Defense On

Prediction of diabetes and analysis of its symptoms among patients of different ages using machine learning approaches.

Course Code: CSE 498

Course Title: Project & Presentation

*Submitted To*

Principle-supervisor

Teacher Name

Lecturer

Department of CSE,

University Name.

*Submitted By*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SL. NO. | NAME | ID | BATCH | DEPT. |
| 01 |  |  |  | CSE |
| 02 |  |  |  | CSE |

Date of Submission: 30th Dec, 2021

**1. Abstract**

Diabetes is a common problem in different ages people and those people who not only suffering diabetes but also, they have suffered from other symptoms which actually affect their health. However, to get a good recognition process we need a large amount of data which may be sufficient to generate good results. It is difficult to collect that amount of data from any health care service. On the other hand, we can use an online existing dataset to establish a proper system to get a better process for Diabetes problems. There are numerous ML algorithms and used for several reasons for examination and prediction.

In this work, how this disease gives better accuracy from the exit classification algorithm. Classification algorithms like SVM, KNN, Random Forest, and Naïve Bayes algorithms give their better performance on the basis of the dataset. Until now there are three types of diabetes which are Type 1, Type 2, and Type 3. We generate accuracy and predict those types by distributing our whole dataset according to the age of the patients. Some algorithms are implemented on the types to find out accuracy. In diabetes prediction, the Random Forest algorithm gives better accuracy than another algorithm 80%.

**Keywords:** Diabetes, ML, Classification Algorithm.

**2. Introduction**

In the Machine Learning process, Scikit-learning is a useful library used for most data processing work. It collects vast information from the data to give a good result. Data collection is one of the most important things in this process, a large amount of data that has every attributes that are needed is not easy. However, the internet is a big platform that always supports us to get solutions for every perspective process. Diabetes is a common term in people, people can affect by this disease in any way and its effects on their bodies by increasing other symptoms or increasing other symptoms disadvantages. The key point is that types of diabetes are divided by symptoms and ages. In many research works, people were trying to find out results differently by comparing those types of symptoms separately.

In our work, we make a process with ML classification algorithms to analyze the results from the dataset that we were collected online. People’s other attributes which are assigned in the dataset are the most important factor. We apply basic dataset queries for cleaning and clustering it. In diabetes, Glucose makes a good impact and other values too. We use a histogram plot to visualize those values to display how values impact people. We also visualize the plot of who have diabetes and how don’t have. Finally, we count algorithms accuracy to know which algorithm is better.

Diabetes is presented in different age group people and those are facing same type of problems, however, some people facing little problem and some facing more according to their daily routine. The dataset we are using, has people from the age 21 to the age 81. All of them are not diabetes positive, so simulating those people’s symptoms which are needed to understand on different age group symptoms. Distributing all ages into three groups because our patients counting start from the age 21. Based on that process we just trying to make a sequence when we distributing them. We create three age groups, first group is from the age 1 to 30, second group is 31 to 60, and last group is 60 to above ages, because age end with age 81. Distribution histogram plot analysis those symptoms.

Diabetes types are deepened in the age group of people. If we separate people of different ages, we can separate those types. However, for type 1 and Type 2 diabetes we don’t need pregnancy and age values, so we drop those values and create a new dataset which only works for type 1 and type 2. On the other hand, we just drop ages from the dataset for type 3 diabetes. And lastly, we count the accuracy of the diabetes types.

The whole process we have done with three very popular ML Classification algorithms which are SVM, KNN, Random Forest, and Naïve Bayes.

**3. Related work**

Maniruzzaman, M., Rahman, M., Ahammed, B. and Abedin, M, et al. [1] shows in their work, they mainly deal with type 1 diabetes. Machine learning selection algorithms and cloud computing process has been used to predict diabetes type 1. The key point is that they detect the patient's insulin level, based on that result they run the whole system. In this thesis, they are using IoT sensors that will signal immediately when insulin is low.

