K NEAREST NEIGHBOUR

- It is a predictive model.
- KNN works on supervised data where, class variable can be discrete or continuous in nature.
- It is based on "similarity" between data points which is measured using "distance metric" such as, Euclidean and Hamming.
- It is non-parametric model. It means that it does not make any assumptions on the underlying data distribution.

PROXIMITY MEASURES I

- Proximity measures are very useful concept in Machine Learning that tells how **similar or dissimilar two data points (or data objects) are.**
- In order to understand proximity measures, consider Figure 71. It is a plot showing relationship between two features namely, Marks1 and Marks2 of 8 students represented by data points P1, P2, P3, P4, P5 , P6, P7 and P8.

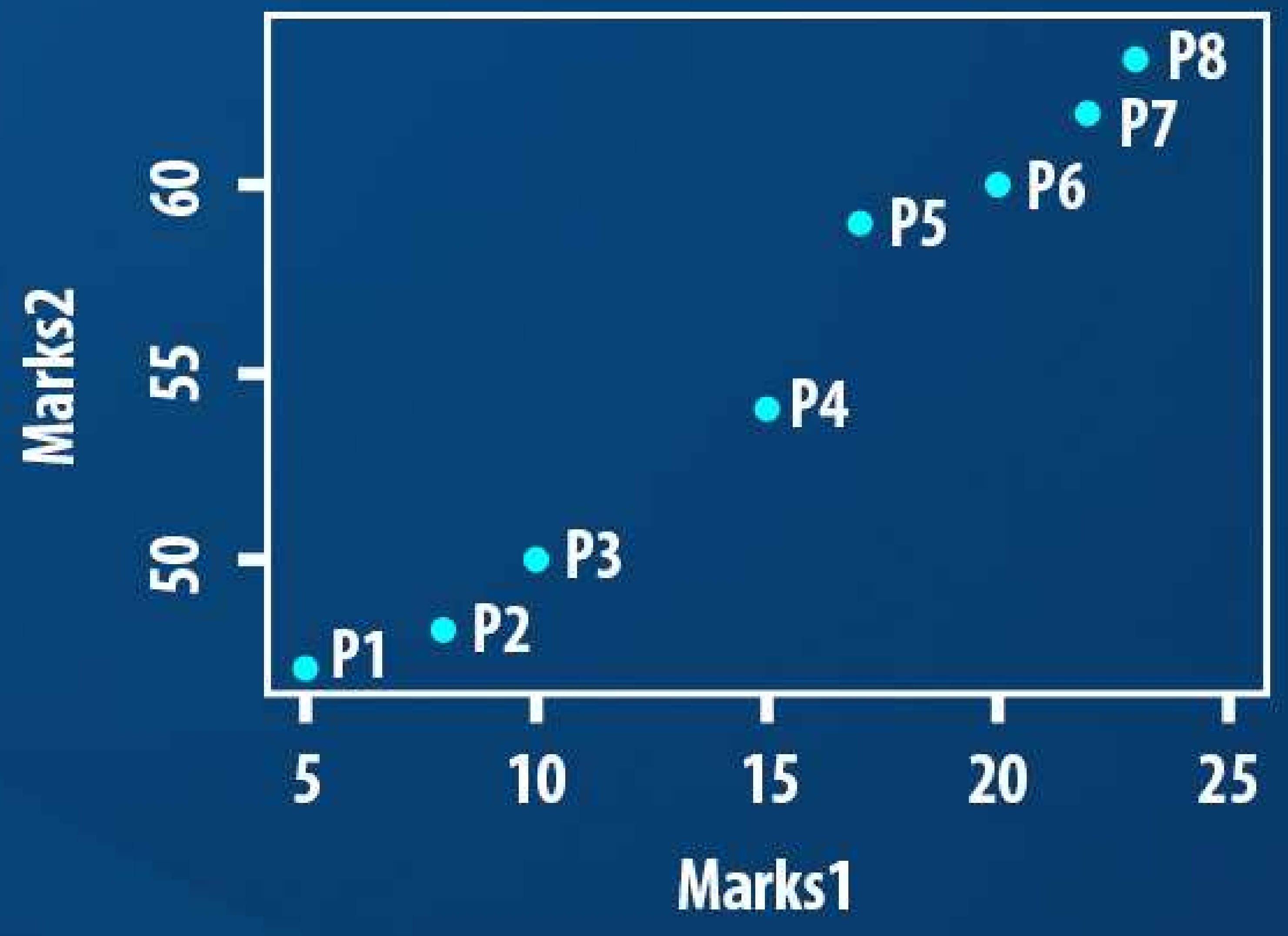


Figure 71: 2D data set showing relationship between features Marks1 and Marks2 of 8 observations

- Proximity measures can address to following questions?

