Project Report

Title: Determination of Carcinogenic Patients Through Data Mining

Name (Reg.No) : Naga Harshith Bezawada 19BCE1547

Abstract:

Cancer is a disease that claims hundreds of thousands of lives every year, and most people know at least one person who suffers from cancer. In fact, it's such a widespread affliction that it's the second-highest leading cause of death in the US, right behind heart disease. So, it should come as no surprise that many of those in the medical field have made it their primary focus for research.

One of the biggest areas of research when it comes to cancer is prediction, because the sooner cancer is detected, the higher the chance that the patient will survive. In most cases, cancer is diagnosed after it has already entered into the advanced stages, which severely decreases the effectiveness of current cancer treatments. Trying to detect cancer in its earlier stages, or even before it happens, gives us a better treatment outcome for the patients.

In this project, we develop a predictive model using two supervised learning algorithms, namely KNN and Logistic Regression, to train and classify whether a patient has cancer or not.

1. Introduction:

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It's overall goal is to extract information from a data set and transform the information into a comprehensible structure for further use.In this project, we will be using a data set that contains diagnostic reports of patients with labels M/B denoting Malignant or Benign cells. The goal of this project is to develop a fairly accurate predictive model that determines if a patient is carcinogenic or not based on the parameters provided through two supervised learning algorithms and to compare the resulting accuracy of the models.

3. Implementation:

Classification, which is a data mining function that assigns items in a collection to target categories or classes, will be used to make a predictive model in this project. The goal of classification is to accurately predict the target class for each case in the data. Here, our goal is to predict if a patient has cancer or otherwise.

Algorithm being applied

1. K-Nearest Neighbours: (refer GITHUB for detailed explanation of Code)

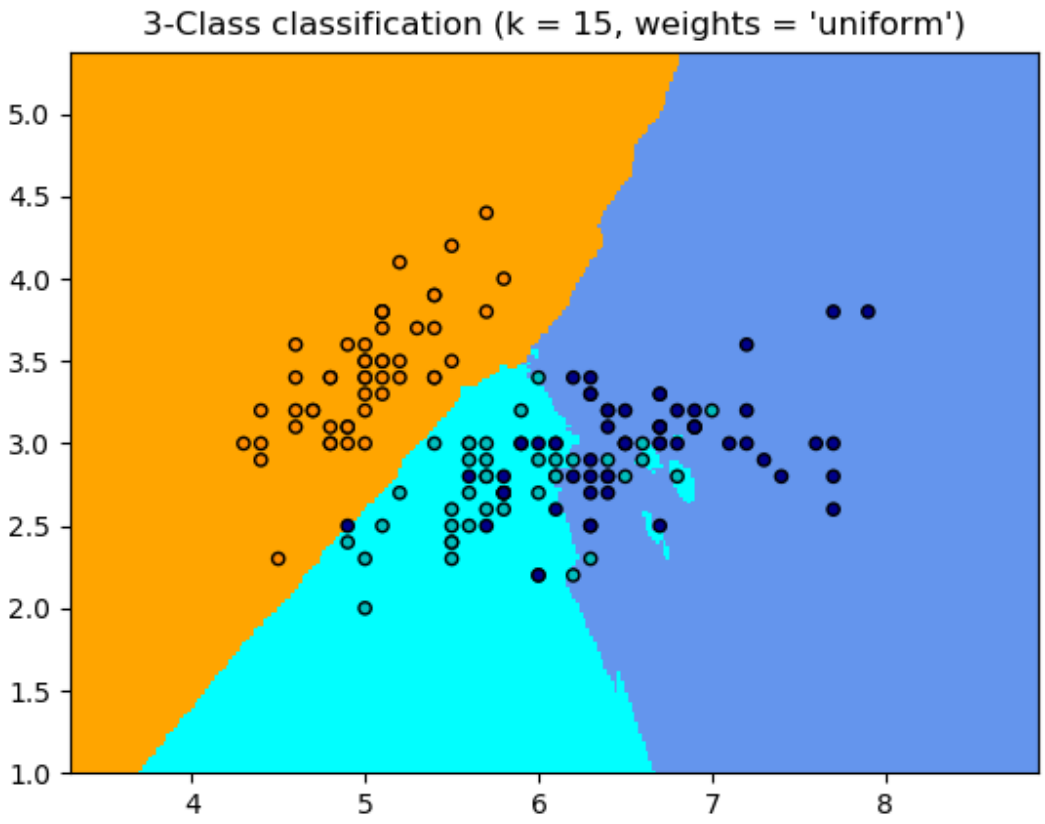
This algorithm considers all samples including the new sample/data as points on a graph and finds the 'k' nearest points to the new point using euclidean distance or manhattan distance.

