**ASSIGNMENT #1:**

**Wine Quality Classification using KNN**

Use the winequality.csv dataset for this first assignment

**NOTE: reference SAMPLE CODES IN ZIP for code guidance**

1. Load the data from the file winequality.csv. (2)
2. Standardize all variables other than Quality. (2)
3. Partition the dataset:
   1. random\_state = 42 (1)
   2. Partitions 60/20/20 (1)
   3. Make sure to stratify! (1)
4. Iterate on K ranging from 1 to 30.
   1. Build a KNN classification model to predict Quality based on all the remaining numeric variables. (2)
   2. Plot the accuracy for **both** the Training and Validation datasets. (4)
5. Which value of k produced the best accuracy in the Training and Validation data sets? (2)
6. Generate predictions for the test partition with the chosen value of k. Plot the confusion matrix of the actual vs predicted wine quality. (4)
7. Print the accuracy of model on the test dataset. (2)
8. Print the test dataframe with the added columns “Quality” and “Predicted Quality” (4)

**ASSIGNMENT #2:**

**Personal Loan Prediction Using Trees**

Use the UniversalBank.csv dataset for this assignment

**NOTE: reference SAMPLE CODES IN ZIP for code guidance**

*This dataset is taken from the website of the book “Data mining for Business Intelligence.” The dataset provides information about many people and our goal is to build a model to classify the cases into those who will accept the offer of a personal loan and those who will reject it. In the data, a zero in the Personal Loan column indicates that the concerned person rejected the offer and a one indicates the person accepted the offer*

1. What is the target variable? (2)
2. Remove the attributes *Row* and *Zip code*. (3)
3. Partition the dataset:
   1. random\_state = 42 (1)
   2. Partitions 70/30 (1)
   3. Make sure to stratify! (1)
4. How many of the cases in the **training** partition represented people who accepted offers of a personal loan? (3)
5. Plot the classification tree Use entropy criterion. Max\_depth = 5, random\_state = 42. (4)
6. On the training partition, how many **acceptors** did the model classify as **non-acceptors**? (3)
7. On the training partition, how many **non-acceptors** did the model classify as **acceptors**? (3)
8. What was the accuracy on the training partition? (2)
9. What was the accuracy on the test partition? (2)

**ASSIGNMENT #3:**

**Mushroom Edibility Using Trees**

*Build a classification model that predicts the edibility of mushrooms (class variable in the dataset).*

*Use the* mushrooms.csv *file*

**NOTE: reference SAMPLE CODES IN ZIP for code guidance**

* cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
* cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
* cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
* bruises?: bruises=t,no=f
* odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
* gill-attachment: attached=a,descending=d,free=f,notched=n
* gill-spacing: close=c,crowded=w,distant=d
* gill-size: broad=b,narrow=n
* gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
* stalk-shape: enlarging=e,tapering=t
* stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
* stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
* stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
* stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
* stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
* veil-type: partial=p,universal=u
* veil-color: brown=n,orange=o,white=w,yellow=y
* ring-number: none=n,one=o,two=t
* ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
* spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
* population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
* habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d
* class: p = poisonous, e=edible

Steps

1. Build a classification tree!
2. Partition the dataset:
   1. random\_state = 42 (1)
   2. Partitions 70/30 (1)
   3. Make sure to stratify! (1)
   4. max\_depth = 6
   5. Use Entropy
3. Print the confusion matrix. Also visualize the confusion matrix using plot\_confusion\_matrix from sklearn.metrics (5)
4. What was the accuracy on the training partition? (2)
5. What was the accuracy on the test partition? (2)
6. Show the classification tree. (4)
7. List the top three most important features in your decision tree for determining toxicity. (6)
8. Classify the following mushroom (6):
   1. cap-shape: x
   2. cap-surface: s
   3. cap-color: n
   4. bruises: t
   5. odor: y
   6. gill-attachment: f
   7. gill-spacing: c
   8. gill-size: n
   9. gill-color: k
   10. stalk-shape: e
   11. stalk-root: e
   12. stalk-surface-above-ring: s
   13. stalk-surface-below-ring: s
   14. stalk-color-above-ring: w
   15. stalk-color-below-ring: w
   16. veil-type: p
   17. veil-color: w
   18. ring-number: o
   19. ring-type: p
   20. spore-print-color: r
   21. population: s
   22. habitat: u

**ASSIGNMENT #4:**

**Vehicle MPGs Using Linear Regression**

Use the auto-mpg.csv dataset for this assignment

**NOTE: reference SAMPLE CODES IN ZIP for code guidance**

1. Summarize the data set. What is the mean of mpg? (2)
2. What is the median value of mpg? (1)
3. Which value is higher – mean or median? What does this indicate in terms of the skewness of the attribute values? Make a plot to verify your answer. (2)
4. Plot the pairplot matrix of all the relevant numeric attributes. (don’t consider No and car\_name)? (2)
5. Based on the pairplot matrix, which two attributes seem to be most strongly linearly correlated? (2)
6. Based on the pairplot matrix, which two attributes seem to be most weakly correlated. (2)
7. Produce a scatterplot of the two attributes mpg and displacement with displacement on the x axis and mpg on the y axis. (2)
8. Build a linear regression model with mpg as the target and displacement as the predictor. Answer the following questions based on the regression model.
   1. For your model, what is the value of the intercept β0 ? (1)
   2. For your model, what is the value of the coefficient β1 of the attribute displacement? (1)
   3. What is the regression equation as per the model? (2)
   4. For your model, does the predicted value for mpg increase or decrease as the displacement increases? (2)
   5. Given a car with a displacement value of 220, what would your model predict its mpg to be? (2)
   6. Display a scatterplot of the actual mpg vs displacement and superimpose the linear regression line. (2)