Statistical Machine Learning Approaches to Liver Disease Prediction

Literature Survey;

Naive Bayes:

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve higher accuracy levels.Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum- likelihood training can be done by evaluating a closed- form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

SVM or Support Vector Machine:

SVM algorithm is a simple yet powerful Supervised Machine Learning algorithm that can be used for building both regression and classification models. SVM algorithm can perform really well with both linearly separable and non-linearly separable datasets. Even with a limited amount of data, the support vector machine algorithm does not fail to show its magic.The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. In short, the hyperplane is (n-1)-D plane for n features.

K-NEAREST NEIGHBOUR (KNN):

- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

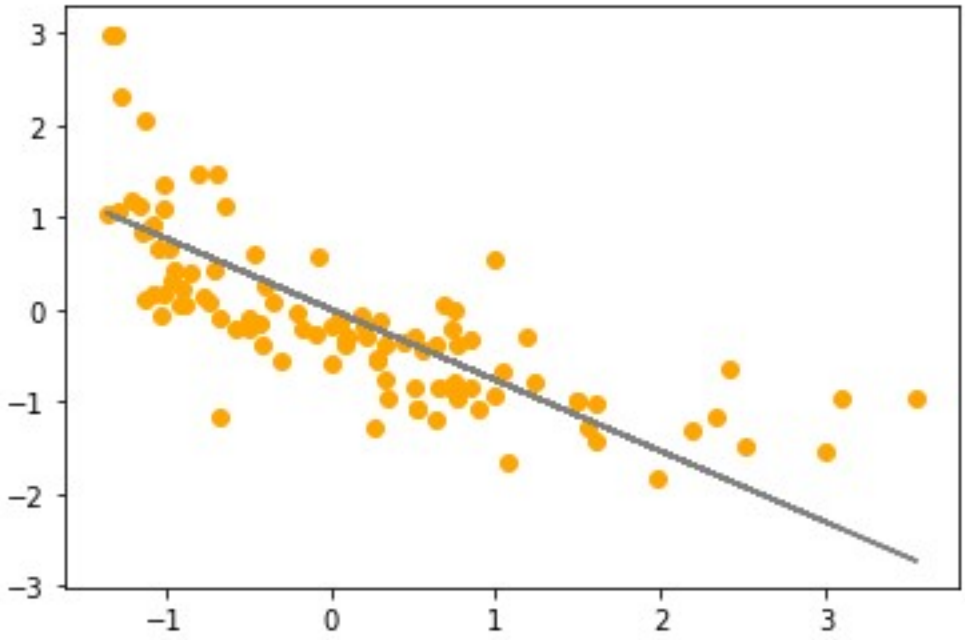
The K-NN working can be explained on the basis of the below algorithm:

- Step-1: Select the number K of the neighbours
- Step-2: Calculate the Euclidean distance of K number of neighbours
- Step-3: Take the K nearest neighbours as per the calculated Euclidean distance.
- Step-4: Among these k neighbours, count the number of the data points in each category.
- Step-5: Assign the new data points to that category for which the number of the neighbour is maximum.
- Step-6: Our model is ready.

LOGISTIC REGRESSION

Regression models enable you to predict the relationship between a dependent and independent variable. These models are at the root of many machine learning analyses and can be used to predict customer behaviour, model events over time, and determine causal relationships between events or behaviours.

There are multiple types of regression models that you can use but linear and logistic regression are the most common. Linear regression attempts to find the best fit line between a dependent variable (often customer behaviour or preference) and one or more independent variables. For example, how much a customer might like a specific product based on previous purchases. It is useful for determining correlations but not causal relationships.



Logistic regression attempts to find the probability that an event occurs or doesn’t occur. For example, whether a customer makes a purchase or doesn’t. It is useful when there is no linear relationship between variables but requires large sample sizes to produce accurate results.