Abstract: -

The increasing ratio of diabetes is found risky across the planet. Therefore, the diagnosis is important in population with extreme risk of diabetes. In this study, a decision-making classifier (J48) is applied over a data-mining platform (Weka) to measure accuracy and linear regression on classification results to forecast cost/benefit ratio in diabetes mellitus patients along with prevalence.

Introduction: -

The data mining is a modern technique of digging into data to locate patterns ideally. It is an astounded spectator of incredible progress in modern health care[1](https://www.nature.com/articles/s41598-019-46631-9#ref-CR1). The undiscovered electronic health records (EHRs) are being used to manipulate the unknown patterns located in the health datasets and healthcare industries, which are enthusiastically supporting the copious amount of data.

We will use Naive Bayes model on the \*\***Prima Indians Diabetes**\*\* data set. The model will predict which people are likely to develop diabetes. All patients here are females at least 21 years old of Pima Indian heritage. The datasets consists of several medical predictor variables.

Dataset:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset.

Data Flow: -

Chart, line chart

Description automatically generated

Methods: -

the assessment framework used in this work for screening and cost benefit implications (performed in six steps) in diabetes patients. Initially, the real-life diabetes mellitus data was acquired and preprocessed afterward; the data was used for evaluation and assessment.

Diagram

Description automatically generated

Dataset Description:-

Several constraints were placed on the selection of these instances from a larger database. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

* Pregnancies: Number of times pregnant
* Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
* Blood Pressure: Diastolic blood pressure (mm Hg)
* Skin Thickness: Triceps skin fold thickness (mm)
* Insulin: 2-Hour serum insulin (mu U/ml)
* BMI: Body mass index (weight in kg/(height in m)^2)
* Diabetes Pedigree Function: Diabetes pedigree function
* Age: Age (years)
* Outcome: Class variable (0 or 1)

Decision tree: -

It is a supervised learning method, which is used for solving classification problems. Decision tree [[37](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6#ref-CR37), [38](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6#ref-CR38)] is a technique which iteratively breaks the given dataset into two or more sample data. The goal of the method is to predict the class value of the target variable. The decision tree will help to segregate the data set and builds the decision model to predict the unknown class labels. A decision tree can be constructed to both binary and continuous variables. Decision tree optimally finds the root node based upon the highest entropy value. This gives decision tree an advantage of choosing the most consistent hypothesis among the datasets.

Diagram

Description automatically generated

### Naïve Bayesian: -

A classification algorithm [[40](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6#ref-CR40), [41](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6#ref-CR41)], a probabilistic classifier which is based on Bayes theorem with the independence assumption between the predictors. Naïve Bayesian method takes the dataset as input, performs analysis and predicts the class label using Bayes’ Theorem. It calculates a probability of class in input data and helps to predict the class of the unknown data sample.It is a powerful classification technique suitable for large datasets. The Bayes Theorem formula calculates the posterior probability for each class using below formula. The Flowchart for Naïve Bayesian.

P(c|x)=P(x|c)P(c)P(x)P(c|x)=P(x|c)P(c)P(x)

P(c|X)=P(x1|c)×P(x2|c)×⋯×P(xn|c)×P(c)

Diagram

Description automatically generatedSteps Involved: -

1. Importing the libraries

2. Data Import

3. Data Pre-processing

4. Study Correlation

5. Train and test data split

6. Build model

7. Make predictions

Machine Learning Model used:

1.Decision Tree

2. Naive Bayes

Libraries Involved:

1. NumPy

2. pandas

3. matplotlib

4. sklearn

5. seaborn

Visualizations:

1. histogram

2. correlation

3. pair plot

4. confusion matrix

References: -

1. Global Report on Diabetes 2016 by World Health Organisation. <http://www.who.int/diabetes/publications/grd-2016/en/>, ISBN 978 92 4 156525 7.
2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018 American Diabetes Association Diabetes Care 2018; 41(Supplement 1): S13–S27. <https://doi.org/10.2337/dc18-S002>.
3. Alberti KG, Zimmet PZ. “Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet Med. 1998;15(7):539–53.