Last Name: ______________________ First Name: ______________________ Student ID: ____________________

AIDI 1002: Machine Learning Programming — Final Exam Fall 2022

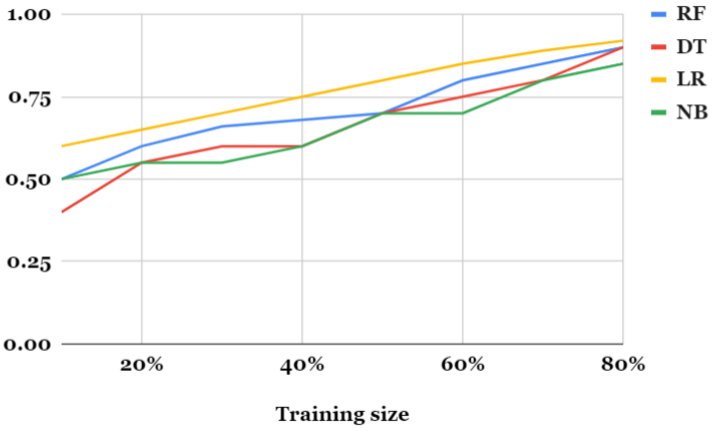
Due Date : December 12, 2022, 2:00 PM - 4:00 PM

Note : Submit two files in the submission folder. First is your colab notebook including your code and outputs and second is the pdf of colab notebook with the following naming convention for both the files.

(File name : $Malik\_Garima\_FinalExam\_AIDI1002.pdf/.ipynb$)

1. (30 Points) Increasing Training Set Size Experiment: Consider the iris dataset for multiclass classification and perform the following steps.

   1. Divide the data into 80% training and 20% testing.
   2. From the training set only take 5% of the data and train the supervised learning models (Logistic Regression, Decision Trees, Random Forest, and Naive Bayes) and test it on the test set created in the previous step.
   3. Repeat the training again with now 10% of the data and keep on adding the 5% until you use the whole training set.
   4. In every training test on the 20% of the test set and report the accuracy and f1-score of the model.
   5. Plot the sample graph for accuracy and f1-score as provided below:



2. (30 Points) Linear Regression: Consider the following N data points (N=10):

$$X = [-1.4, -1.6, -1.3, 0.2, 2.0, -1.1, 0.0, 0.3, -0.9, -1.8]$$
$$r = [6.9, 7.8, 8.0, 5.8, 1.9, 7.3, 5.8, 5.8, 8.2, 9.6]$$

Note that these data points are ordered, so that $(X_1, r_1) = (-1.3, 6.9)$ and $(X_{10}, r_{10}) = (-1.8, 9.6)$. For the above data points, fit a linear regression model, $f(x) = w_0 + w_1 x$, by estimating the values of w_0 and w_1, so that $r_i = f(X_i) + \varepsilon_i$.

Hint: $w_0 = \bar{r} - w_1 \bar{X}$, $\quad w_1 = \dfrac{\sum_{i=1}^{N} X_i r_i - \bar{X}\bar{r}N}{\sum_{i=1}^{N}(X_i)^2 - N(\bar{X})^2}$ where $\bar{r} = \dfrac{1}{N}\sum_{i=1}^{N} r_i$, $\bar{X} = \dfrac{1}{N}\sum_{i=1}^{N} X_i$

Answer:

$w_0 =$	$w_1 =$

- Compute root mean squared error (RMSE) and median absolute error (MAE) values for given training data set (X, r). Hint: $RMSE = \frac{1}{N}\sum_{i=1}^{N}(f(X_i) - X_i)^2$, $MAE = numpy.median(abs(e_i))$ where $e_i = f(X_i) - X_i$

  Answer:

$RMSE =$	$MAE =$

- What are the linear regression model predictions for the following test data?

```
Z = [−0.6, 1.8, −0.1, 1.1, −1.7]
```

  Answer:

$f(Z_1) =$	$f(Z_2) =$	$f(Z_3) =$	$f(Z_4) =$	$f(Z_5) =$

- Consider the following values to be true labels for data points Z.

```
u = [5.1, −0.2, 6.5, 2.2, 8.3]
```

  Compute RMSE and MAE values for given test data set (Z, u).

  Answer:

$RMSE =$	$MAE =$

3. (40 points) K-Means Clustering: Consider the 30 data points and their corresponding class labels stored in a dictionary named “data_dict”.

```
data_dict = {(2.0, 3.43, 4.37):2, (2.49, 4.28, 4.83):2, (2.58, 4.36, 4.48):2, (2.66, 4.45, 5.95):2,
(2.82, 3.66, 4.51): 2, (3.03, 4.37, 5.07): 2, (3.27, 4.54, 4.57): 2, (3.41, 3.94, 5.35): 2,
(3.53, 4.32, 5.41): 2, (3.53, 4.6, 6.8): 1, (3.61, 4.25, 5.21): 1, (3.61, 4.78, 5.47): 1,
(3.72, 5.44, 5.88): 1, (3.87, 4.96, 4.52): 2, (4.13, 5.29, 6.6): 1, (4.25, 5.97, 5.48): 1,
(4.61, 4.9, 5.11): 1, (4.73, 4.4, 6.78): 1, (4.97, 4.25, 5.0): 1, (4.98, 5.27, 6.79): 1,
(5.08, 3.51, 4.69): 3, (5.15, 3.58, 4.2): 3, (5.67, 2.27, 4.65): 3, (5.67, 3.81, 5.75): 3,
(5.94, 2.34, 4.12): 3, (6.06, 3.16, 4.36): 3, (6.09, 3.19, 4.02): 3, (6.43, 3.42, 4.18): 3,
(6.56, 2.7, 4.03): 3, (6.79, 3.46, 4.81): 3}
```

For instance, the first point has coordinates $(x_1, x_2, x_3) = (2.0, 3.43, 4.37)$ and belongs to class 2. In total we have three classes: 1, 2, and 3.

As a discriminant function, consider a distance function based on below center coordinates (encoded as a dictionary of values) for each class labels.

```
centers_dict = {}
centers_dict[(3,4,5)] = 1 # center coordinates for class 1, i.e., c1=4, c2=5, c3=6
centers_dict[(4,5,6)] = 2 # center coordinates for class 2, i.e., c1=3, c2=4, c3=5
centers_dict[(6,3,5)] = 3 # center coordinates for class 3, i.e., c1=6, c2=3, c3=5
```

Note that a discriminant function based on Minkowski distance can be written as

$$g(x) = (\sum_{i=1}^{n} |c_i - x_i|^p)^{1/p} \quad \text{where } x_i = (x_1, x_2, x_3), \quad c_i = (c_1, c_2, c_3)$$

Based on above discriminant functions, perform a K-Means Clustering task over 30 points in data_dict and then compare it with true labels. What is the number of correctly classified instances for each value of p in distance measure?

$p = 1:$	$p = 2:$	$p = 3:$