independently* on Google Colab or fails to give the correct answers, it cannot get the points. Read this assignment and then work on the corresponding Colab notebook. Submit your completed notebook.

1 [40] Classification Algorithms

Given a data set from the Notre Dame Fighting Irish Football database on FBSchedules, please use **ID3 Decision Tree** models, **Naïve Bayes** models, and other types of models to predict whether the team will win or lose a game. The data table is given on a full page.

Data set: Each data object is a game. We have four attributes about the game:

- **Is Night Game (NG)** (after 7:30PM ET): a 2-value attribute ("Yes", "No");
- **Is Home or Away (HorA)**: a 3-value attribute ("Home", "Away", "Neutral");
- **Is Opponent in AP Top 25 at Preseason (Top25)**: a 2-value attribute ("In", "Out");
- **Media**: a 3-value attribute ("1-ABC", "2-NBC", "3-XYZ"). "XYZ" means others, including Peacock, ESPN, CBS, and P12N.

The label is **Win or Lose**. It is a binary label ("Win", "Lose").

Training set: the games in the seasons 2022 and 2023 with IDs 1–26 in the data table.
Test set: the games in season 2024 with IDs 27–41 in the data table.

1. [2] Preprocess the data table in the format of a LaTeX table to be used for the machine learning study.
2. [3] Use sklearn.preprocessing.OneHotEncoder to convert the data table into np.array for sklearn use. There were four attributes. Now how many features (i.e., columns) are there in the arrays (e.g., training vectors, test vectors)?
3. [5] Define the positives and negatives considering the label imbalance in this binary classification task. Suppose the first baseline predicts W/L at 50%. Report its accuracy, precision, recall, and F1 score. Suppose the second baseline predicts W/L at the probability based on the W-L ratio in the 26 training data points. Report its accuracy, precision, recall, and F1 score.

4. [5] For each feature, calculate the correlation between it and the label. Find the top three strongest correlated features, positively or negatively. Find the least correlated feature with the Win/Lose label.

5. [5] Use sklearn.tree.DecisionTreeClassifier to build a model based on the 26 training data points. Draw the final decision tree.

6. [2] Use the tree model to make predictions on the test data points. Report accuracy, precision, recall, and F1 score.

7. [5] Build two new tree models: one is on the 13 training points from season 2022, and the other is on the 13 data points from season 2023. Compare their final trees against the one that was trained on all the 26 data points. Compare their predictions on the test data points. Compare the accuracy, precision, recall, and F1 score of these two models against the original one.

8. [5] Use sklearn.naive_bayes.GaussianNB to build a Naive Bayes model using the 26 training data points to predict the labels of the test data points. Report accuracy, precision, recall, and F1 score.

9. [5] Use at least **THREE** other types of classification algorithms in sklearn, such as random forest, AdaBoost, support vector machine, and multi-layer perceptrons (neural network) to build the models, make predictions, and report their accuracy, precision, recall, and F1 score.

10. [3] Summarize so far what classifiers perform better than the two baselines.

2 [20] Clustering Algorithms

The dataset contains a comprehensive list of popular songs listed on Spotify. The dataset also contains information such as track name, artist(s) name, release year, Spotify playlists and charts, streaming statistics, Apple Music presence, and various audio features.

- Import the dataset using 'pandas' and assign the dataframe to 'df'.
- Extract the numeric columns of the dataframe as a list, and assign it to the variable 'numeric_columns'.
- Normalize the data using 'MinMaxScaler'. Assign the normalized data to 'df'.
- Create the variable 'X' with the numeric columns.
- Create the variable 'X_simple' extracting 'in_spotify_playlists' and 'in_apple_playlists' from 'X'.
- Let's create four clusters using K-Means. Create the model 'kmeans' using the function 'KMeans'. The number of clusters is 3. Train 'kmeans' using 'X_simple'.

- Predict the clusters using the ‘kmeans‘ model and the dataframe ‘X_simple‘. Assign the clusters to the variable ‘y_kmeans‘.

- Extract the centroids and assign the variables to ‘centroids‘.