Rahman, M., Arman Nabid, R. and Hossain, M., et al. [2] shows in their work, They predict diabetes disease different section on the basis of different attributes in the dataset. They use machine learning approaches to find out better result. Among those classification algorithms they discover XGBoost algorithm which gives better accuracy than others.

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., Deng, K., Yan, D., Tang, H. and Lin, H., et al. [7] shows in their work, they selected which symptoms are actually related to grow diabetes disease. They focused on type 2 diabetes according to the dataset attributes. In their thesis they implement some PNN algorithms back-propagation, Bayesian Regulation, ANN, and GRNN as well as they count accuracy from those algorithms. However, they also include another algorithm MLP which shows better results than PNN algorithms. In addition, they separately depend on MLP accuracy for type 2 diabetes.

Suzanne, M., et al. [6] Shows in their work, they made a computational system by the help of fusing multifarious types of physical exercise data. To categorized healthy diabetes people they implement XGBoost model in the system. On the other hand, logistics regression was used to find out those people’s diabetes risks. Whatever they did their main goal was to categorized those key factor which influence patients to control their conditions.

**4. Methodology**

In our research work, we are working with different kinds of methods which give us better results among all methods. We use four popular Machine Learning classification algorithms to make the work happen, those algorithms are SVM, Random Forest, KNN, and Naïve Bayes. All of those algorithms are ML classification algorithms. However, we use those algorithms to calculate the accuracy of our patient’s values in the dataset. Patients who have diabetes are suffered from many symptoms which actually connected with diabetes. Those symptoms increase another disease in the patient’s body and also those are depending on people’s daily routine. Our used algorithms are popular in the processing sectors. On the other hand, algorithms can do better on the basis of dataset values. A good amount of data can give good results.

The age difference is a big factor in diabetes. Though all people face the same symptoms the level of those symptoms is not the same. We work on the basis of age values. In this thesis, we can see how people of different ages suffer from the level of their condition. From the diabetes value of the symptoms, we categorize and find out the history behind them.

Diabetes has three types, type 1, type 2, and type 3. We analyze those types according to age and implement 4 Machine Learning Classification algorithms to analyze them. We use separate files to execute the different results. Separate dataset into three datasets to analyze genuine work.

**4.1 SVM Algorithm**

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they’re able to categorize new text.

We can compare this classification algorithm with newer algorithm like Neural Network which is actually separated in two steps: Higher speed and better performance.

**4.2 Random Forest Algorithm**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

**4.3 KNN algorithm**

K Nearest Neighbor algorithm falls under the Supervised Learning category and is used for classification and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets. As the name (K Nearest Neighbor) suggests it considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new Datapoint.

**4.4 Naïve Bayes Algorithm**

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

**4.5 Proposed Methodology**

We read the whole dataset first. In our dataset, we have 768 patient values and 9 attributes. After cleaning and clustering, we process the dataset to use our further work. Histogram plot visualizes each attributes values distribution and we also create 2 other histogram plots for the people who have diabetes and how don’t have dietetics. We separate the whole data into two terms one is training which will carry 80% and another is testing and testing carries 20%. After training and testing we get our accuracy level.

