FINAL EXAM - IT 418/518

Spring 2020

250 Points; 2 Hours

Note: This exam is an open book exam. However, you are not allowed to speak to other students during the exam.

Submission Instructions: Submit your code, output, and explanations in a Jupyter Notebook. Export your notebook as a PDF. Name your files as YourLastName_IT418_Final.pynb and YourLastName_IT418_Final.pdf. Upload your files in D2L.

Clearly state the question number prior to your answer. Use Markup and HTML to show your question number clearly. **All your answers must be in the alphabetical order (a-z).** You must submit the Jupyter Notebook (not a Python file) and a PDF. Your Jupyter Notebook and PDF must contain the codes, outputs, and your explanations. Properly comment your code. **20 points will be deducted from your score if you do not follow any of the instructions.**

****

Imagine that you are a data scientist at a hospital. Your goal is to develop a model so pregnant women who have a greater likelihood of developing gestational diabetes could be identified. All patients in the dataset are women of at least 21 years of age. The dataset (patients.csv) has several independent variables and a dependent variable (labeled as Diagnosis). For this project, you will use the provided dataset (patients.csv). The following is the description of each variable in the dataset:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration 2 hours after an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- Pedigree: Diabetes pedigree function
- Age: Age (years)
- Diagnosis: Class (diabetes/not diabetes, 1,0)

Create a binary classification model using the provided dataset to predict diabetes. Provide code and solve each of the following steps. Use the Markdown option to write your explanations in the Jupyter Notebook. This exercise will utilize Python libraries such as numpy, matplotlib, seaborn, pandas, and scikit Learn.

a) Load the dataset in Jupyter Notebook. (1)
b) Using head(), examine the data. (1)

c) Provide mean, median, standard deviation, and quartiles for each independent variable. Explain your results. (8)
d) Find missing values for each independent variable and fill them with median values. (5)
e) Find outliers for each independent variable using the IQR rule. (5)
f) Replace outliers with median values. (4)
g) Set Matplotlib to plot inline. (1)
h) Create a histogram for each variable to see distribution. (5)
i) Explain the histograms. (5)
j) Create a boxplot for each of the variables. (5)
k) Explain the boxplots (5)
l) Create a heat map to see the correlation between variables using seaborn. Which variables are most correlated? (10)
m) Find the best performing features using feature extraction in scikit Learn. (10)
n) Standardize your features to Gaussian distribution. (5)
o) Split the dataset into 60/40 training and testing. (5)
p) Create a logistic regression model (call it LRM1) using your best features. Describe your model. (15)
q) Create classification report of your model. (5)
r) Describe your classification report (precision, recall, F1 score, and support). (10)
s) Create the accuracy score of your model. Describe the accuracy score. (10)
t) Create another logistic regression model (call it LRM2). Use all the independent features this time (instead of your best performing features). (10)
u) Compare the two models (LRM1 and LRM2) based on the classification report and accuracy score. Which one is a better model? Why? (10)
v) Create a Naïve Bayes model (call it NBM) using 60/40 split. (10)
w) Create classification report of your NBM model. (5)
x) Describe your classification report of NBM (precision, recall, F1 score, and support). (10)
y) Create the accuracy score of your NBM model. Describe the accuracy score. (10)
z) Compare the logistic regression (LRM1 or LRM2) with the Naïve Bayes model (NBM). Which one is better? Why? (20)
aa) What would be your suggestions for further improving the accuracy of your chosen model? (20)
bb) What would be the pitfalls or weaknesses of your model if the hospital decided to deploy it to predict diabetes? (20)
cc) If you were to present your analysis and findings to the CEO of the hospital, what would be your top five key points? (20)

IT 518 only – (20 points will be deducted for incorrect answer or non-submission)

dd) How did you check for multicollinearity in your model? (10)
ee) Create a data visualization of your choice that can show an interesting insight. Your visualization can be of any kind if it is not covered in a previous question. (10)

#######