Euclidean distance is the length of the line segment connecting two points, it is used when all the features are of the same type.

In Cartesian coordinates, if p = (p_1, p_2,..., p_n) and q = (q_1, q_2,..., q_n) are two points in Euclidean space, then the distance (d) from p to q, or from q to p is given by the Pythagorean formula,

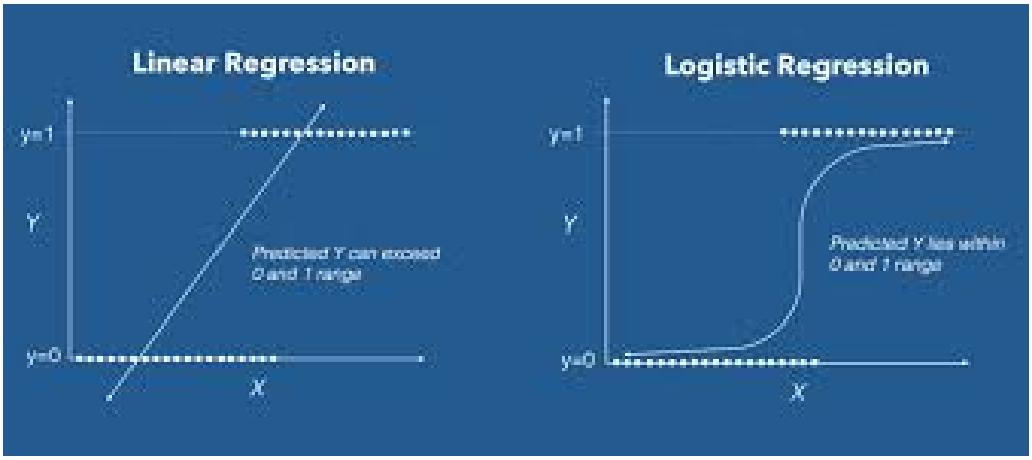
$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$



2. Logistic Regression (not yet implemented) :

It is based on a supervised classification algorithm. In a classification problem, the target variable(or output) Y, can take only discrete values for a given set of features(or inputs) X. Logistic regression builds a regression model to predict the probability that a given data entry belongs to the category numbered "1" or "0". In Logistic regression models data follows sigmoid function.



4. Results and Discussion:

Based on conclusive evidence we can say that through K Nearest Neighbours the accuracy is variable as we use a random test set and training set randomly but can go to an accuracy of 97.8% and an average accuracy of 96.8%.. Thereby proving to be more efficient than logistic regression.

5. Conclusion and Future Work

Increasing the accuracy of KNN by using Normalisation, reliability and clustering techniques.
In the long run, prototyping a practitioner-friendly user interface where data obtained from processing imagery is immediately sent through the ML model and a result is predicted. I.E., a full fledged software solution to our underlying problem that's fairly accurate and feasible for use in the field of medical technology.

References

1. Procedia Computer Science (2019), Application of data mining techniques to predict breast cancer
2. This was written in reference with : Dursun Delen, Glenn Walker, Amit Kadam(2019) Predicting breast cancer survivability a comparison of three data mining methods
3. Walid cherif (2019), Optimisation Of KNN algorithm by clustering and reliability coefficients : application to breast - cancer diagnosis