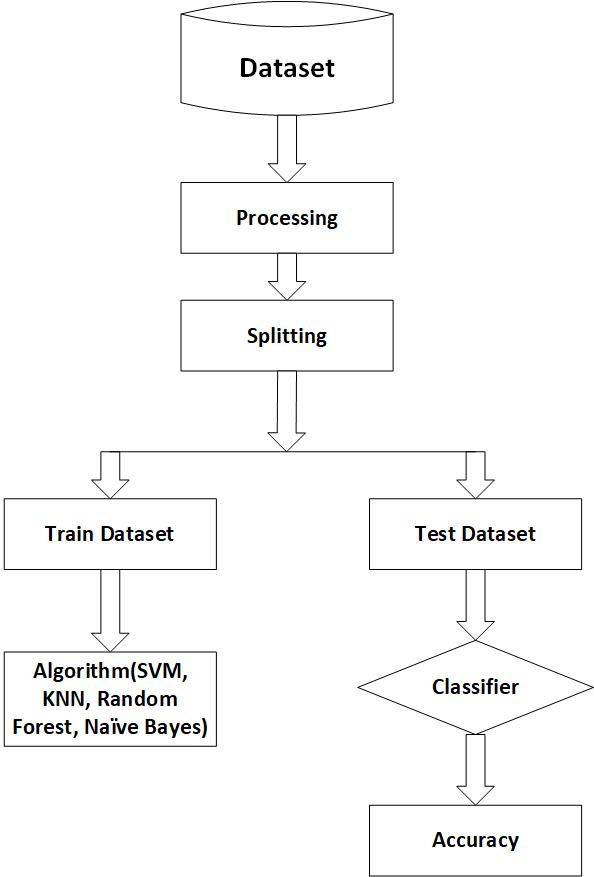


Fig 1: Proposed Methodology

**4.6 Flowchart**

**4.6.1 Age Group**

In the dataset, there are different ages people who have diabetes or do not have diabetes who is separate them into three age groups fast age group is from 1 to 30, seconds from 31 to 60 and last is from 61 to above. In our dataset age starts from 21 and finished at age 81. That is why, we change our condition like we do not face age for 1, we just separate the dataset, which is below 30, as well as, we separate from the last group of age people who is greater than 61. With this condition, we separate the dataset and again save and read the new dataset, then drop all the rows that do not have diabetes. Finally, with a distribution plot, we visualize all dataset diabetes symptoms.

