**Paper Title:** Diabetes prediction using supervised machine learning.

**Publisher:** ELSEVIER

**About:**

* Diabetesis a disease that can lead to blindness, kidney failure, and heart attacks, as well as death.
* According to the International Diabetes Federation, there were 463 million diabetics in 2019.
* If predictions are correct, this number will rise by 578 million by 2030, reaching 700 million by 2045.
* According to an article published by the Ministry of Health of the Republic of Indonesia in 2020, the ten countries with the highest diabetes rates in 2019 include Indonesia.
* The ability of experts is required to determine the type of diabetes disease, because of their delay in discovering what disease they have, many people who are examined have a disease that can be described as severe.
* Diabetes detection technology is required to prevent severe conditions, in today's medical world, doctors can use it to quickly and accurately interpret diseases.
* We can use machine learning to prevent the death by making an artificial inteligent model that can predict diabetes disease and the method that be used is comparison between the KNN and Naive Bayes algorithms to see which algorithm suit the best for diabetes prediction.
* The study concluded by comparing two k-Nearest Neighbor algorithms and the Naive Bayes algorithm to predict diabetes based on several health attributes in the dataset using supervised machine learning.
* The object of this study is the Pima Indians Diabetes dataset, which contains 8 independent variables and 1 dependent variable.
* In this study, eight factors that are characteristic of diabetes are used to determine whether a person has diabetes or not, namely pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age.
* With so many variables to consider in the diagnosis of diabetes, a method that produces results quickly and efficiently is required.
  + To compare the classification accuracy of various machine learning algorithms if similar data is purchased.
  + To learn how some machine learning algorithms classify.
  + To categorize diabetes using a variety of machine learning algorithms
  + Significance of the Study
  + Contribute ideas for researchers and those interested in using machine learning to classify other diseases.
  + To make a scientific contribution to the field of health sciences by raising readers' awareness of the characteristics of diabetes. This article can be used as a resource for machine learning algorithms.

**Proposed model classification:**

* In this problem, the authors use two algorithm for comparing, K-Nearest Neighbour and Naïve Bayes. K-nearest Neighbor is one of the simplest Machine Learning Algorithm based on Supervised Learning techniques.
* The K-NN algorithm assumes the similarity between new cases/data with available cases and puts new cases into the category that is most similar to the available categories.
* The K-NN algorithm stores all available data and classifies new data points based on similarity. This means that when new data appears it can be easily classified into the well suite category using the K-NN algorithm.
* Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. This technique is easiest to understand when it is explained using binary or categorical input values. Called naive Bayes or idiot Bayes because the calculation of the probability for each hypothesis is simplified to make the calculation workable. This is a very strong assumption that is most unlikely in real data, namely that the attributes do not interact.
  + K-Nearest Neighbor: The KNN algorithm is one of the supervised learning techniques, it will calculate the proximity between the old case (training data) and new cases (testing data). The steps of the KNN algorithm are:
    - Determine the parameter k, as nearest neighbors (number of nearest neighbors);
    - Calculate the square of the distance between the new data and the training data using Eucliden.
    - Sort the closest neighbors that have the minimum distance.
    - Classify the nearest neighbor according to the value of k.
    - Determine the classification results based on the majority label.
  + Naïve Bayes: The Naive Bayes classifier is a simple probabilistic classifier based on the application of the Bayesian theorem (from Bayesian statistics) with the assumption of strong (naive) self-determination. Naïve Bayes can also be called Simple Bayes and Independence Bayes. This algorithm can predict the probability of class membership, such as the probability of data given a certain class label. The Naive Bayes classifier assumes that the presence (or absence) of certain features (attributes) of a class is not related to the presence (or absence) of other features when a class variable is given. The following is the Naïve Bayes equation:



The advantages of using the Naïve Bayes algorithm are:

* + Fast and highly scalable model
  + Balancing linear with number of predictors and rows.
  + The Naïve Bayes procedure is parallel.
  + Naïve Bayes can be used for binary and multiclass classification.

**Conclusion:** The study concluded by comparing two k-Nearest Neighbor algorithms and the Naive Bayes algorithm to predict diabetes based on several health attributes in the dataset using supervised machine learning. According to the results of our experiments and evaluating alghorithm using Confusion Matrix, the Naive Bayes algorithm outperforms KNN, with an average value of 76.07 percent accuracy, 73.37 percent precision, and 71.37 percent recall and in Naive Bayes and an average value of 73.33 percent accuracy, precision 70.25 percent, and recall of 69.37 percent in KNN. As a result, it can be concluded that the Naive Bayes algorithm is preferable to the KNN algorithm for predicting diabetes using the Pima Indians dataset.

**Future Research**: Can be done by adding other algorithm like neural network and other techniques in order to produce an accuracy value and better precision also by Adding technique Particle Swarm Optimization for optimize the results and using application program development.