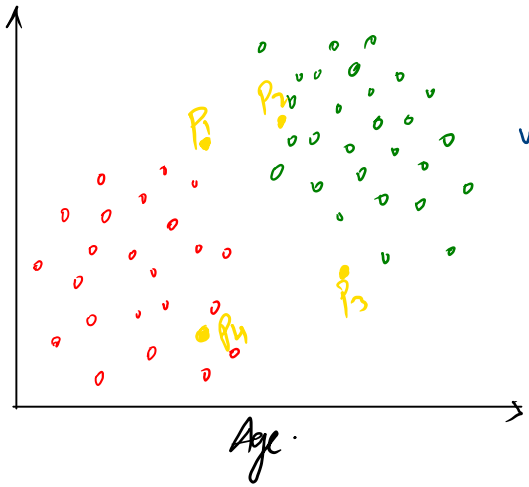
KNN - Model :-

KNN or K-Nearest Neighbours is a supervised machine learning model which is used for both the regression and classification problems.

We are going to understand the classification first for the KNN model.

Let us suppose that we have two independent variables salary and age, and, we have to to predict whether the person will default or not. So, let us plot the training data for this problem statement.



The yellow points are the points that we want to predict.

In the KNN model, we first select the first point to predict the fraud.

For P_1, first we select the k nearest neighbours, that we want to take for the prediction. we can select any numbers of neighbors that we want, for this model, we will select 5. Now we check that the majority of the neighbors are of which class. If the majority is of class 1, then we give the predicted value as 1 or vice-versa.

Similar to the classification, we take the value of k nearest neighbors and then we take their mean in case of a regression.

Now, it is very important that we find the correct or optimal value of k. The k is calculated by performing some hit and trial or experimentation.

The nearest neighbors can be found using different types of methods which can be as follows:

- → Manhattan Distance
- → Euclidian Distance
- → Minkowiski Distance
- → Hamming Distance

& many other methods.

Now, the main thing to remember when we use the KNN model is that it is highly dependent on the the distances. So, there are some factors which can effect the performance of this model. Lets suppose that we have two variables in our data, salary and age. Salary is in lakhs whereas age is in tens. So, if we build our model using this, then the k-neighbors will be affected largely and the model will not perform correctly.

So, before using the KNN model, it is very important that we check the scale of the data and perform proper scaling.

For the scaling of the data, we can use the different types of scalers depending upon our data.

Now, in case of KNN regression, we can take the value of k to be both even and odd as it will not affect the prediction. But now, let us suppose that we have a classification problem. In this case, we have two variables that we want to predict 0, 1. In this case, if we take the value of k to be even, for example 4 or 6, then it is possible that 3 neighbors are positive and other 3 are negative. So, in this case the prediction will be hampered as the value to be predicted have equal neighbors to choose from. So, it is better or best option that in case of a classification problem when we want to use knn, we always use value of k to be odd.

Underfitting and Overfitting in KNN:-

Overfitting means when any machine learning model or algorithm is learning all the details and information form the given data. In this case, the model also learns the noise and errors From the data. In such situation where overfitting is happening, the model will perform very good on the Traning data, but in case of testing data, it will perform very poorly, beating the main motive of machine learning.

Unlike overfitting, Underfitting occurs when the model is unable to learn all the details and information provided by the dataset. It happens when model has not trained for enough time or haven't learned enough data to get the relationship between the input or independent variables and the dependent variable. In this case, both performance while training and testing are poor.

For the knn model, overfitting occurs when the value of k is too small. In this case, model will try to learn as much information as possible as number of neighbors are less and it will result to overfitting.
When the value of k is too large, then the model will be unable to learn much data as it will identify everything as similar and thus will result to underfitting.

Choosing the correct K :-
In knn model, k is a parameter on which the learning is dependent. We can select the value of k by doing proper hyperparameter tuning and comparing the correct accuracy measure depending upon the problem statement.

Advantages and Disadvantages :-

It is easy, good on small datasets, good as a benchmarking model.
Not good for large datasets, when we have high dimension we need to do feature selection; Value of k effects the model performance.