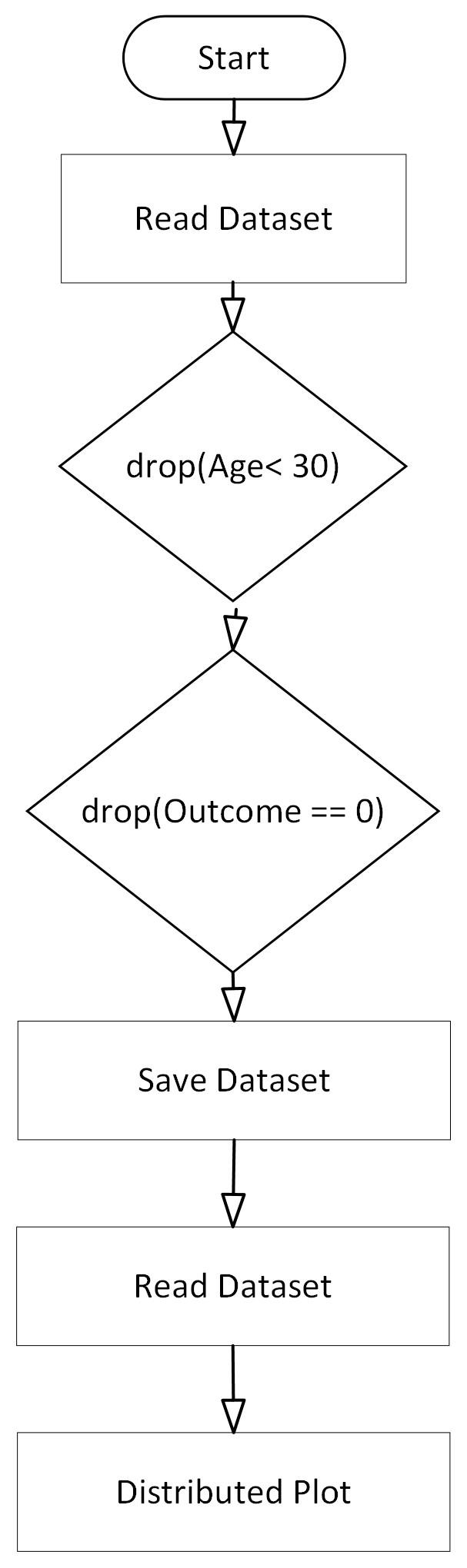


Fig 2: Age Group Flowchart

**4.6.2 Diabetes Type**

After facing all the conditions for our type 1, type 2, and type 3, we find out accuracy with classification algorithms. Separate the dataset according to age and move them to another new dataset. Diabetes types can separate according to the age limit. We can also use glucose level for that, but glucose level. For type 1 and type 2 we drop the pregnancies column from the dataset. And finally, with algorithms SVM, KNN, Random Forest, and Naïve Bayes, we calculate accuracy. In addition, for every type of work, we have to use separate files. Among the four algorithms, the random forest gives a better result. It gives for type 1 66%, for type 2 80% and for type 3 67%.

