Logo, company name

Description automatically generated

**CSCI 417**

**Divorce Prediction Project**

**Presented by**

**Mostafa Osama**

**Mostafa Mahmoud**

**Anthony Milad**

**Joseph Raafat**

**Presented to:**

**Dr. Ghada Ahmed**

**January 2022**

**Problem Statement**

Have you ever experienced a shattered heart or pondered what makes a relationship last? You might find this dataset useful. This dataset includes information on 170 couples and the Divorce Predictors Scale (DPS) factors based on Gottman couples’ therapy. The couples come from different regions, and the records were obtained through in-person interviews with married or divorced couples. On a scale of 0 to 4, the answers were tallied as follows: "Never," "Seldom," "Average," "Frequently," and "Always.". each attribute we have in the data is a question asked to the 170 couples and each row is one of the answered that we mention earlier. So, our project to detect the divorce based on the attributes we got in the dataset.

1. **Models Archietiture**
   1. **Logistic Regression**

Use logistic regression if you are trying to do binary classification and want to get the probability that the input data belongs to that class. Logistic regression is a simple and effective classifier for training models. It is a good candidate for predicting a discrete relationship, for example, predicting whether a tumor is malignant or benign given some measured properties.

* 1. **Decision Tree**

The main benefits of using a decision tree in machine learning is its simplicity, as the decision-making process is easy to visualize and understand. However, decision trees in machine learning can become overly complex by generating very granular branches

* 1. **KNN**

The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, you can use the KNN algorithm for applications that require high accuracy but that do not require a human-readable model. The quality of the predictions depends on the distance measure.

* 1. **SVM**

If you want to do linear classification but the input data is not linearly separable, use Kernelized support vector machines.

* 1. **Naïve bayes**

A Naive Bayes classifier can be a good choice if you have a relatively small dataset and want to do a fast classification.

* 1. **Random Forest**

the random forest classifier can be used to solve for regression or classification problems. The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample.

1. **Evaluation Strategies** 
   1. **F1 Score**

The F1 score can be interpreted as a harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is: F1 = 2 \* (precision \* recall) / (precision + recall)

* 1. **Confusion Matrix**

A confusion matrix is such that C i, j is equal to the number of observations known to be in group and predicted to be in group. Thus, in binary classification, the count of true negatives is C 0, 0, false negatives are C 1, 0, true positives are C 1, 1 and false positives is C 0, 1.

* 1. **Accuracy Score**

The accuracy score method of the sklearn. metrics package assigns subset accuracy in multi-label classification. It is required that the labels the model has predicted for the given sample and the true labels of the sample match exactly. Accuracy describes the model's behavior across all classes.

**Outputs**

**SVM Accuracy**

****

**Decision Tree Classifier Accuracy**

**Background pattern

Description automatically generated**

**Random Forest Classifier Accuracy**

**Background pattern

Description automatically generated with medium confidence**

**KNN Accuracy**

**Background pattern

Description automatically generated with medium confidence**

**Naïve Bayes Accuracy**

**Background pattern

Description automatically generated**

**Logistic Regression Accuracy**

**Graphical user interface, application

Description automatically generated**