1. How far student P_1 is from P_6 ? More formally, what is the distance between P_1 and P_6?
2. How much similar are students namely, P_3, P_4 and P_7 ? What is the similarity between students?

POPULAR PROXIMITY MEASURES I

The two popular methods to calculate proximity between two data points are described below. These methods are called as "distances metrics"

1. Euclidean Distance
2. Hamming Distance

1. Euclidean Distance:

   It is particularly used to calculate the distance between two data points. The larger distance between data points represents "dissimilarity" whereas, a small distance indicates "similarity" between observations. Euclidean distance is between two points P_1 and P_2 is calculated using Equation 38.

$$\text{dist}(P_1, P_2) = \sqrt{\sum_{k=1}^{n} (P_{1_k}, P_{2_k})^2}$$

Where, n represents the number of features in the data set and P_{1_k}, P_{2_k} indicates k^{th} value of the attribute. Euclidean distance can only be calculated for numerical attributes.

POPULAR PROXIMITY MEASURES II

Example Illustration:

In Table 13, data set of 3 data points with two features is presented.

Data Points	X_1	X_2
P_1	0	1
P_2	2	1
P_3	2	2

$$\Delta(P_1,P_2)=\sqrt{(0-2)^2+(1-1)^2}$$

Table 13: Hypothetical data set with 3 observations and 2 features

Using this data set and Equation 38, we calculate distance between pair of data points. These distances are represented in Table 14.

	P_1	P_2	P_3
P_1	0	2	2.2
P_2	2	0	1
P_3	2.2	1	0

Table 14: Euclidean distance between each pair of data points

- Data point P_1 is closer to data point P_2 than P_3.
- Proximity of P_2 is more with P_3 than P_1.
- Data point P_3 is closer to P_2 than P_1.

Euclidean distance cannot be used for categorical data values.

POPULAR PROXIMITY MEASURES III

Hamming Distance

- It is also used to find "dissimilarity" and "similarity" between two data points but only when attributes are categorical or string in nature. If the attribute value of for two data points matches, hamming distance is given as 0 otherwise it is 1.
- Hamming distance between two data points P_1 and P_2 is calculated using Equation 39.

$$dist(P_1, P_2) = \sum_{k=1}^{n} [\, P_{1_k} = P_{2_k} 0,\ P_{1_k} \neq P_{2_k} 1 \,]$$

Example Illustration

Consider Table 15 where information on three species P_1 ,P_2 and P_3 is presented with attributes namely, Size, Type and Color. For this data set, suppose we want to find out distance between P_1 ,P_2 and P_1 ,P_3.

Data Points	Size	Type	Color
P_1	big	cat	black
P_2	small	cat	white
P_3	big	dog	black

Table 15: Hypothetical data set with 3 observations and 3 features.

dist $(P_1,P_2) = 2$

dist $(P_1,P_3) = 1$

Here data point P_1 is more closer to P_3 than P_2.

CONCEPT OF K NEAREST NEIGHBOUR I

Let us understand concept of KNN using a simple example.

Data Points	X	Y
P_1	1.5	2.2
P_2	1	1
P_3	2	1
P_4	1.5	2.0
P_5	2	2
P_6	2.5	3.0
P_7	2.7	2.5
P_8	1.5	1.5