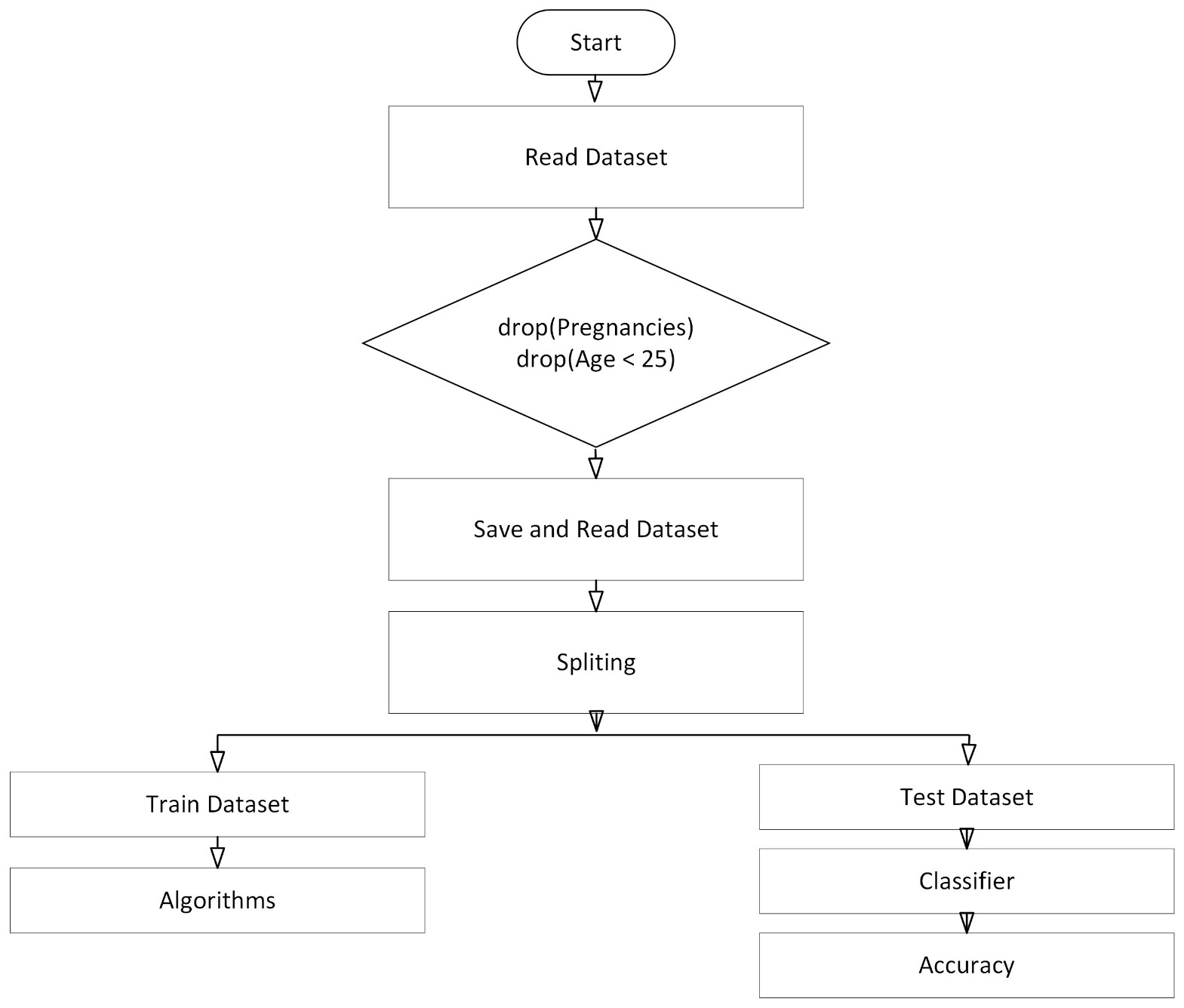


Fig 3: Diabetes Types Flowchart

**5. Result**

The whole process is executed step by step like a plan. Jupyter Notebook used to run python code and use machine learning models to create a genuine system. With the Kaggol dataset, we process the data with the proper condition and find out balanced dataset through visualization. Mathplotlib gives a nice histogram bar plot, through that plot we can understand the problem's results. I dataset we have 9 attributes, but not all of them are needed for analysis. We use only those attributes which are the symptoms for diabetics. We select those attributes and one by one visualize them according to positive and negative diabetes counts.

There are many Machine learning classification algorithms. However, we use 4 of them which are SVM, KNN, Random Forest, and Naïve Bayes. To train and test the dataset for accuracy, we gave 80% for training and 20% for testing.

**5.1 Dataset**

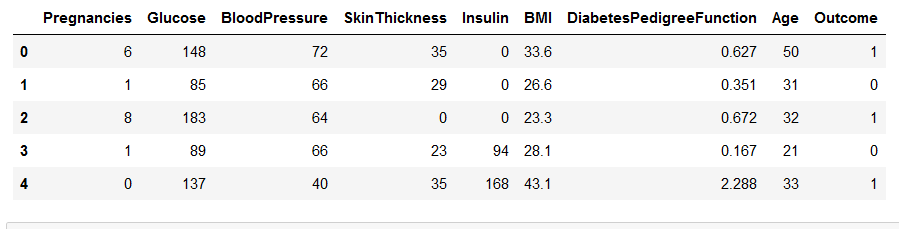
Dataset was formally occupied from (Kaggol). The main purpose of this dataset is to employ signs of progress throughout the analysis process in the case of those who have diabetes. Our dataset has 8+1 = 9 attributes and 768 records of patients. 

