**CMPSC 445: Applied Machine Learning in Data Science**

**Spring 2023**

**Project-1: Iris classification using KNN**

**Gabriel Nulman – CMPSC Senior  
Professor V. Elangovan – Department Chair CMPSC**

Goal: For the iris dataset, develop a KNN classifier to classify different iris species.

Brief information about the data:

This dataset consists of 4 attributes: sepal-length, sepal-width, petal-length and petal-width. Our goal is to predict the iris class given the 4 attributes. There are 3 classes in the dataset: Iris-setosa, Iris-versicolor and Iris-virginica. For more detailed information about the dataset, visit the above website.

1. **Introduction:**

* KNN is a simple supervised machine learning algorithm used primarily for regression and classification. In this project, we will use it for regression to predict values based on k nearest neighbors. We will also test for errors to obtain the most optimal k value(s).
* The Iris dataset is another dataset present within the Seaborn library. This 150-record dataset is very easy to navigate, in addition to have 3 specification classes which correspond to the three flower species.

1. **Design and Implementation:**

* Below you can find the design and implementation for this project.

1. Import the dataset and print a sample of the iris dataset.

Table

Description automatically generated

1. Perform various appropriate plots (box plot, bar plot, count plot, distplot, clustermap, pairplot) to visually analyze the data. Record your observations i.e., using your plots, what can you understand about this data?

Chart, box and whisker chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

1. Preprocessing - split the data for training and testing i.e., 70(training):30(testing)

Graphical user interface, text, application

Description automatically generated

1. Scale the features i.e., perform normalization. From hereafter, you should be working on the scaled features.

Graphical user interface, text, application

Description automatically generated

1. Perform the training using KNN algorithm.

Graphical user interface, text, application

Description automatically generated

1. Performance evaluation. Print the table and briefly explain your understanding.

Table

Description automatically generated

1. Choose appropriate k value by comparing the error rate. Show appropriate plot/s.

Graphical user interface, text, application, email

Description automatically generated

Chart, line chart

Description automatically generated

1. Now choose only the first 2 attributes of the iris dataset i.e., sepal-length, sepal-width and repeat the steps 3 to 7.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

1. Compare the performance evaluation of the classifier with 4 attributes and classifier with 2 attributes. Explain your understanding.
   1. Based on the plots, the 4 attribute dataset proved to be more accurate than the 2 attribute dataset. This may be due to valuable information being removed on the second iteration without the last 2 attributes.

**Conclusion:**

In this project, I learned how to use pandas, seaborn, and pyplot. Even though I had prior experience with Pandas and Dataframes, I never used it in the application for data science and predicting values and related tasks.