Figure 72: Hypothetical data set with 8 observations and 2 features

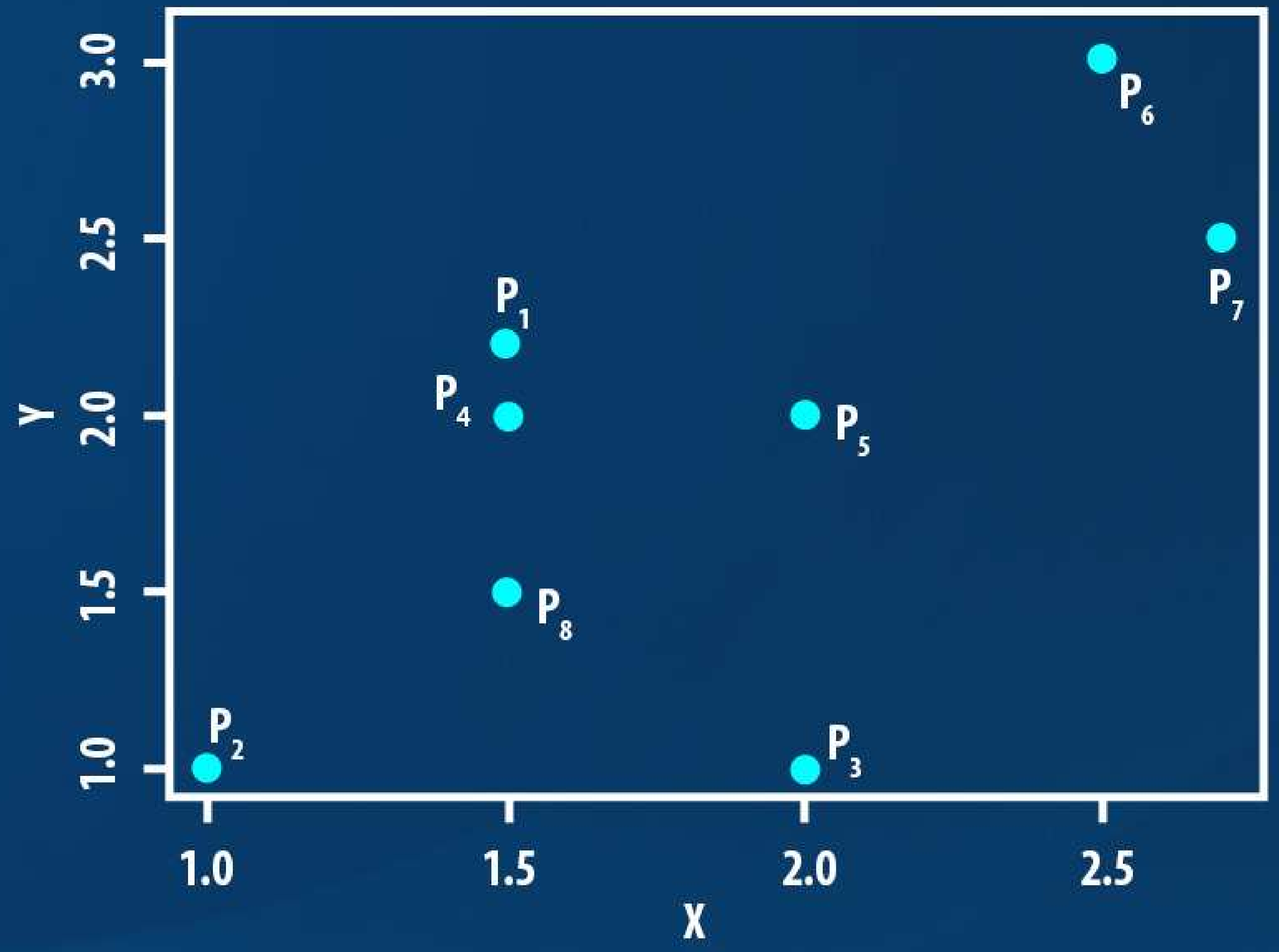
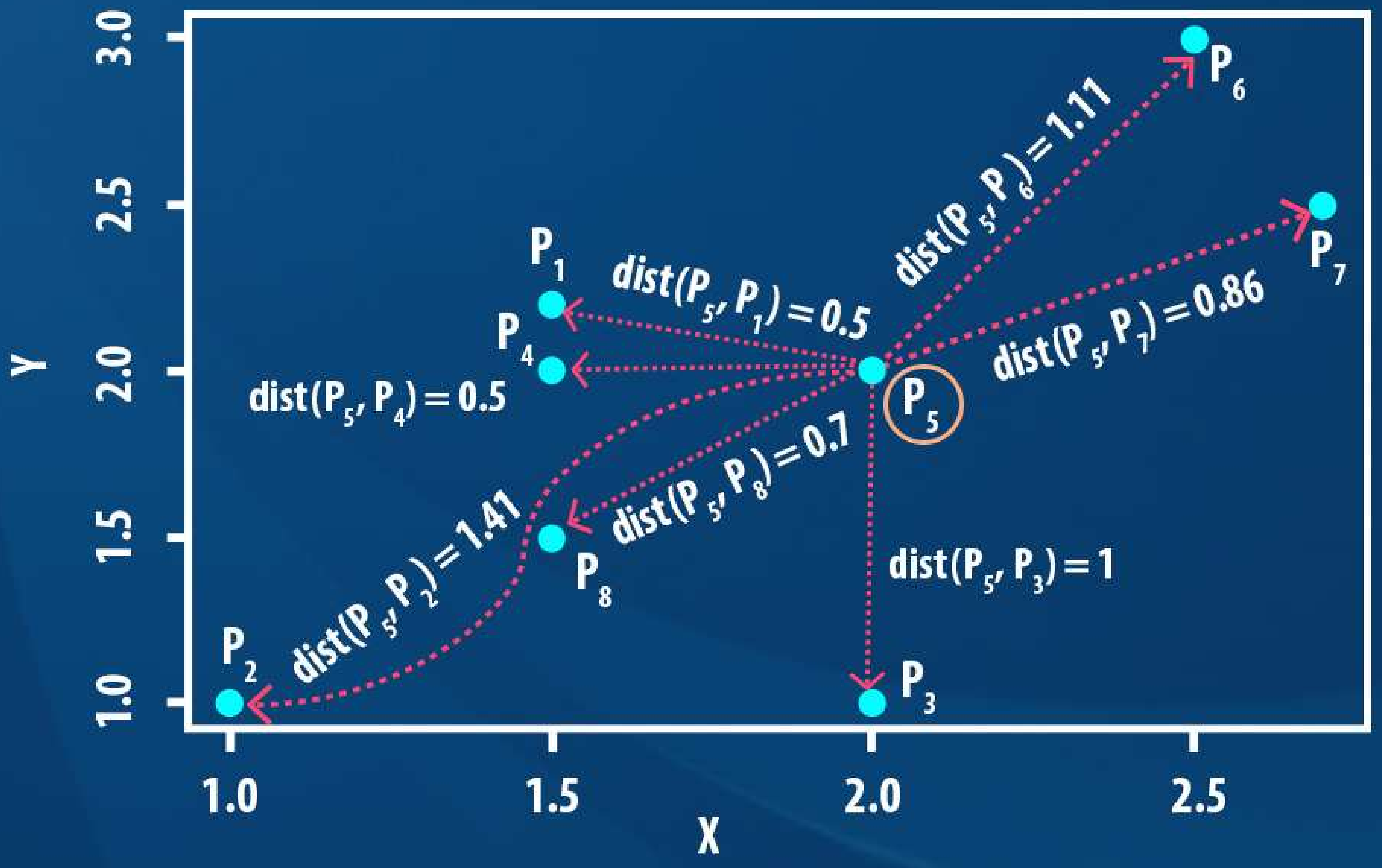