Fig 4: Dataset Information

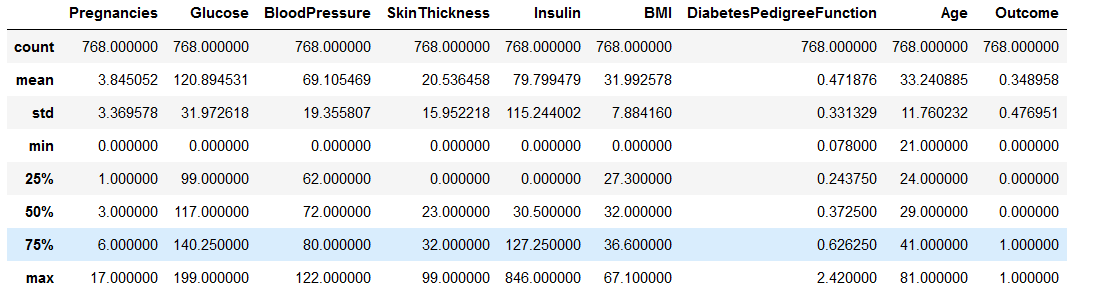
**5.2 Result Analysis** 

Fig 5: Mean and Standared Deviation of Dataset

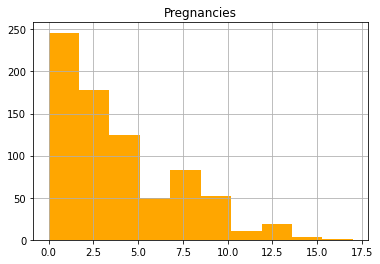
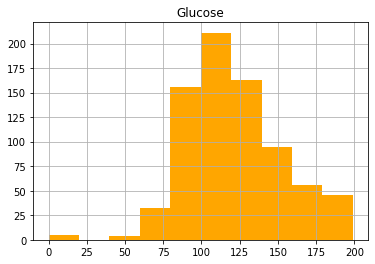
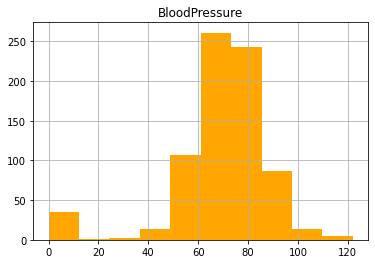
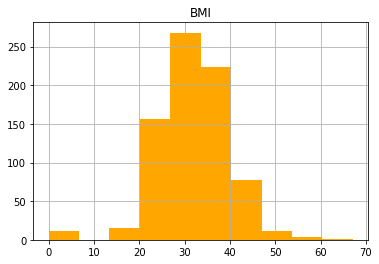
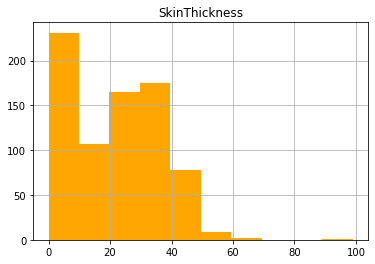
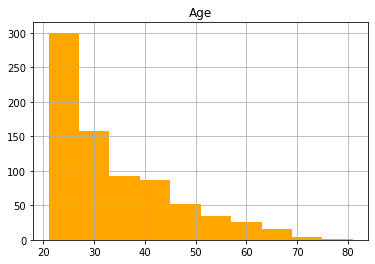
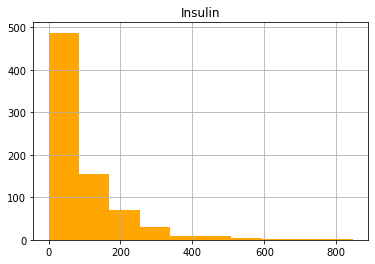
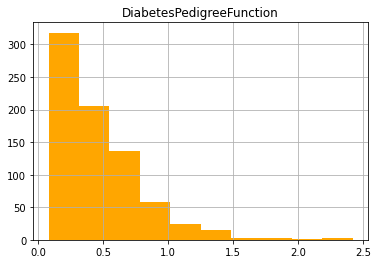
****       

Fig 6: Histogram Plot after cleaning

**5.3 Patients Percentage**

65% percent patients are negative and only 35% are positive.

[500, 268]

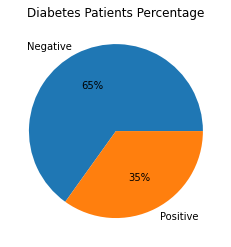


Fig 7: Diabetes Patients Percentage

**5.4 Heat Map**

****

Fig 8: Heat Map Plot

**5.5 Histogram Plot**

We divided the visualization of the plot into two sections so that we could understand and compare them. This comparison can sense us about the condition of diabetes. We select separately the attributes from the dataset and analysis on them.