- Using the function ‘plt.scatter‘, print the datapoints and assign their color using the predicted cluster (i.e., ‘y_kmeans‘). Plot the centroids with another overlapping scatter plot.

- The following code includes the function ‘calculate_sse‘ which computes all the K-Means models from 1 to ‘max_num_clusters‘. We calculate each model's sum of squared errors (SSE) and plot them. Run the code.

- Run the function using ‘X_simple‘ and with 10 as the value for ‘max_num_cluster‘.

- We will now use DBSCAN to find clusters from ‘X_simple‘. Using the function ‘DBSCAN‘ create an initial set of clusters. Use the parameters ‘eps=0.05‘ and ‘min_samples=3‘. Assign the results to the variable ‘clustering‘.

- Create the variable ‘y_dbscan‘ with the cluster values from ‘clustering‘.

- Print the plot using the function ‘plt.scatter‘. Color the datapoints using ‘y_dbscan‘.

ID	Date	Opponent	Is Night Game?	Is Home or Away?	Is Opponent in AP Top 25 at Preseason?	Media	Label: Win/ Lose
1	9/3/22	OSU	Yes	Away	In	1-ABC	L
2	9/10/22	Marshall	No	Home	Out	2-NBC	L
3	9/17/22	California	No	Home	Out	2-NBC	W
4	9/24/22	UNC	No	Away	Out	1-ABC	W
5	10/8/22	BYU	Yes	Neutral	In	2-NBC	W
6	10/15/22	Stanford	Yes	Home	Out	2-NBC	L
7	10/22/22	UNLV	No	Home	Out	3-XYZ	W
8	10/29/22	Syracuse	No	Away	Out	1-ABC	W
9	11/5/22	Clemson	Yes	Home	In	2-NBC	W
10	11/12/22	Navy	No	Neutral	Out	1-ABC	W
11	11/19/22	BC	No	Home	Out	2-NBC	W
12	11/26/22	USC	Yes	Away	In	1-ABC	L
13	12/30/22	SC	No	Neutral	Out	3-XYZ	W
14	8/26/23	Navy	No	Neutral	Out	2-NBC	W
15	9/2/23	TSU	No	Home	Out	2-NBC	W
16	9/9/23	NC State	No	Away	Out	1-ABC	W
17	9/16/23	CMich	No	Home	Out	3-XYZ	W
18	9/23/23	OSU	Yes	Home	In	2-NBC	L
19	9/30/23	Duke	Yes	Away	Out	1-ABC	W
20	10/7/23	Louisville	Yes	Away	Out	1-ABC	L
21	10/14/23	USC	Yes	Home	In	2-NBC	W
22	10/28/23	Pitt	No	Home	Out	2-NBC	W
23	11/4/23	Clemson	No	Away	In	1-ABC	L
24	11/18/23	WF	No	Home	Out	2-NBC	W
25	11/25/23	Stanford	Yes	Away	Out	3-XYZ	W
26	12/29/23	OregonS	No	Neutral	In	3-XYZ	W
27	8/31/24	TAMU	Yes	Away	In	1-ABC	W
28	9/7/24	NIU	No	Home	Out	2-NBC	L
29	9/14/24	Purdue	No	Away	Out	3-XYZ	W
30	9/21/24	Miami(OH)	No	Home	Out	2-NBC	W
31	9/28/24	Louisville	No	Home	Out	3-XYZ	W
32	10/12/24	Stanford	No	Home	Out	2-NBC	W
33	10/19/24	GT	No	Away	Out	3-XYZ	W
34	10/26/24	Navy	No	Neutral	Out	1-ABC	W
35	11/9/24	FSU	Yes	Home	In	2-NBC	W
36	11/16/24	Virginia	No	Home	Out	2-NBC	W
37	11/23/24	Army	Yes	Neutral	Out	2-NBC	W
38	11/30/24	USC	No	Away	In	3-XYZ	W
39	12/20/24	Indiana	Yes	Home	Out	1-ABC	W
40	1/2/25	Georgia	No	Neutral	In	3-XYZ	W
41	1/9/25	PSU	Yes	Neutral	In	3-XYZ	?