Figure 73: 2D plot of data points in table above

CONCEPT OF K NEAREST NEIGHBOUR II

Concept of K nearest Neighbour

We represent distance of data point P_5 from all other data points in Figure below. Where the distance is measured using Euclidean metric.



1. **Which are the two nearest (closest) data points from P_5 ?**

   Ans: P_1 and P_4 (here $k = 2$)

2. **Which are the four nearest (closest) data points from P_5 ?**

   Ans: P_1, P_4, P_7 and P_8 (here $k = 4$)

K nearest neighbour stands for number of k closest data points.

KNN FOR CLASSIFICATION I

The figure below shows the class wise distribution of data points P_1, P_2, P_3, P_4, P_5, P_6, P_7 and P_8. Let there be two classes namely C_1 (blue) and C_2(green). Each data point in the figure is either representing C_1 class or C_2 class. Consider this data set with 2 classes and 8 data points as the training set for KNN classifier.

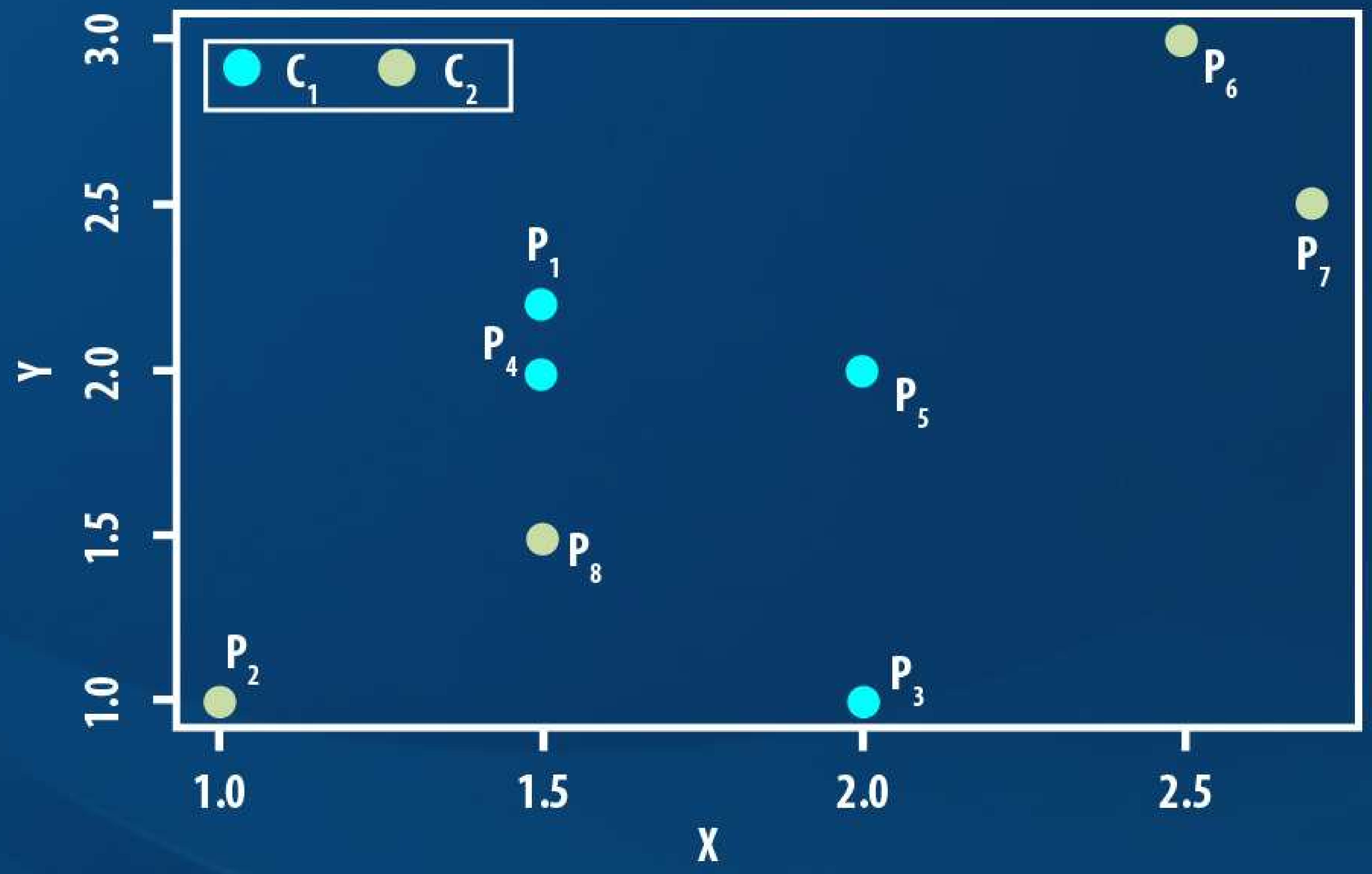


Figure 73: 2D plot of data points in table above

KNN FOR CLASSIFICATION II

Assume a new data point $T_1 = (1.8, 2.7)$ appears for which class has to be predicted based on the training data set represented in Figure below.

