**A**

**MINI PROJECT**

**REPORT ON**

**Predicting Diabetes Using Pima Indian Diabetes Dataset**

**SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE.**

**FOR**

**LAB PRACTICE II**

**Data Mining and Ware-Housing**

**BACHELOR OF ENGINEERING (COMPUTER ENGINEERING)**

**SUBMITTED BY**

**Name: Prajakta kamble PRN – 72000288F**

**Name: Rutuja Bachhav PRN – 72000282G**

**Name: Sayma Kazi PRN – 7200089D**

****

**DEPARTMENT OF COMPUTER ENGINEERING**

**D.Y.PATIL COLLEGE OF ENGINEERING AKURDI, PUNE-44.**

**SAVITRIBAI PHULE PUNE UNIVERSITY, 2021-22 SEM-I**

**INDEX**

1. **Project Details**
   1. Project Title
   2. Problem Statement
2. **Abstract**
3. **Introduction**
4. **Technical Details**
5. **Algorithm Used (with Purpose)**
   1. **Correlation Matrix**
   2. **Random Forest**
   3. **Naïve Bayes**
6. **Output Snapshot**
7. **Conclusion**

**1.Project Details**

1.1 Project Title -

**Predicting Diabetes Using Pima Indian Diabetes Dataset**

1.2 Problem Statement -

Using the UCI PIMA Indian Diabetes dataset to predict a person has diabetes or not using the medical attributes provided.

The first is to analyse the data, and see if it is possible to gleam any further information from the data to determine correlation between parameters and diabetes. The second is to attempt to get the best accuracy score using various supervised learning machine learning algorithms.

**2.Abstract -**

Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affect other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by con- structing models from datasets collected from patients. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes.

**3. Introduction**

Diabetes is noxious diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particu- larly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabet- ics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected

data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however its tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

**4. Technical Details**

**About Pima Diabetes Dataset:**

The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

About models

We have created three working models as follows:

1. Correlation Analysis
2. Random Forest

Naive Bayesian

**5. Algorithm Used (with Purpose)**

**1.Correlation Matrix:** This Operator determines correlation between all attributes. Correlation is a statistical technique that can show whether and how strongly pairs of attributes are related.

* **Set Data:** This operator sets the value of one or more attributes of the specific example. The example is specified by the *example index* parameter. The *attribute name* parameter specifies the attribute whose value is to be set.
* **Numerical to Polynomial:** This operator changes the type of selected numeric attributes to a polynomial type. It also maps all values of these attributes to corresponding polynomial values.
* **Real to Integer:** This operator changes the type of the selected real attributes to integer type. It also maps all values of these attributes to integer values.
* **Generate Aggregation:** This operator generates a new attribute by performing the specified aggregation function on every example of the selected attributes.
* **Discretize by Size:** This operator converts the selected numerical attributes into nominal attributes by discretizing the numerical attribute into bins of user-specified size.
* **Retrieve:** This Operator can access stored information in the Repository and load them into the Process. This Object is often an Example Set but it can also be a Collection or a Model.
* **Replace Missing Values:** This Operator replaces missing values in Examples of selected attributes by a specified replacement. Missing values can be replaced by the minimum, maximum or average value of that attribute.
* **Set Role:** This Operator is used to change the role of one or more attributes. The role of an attribute describes how other Operators handle this attribute.

**2.Random Forest**

* **Retrieve**
* **Replace Missing Values**
* **Set Role**
* **Split Data:** This operator produces the desired number of subsets of the given ExampleSet. The ExampleSet is partitioned into subsets according to the specified relative sizes.
* **Random Forest:** This Operator generates a Random Forest model, which can be used for classification and regression.

**Apply Model:** This Operator applies a model on an ExampleSet. A model is first trained on an ExampleSet by another Operator, which is often a learning algorithm. Afterwards, this model can be applied on another ExampleSet.

