ITP 449
Final Project, Fall 2020

100 points

For each one of the following questions, write Python code in PyCharm.

- For each question, create a *new* Python file. Name each
  lastname_firstname_finalproject_q#.py
- Create a header in each file using *comments* to display your name and HW information. After that write your Python code.
  # Tommy Trojan
  # ITP 449 Fall 2020
  # Final Project
  # Q1
- Create a Word document. Insert screenshots of your output
- Zip the python files and the Word doc. Submit it on Blackboard.

Problem 1 (25 points)

Wine Quality classification using KNN

For this problem you will be doing classification with KNN. The goal is to predict the quality of wine given the other attributes.

random_state = 2020, *Partitions 60/20/20*, stratify = y

A. Load the data from the file *winequality.csv*.
B. Standardize all variables other than *Quality*.
C. Partition the dataset
D. Build a KNN classification model to predict Quality based on all the remaining numeric variables.
E. Iterate on K ranging from 1 to 30. Plot the accuracy for the train A and train B datasets.
F. Which value of k produced the best accuracy in the train A and train B data sets?
G. Generate predictions for the test partition with the chosen value of k. Print the confusion matrix of the actual vs predicted wine quality.
H. Print the test dataframe with the added columns “Quality” and “Predicted Quality”
I. Print the accuracy of model on the test dataset.

	Fixed acidity	Volatile Acidity	...	Quality	PredictedQuality
1535	-0.758172	0.123905	...	6	6
1528	-0.241094	-1.328579	...	6	5
659	-0.700719	1.743983	...	4	5
1102	-1.275249	-0.267148	...	6	6
96	-0.873078	1.380862	...	5	5
...	...	...	...	...	...
241	2.114480	-0.825796	...	6	6
863	-0.643266	0.514959	...	5	5
668	0.865420	1.129471		5	5

Problem 2 (25 points)

Load the "*UniversalBank.csv*" (this dataset is taken from the website of the book "Data mining for business intelligence" by Shmueli, Patel and Bruce, 1st ed, Wiley 2006). The data set 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 that the person accepted the offer. Answer the following questions:

A. What is the target variable?
B. Ignore the variables *Row* and *Zip code*.
C. Partition the data *70/30, random_state = 2020, stratify=y*
D. How many of the cases in the ***training*** partition represented people who accepted offer of a personal loan?
E. Plot the *classification tree* Use *entropy* criterion. *max_depth = 5, random_state = 2020*
F. On the ***training*** partition, how many **acceptors** did the model classify as **non-acceptors**?
G. On the ***training*** partition, how many **non-acceptors** did the model classify as **acceptors**?
H. What was the accuracy on the training partition?
I. What was the accuracy on the test partition?

Problem 3 (25 points)

Your task is to build a classification model that predicts the edibility of mushrooms (*class* variable in the dataset). You have been provided with a dataset as a *mushrooms.csv* file. Here is a description of the attributes:

Attribute description:

1. cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
2. cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s
3. cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
4. bruises: bruises=t, no=f
5. odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
6. gill-attachment: attached=a, descending=d, free=f, notched=n
7. gill-spacing: close=c, crowded=w, distant=d
8. gill-size: broad=b, narrow=n
9. 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
10. stalk-shape: enlarging=e, tapering=t
11. stalk-root: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
12. stalk-surface-above-ring: fibrous=f, scaly=y, silky=k, smooth=s
13. stalk-surface-below-ring: fibrous=f, scaly=y, silky=k, smooth=s
14. stalk-color-above-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
15. stalk-color-below-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
16. veil-type: partial=p, universal=u
17. veil-color: brown=n, orange=o, white=w, yellow=y
18. ring-number: none=n, one=o, two=t
19. ring-type: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
20. spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
21. population: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
22. habitat: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d
23. class: p = poisonous, e=edible

Build a *classification tree.* Random_state =2020. Training partition 0.7. stratify = y, max_depth = 6

A. Print the confusion matrix. Also visualize the confusion matrix using *plot_confusion_matrix* from *sklearn.metrics*
B. What was the accuracy on the training partition?
C. What was the accuracy on the test partition?
D. Show the classification tree.
E. List the top three most important features in your decision tree for determining toxicity.
F. Classify the following mushroom.

Class	?
cap-shape	X
cap-surface	S
cap-color	N
Bruises	T
Odor	Y
gill-attachment	F
gill-spacing	C
gill-size	N
gill-color	K
stalk-shape	E
stalk-root	E
stalk-surface-above-ring	S
stalk-surface-below-ring	S
stalk-color-above-ring	W
stalk-color-below-ring	W
veil-type	P
veil-color	W
ring-number	O
ring-type	P
spore-print-color	R
population	S
Habitat	U

Problem 4 (25 points)

Load the data from the file *auto-mpg.csv.* The file contains information about various cars made between 1970 and 1982. The file contains 398 rows of data. Table below shows an extract of the first 10 rows to give you an idea of the data.

No	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	car_name
1	28	4	140	90	2264	15.5	71	chevrolet vega 2300
2	19	3	70	97	2330	13.5	72	mazda rx2 coupe
3	36	4	107	75	2205	14.5	82	honda accord
4	28	4	97	92	2288	17	72	datsun 510 (sw)
5	21	6	199	90	2648	15	70	amc gremlin
6	23	4	115	95	2694	15	75	audi 100ls
7	15.5	8	304	120	3962	13.9	76	amc matador
8	32.9	4	119	100	2615	14.8	81	datsun 200sx
9	16	6	250	105	3897	18.5	75	chevroelt chevelle malibu
10	13	8	318	150	3755	14	76	dodge d100

A. Summarize the data set. What is the mean of mpg?
B. What is the median value of mpg?
C. 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. Hint: Look
D. Plot the pairplot matrix of all the relevant numeric attributes. (don't consider *No*)?
E. Based on the pairplot matrix, which two attributes seem to be most strongly linearly correlated?
F. Based on the pairplot matrix, which two attributes seem to be most weakly correlated.
G. Produce a scatterplot of the two attributes mpg and displacement with *displacement* on the x axis and mpg on the y axis.
H. Build a linear regression model with mpg as the target and *displacement* as the predictor. Answer the following questions based on the regression model.
    a. For your model, what is the value of the intercept $\beta 0$?
    b. For your model, what is the value of the coefficient $\beta 1$ of the attribute displacement?
    c. What is the regression equation as per the model?
    d. For your model, does the predicted value for mpg increase or decrease as the displacement increases?
    e. Given a car with a displacement value of 220, what would your model predict its mpg to be?
    f. Display a scatterplot of the actual mpg vs displacement and superimpose the linear regression line.
    g. Plot the residuals.