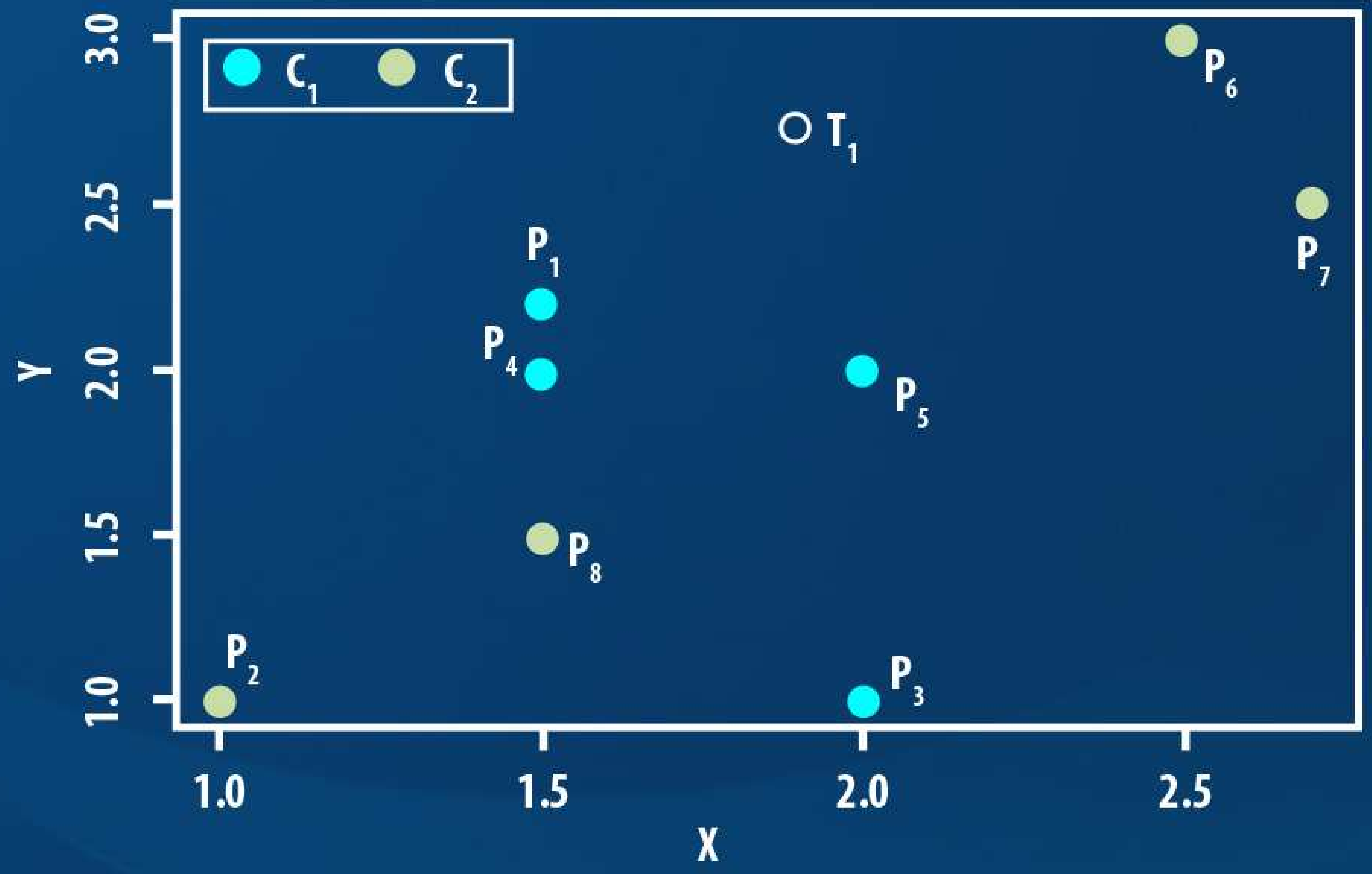


Figure 73: 2D plot of data points in table above

KNN FOR CLASSIFICATION III

In order to apply KNN, distance score of T_1 will be computed with every data point in the training set using Euclidean distance.