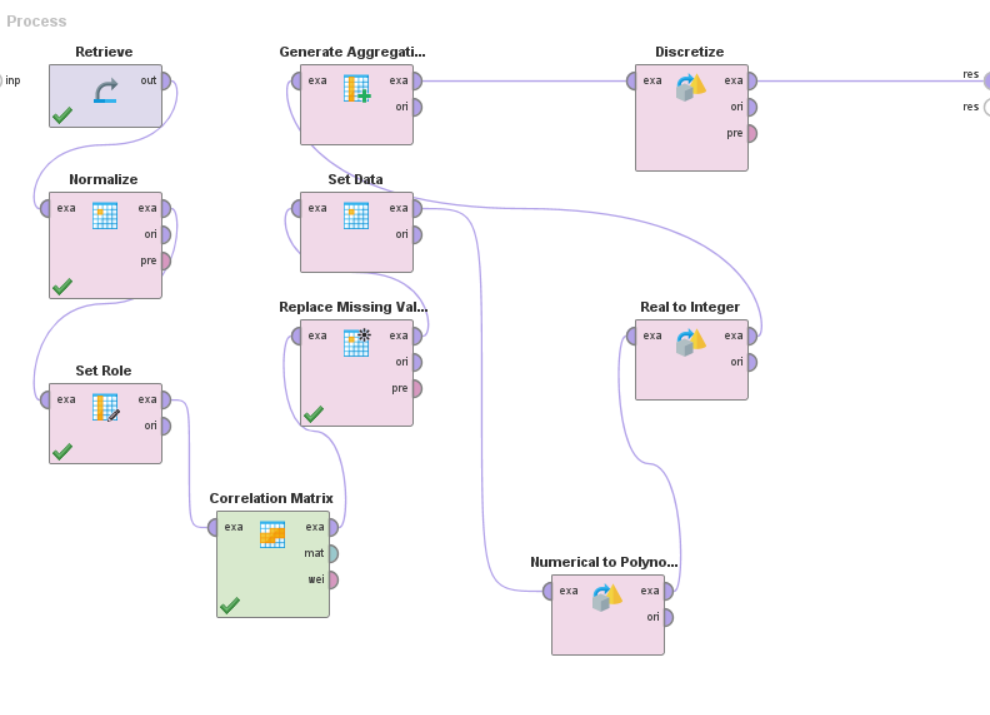
**3.Naïve Bayes**

* **Retrieve**
* **Replace Missing Values**
* **Set Role**
* **Discretize by Size**

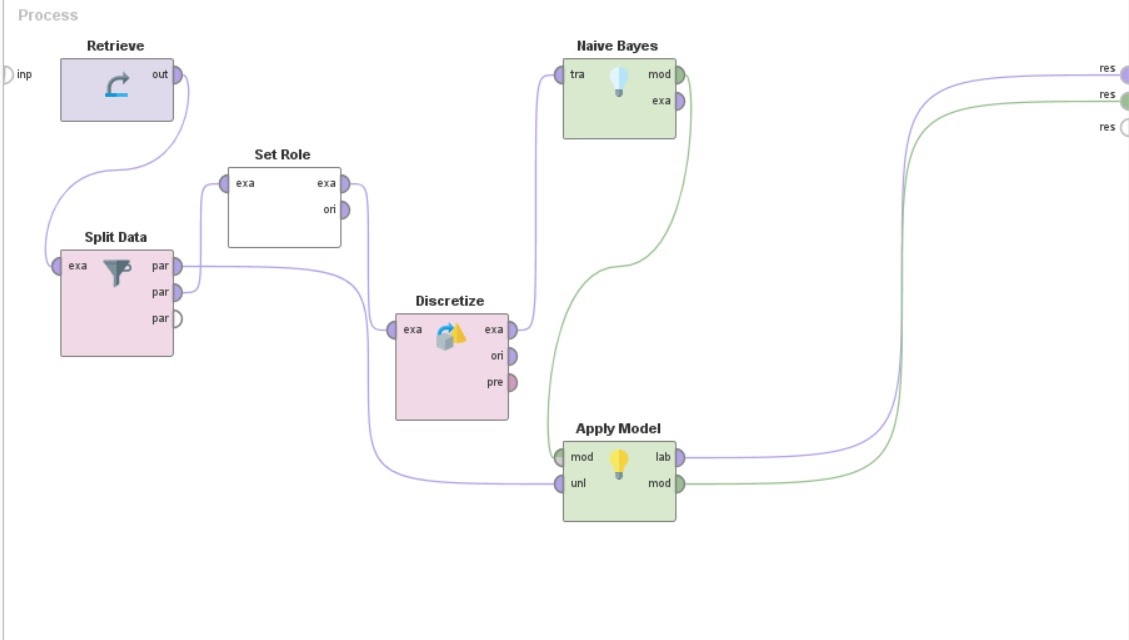
**Naive Bayes:** Naive Bayes is a high-bias, low-variance classifier, and it can build a good model even with a small data set. It is simple to use and computationally inexpensive. Typical use cases involve text categorization, including spam detection, sentiment analysis, and recommender systems.

