# **MACHINE LEARNING AND DEEP LEARNING**

**Model 1 - Decision Tree**

**Data**

The data is a diabetes dataset that has several input features and predicts if the person or the patient is diabetic or not . The input features are:

* Pregnancies : number of pregnancies had.
* Glucose level: blood glucose level
* blood pressure: normal blood pressure of the person
* skin thickness: How thick the skin is of a person in millimeters.
* Insulin : insulin production/ insulin in blood
* BMI : body mass index
* Diabetes Pedigree Function: calculates the likelihood of the person having diabetes.
* Age of the person.

Since Decision Tree is a supervised learning model, the data has 1 output variable that the Decision Tree model classifies the predicted output into one of the two possibilities.

* Diabetic - 1
* Non-diabetic - 0

**Why decision Tree?**

**Task: Classification Task**

**Modeling Technique: Predictive Analytics**

The data has an output variable hence only supervised learning algorithms can be applied. This is a classification task where we have to classify if they are diabetic or not hence regression cannot be used as the output variable is not numeric but a categorical variable. This is a class classification problem. Out of the many supervised Learning algorithms hence Decision Tree works well for classification tasks. Since Decision Trees can handle categorical variables very well and is efficient for smaller dataset hence this makes it the best possible algorithm in this case.

**Model 2 – K-Means Clustering**

**Data**

This dataset is a driver dataset. The input features are:

* Mean\_dist\_day
* mean\_over\_speed\_perc

**Why K-Means Clustering?**

**Task: Clustering**

**Modeling Technique: Descriptive Analytics**

The dataset has no output variables hence unsupervised learning algorithms are used. Since all the input features are numeric, using the K-means algorithm makes more sense. Since we have no hierarchy in the data, using hierarchical clustering doesn’t make sense. DB Scan can also be used to predict the clusters but since the dataset is small K-Means works efficiently and saves time.