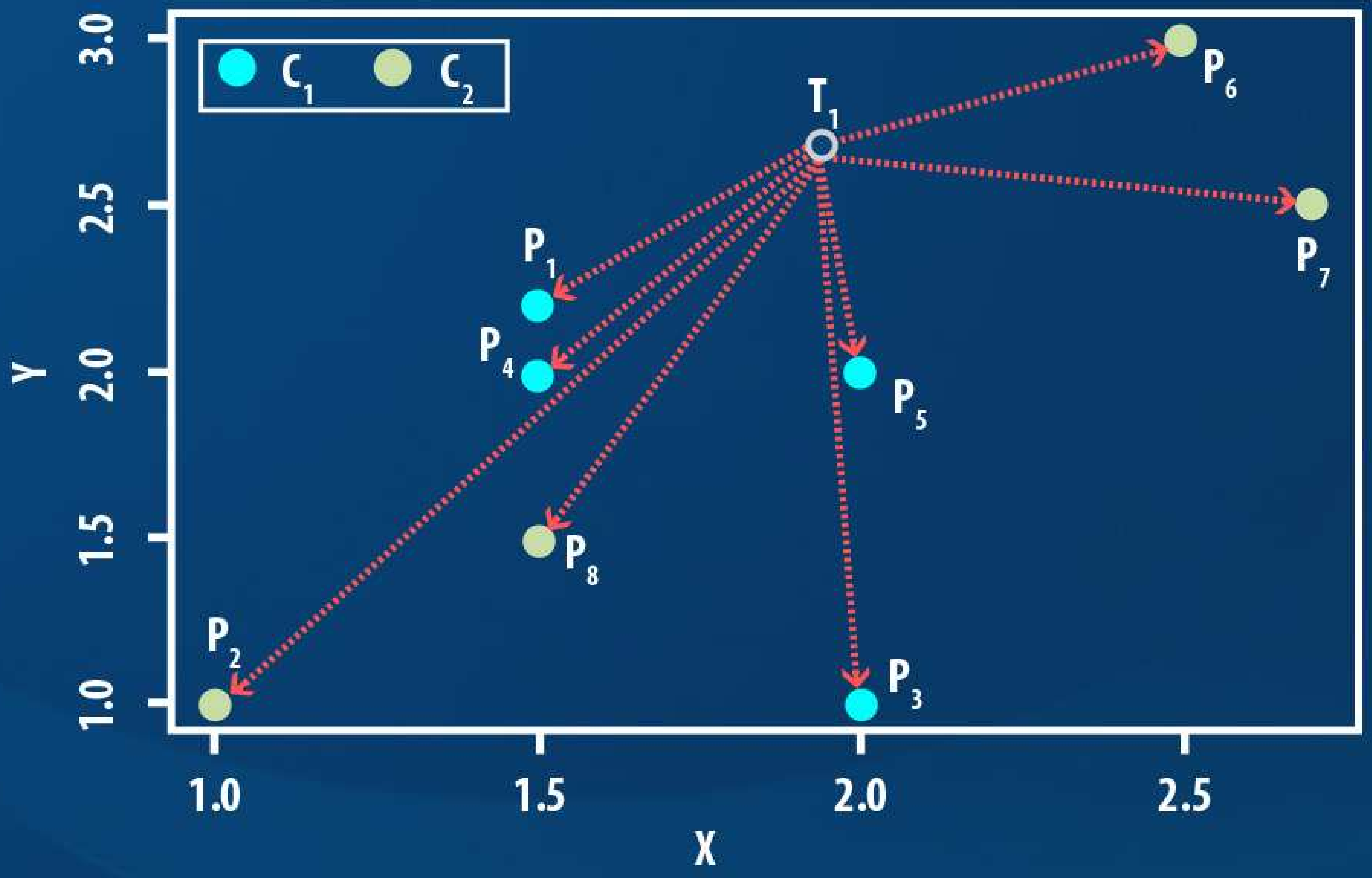


Figure 73: 2D plot of data points in table above

KNN FOR CLASSIFICATION IV

Based on the input parameter *k*, top *k* nearest neighbours of T_1 are identified. Let suppose, $k = 5$. The 5 nearest neighbour of T_1 are highlighted in black arrow in figure below.