**6. Output Snapshot**

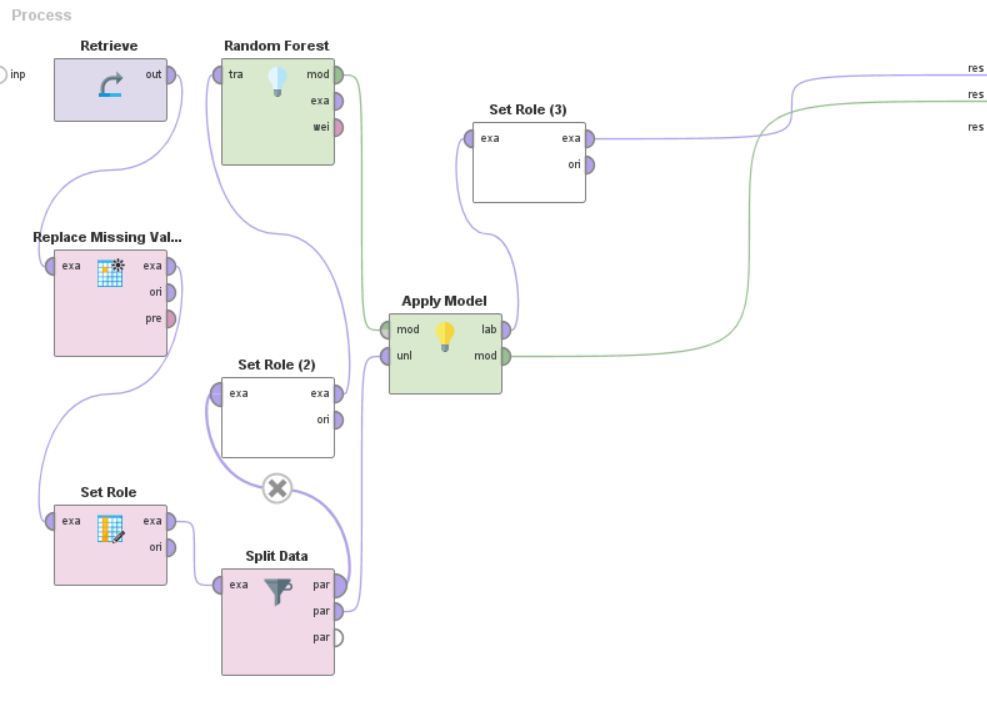
**Correlation analysis**



**Naïve Bayes:**



**Random Forest:**



**7. Conclusion**

We have studied about data mining and warehousing and performed Diabetes prediction using different models.

We can see that based on workings, there is a strong positive relation between glucose and diabetes, but we cannot be certain of causality. Secondly, Random Forest Algorithm appears to have the best overall metrics in terms of predicting the Diabetes classification for Pima Indians.