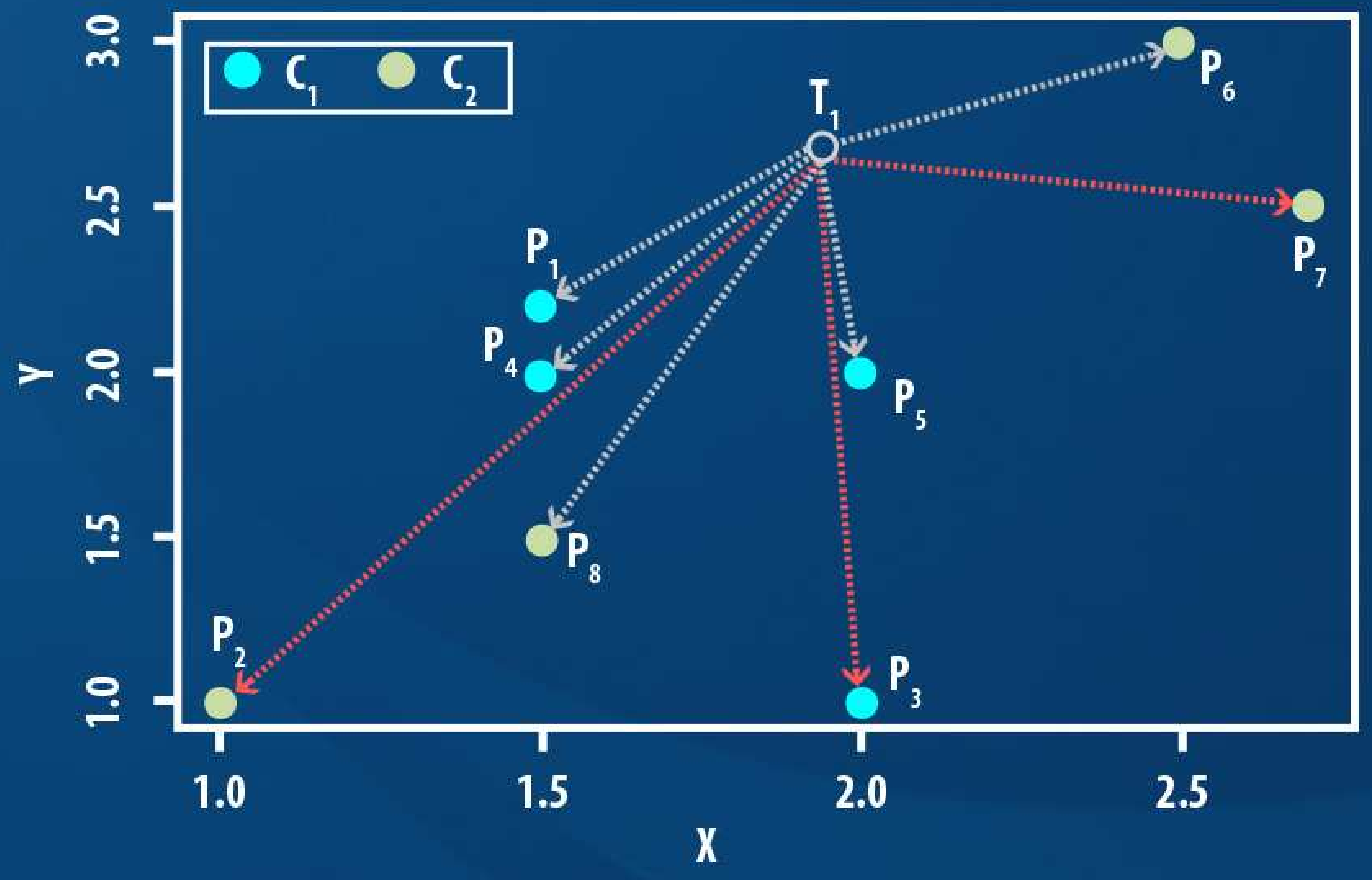


Figure 73: 2D plot of data points in table above

Out of the 5 nearest neighbours of T_1

1. 3 data points belong to C_1 (blue class)
2. 2 data points belong to C_2 (green class)

Since C_1 has majority of closeness with T_1 so, T_1 is also classified as C_1

HOW TO CHOOSE THE VALUE *K* ?

Choosing the right value of k is important for best result. However, it highly depends on the size of data set. There are mainly two approaches to choose the value of *k* The value of *k*

1. Try different k values and use Cross-Validation to see which K value is giving the best result

2. Choose $k =$ sqrt (N). Where N is the number of samples in the data set

KNN FOR PREDICTION (SUMMARY)

1. KNN is a very flexible model. It can fit to almost all applications.
2. Very simple technique. Easy to explain and understand/interpret.
3. It is a white box model. Meaning that KNN not only predicts but KNN can also explain the reason to its prediction.
4. Most of the time it gives high accuracy.
5. It is versatile in nature,i.e., it useful for classification or regression.
6. This algorithm is computationally expensive for the reason that distance between every pair of data points is computed.
7. The performance of KNN may degrade when data set has attributes of different scale. It is a good practice to scale/normalise data set before applying KNN.