**5.6 Pregnancies**

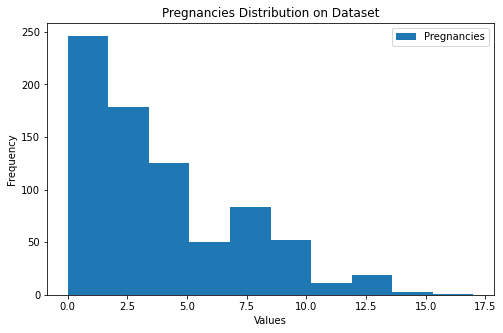
****

Fig 9: Pregnancies Distribution on Dataset

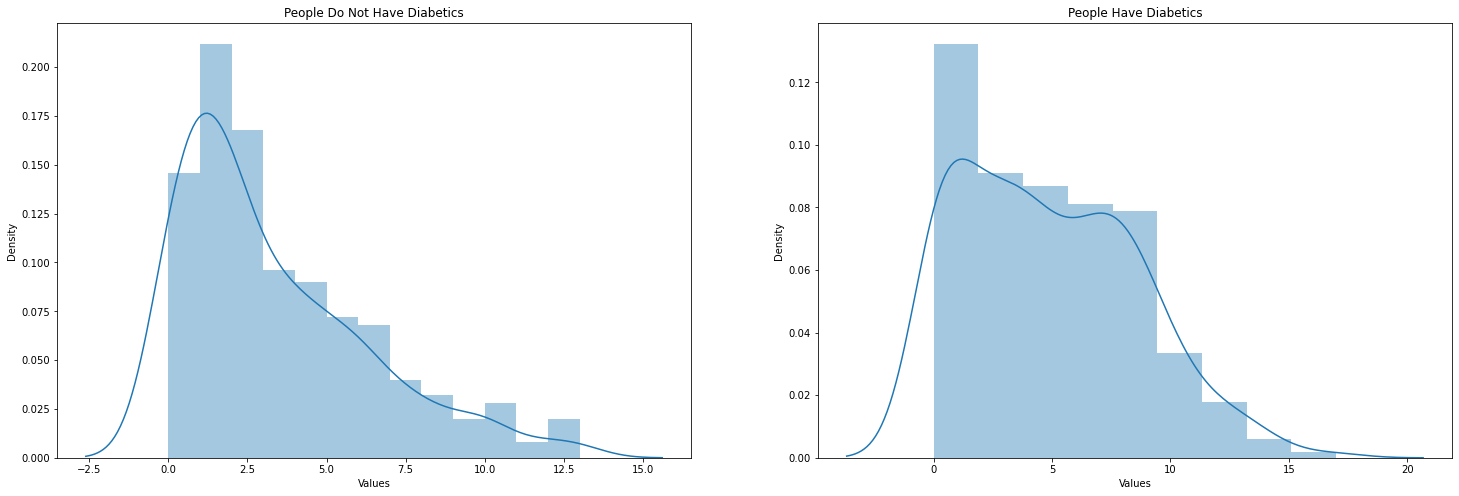


Fig 10: Pregnancies Distribution on positive and negative Diabetes.

**5.7 Classification Algorithm Accuracy**

Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors, and Naïve Bayes algorithms are applied to the dataset to calculate the accuracy. We train and test the dataset before applying it to algorithms.

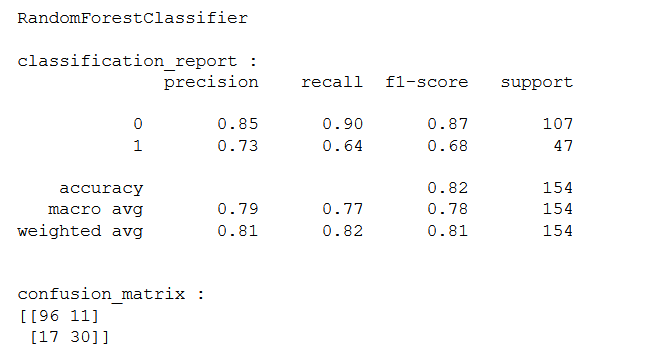


Fig 11: Confusion matrix for Random Forest classification.

**5.8 Types**

**5.8.1 Type 1**

|  |  |
| --- | --- |
| Algorithm | Accuracy |
| SVM | **67%** |
| Naive Bayes | **63%** |

Table 1: Diabetes Type 1 Accuracy

**5.8.2 Type 2**

|  |  |
| --- | --- |
| Algorithm | Accuracy |
| SVM | **67%** |
| Naive Bayes | **70%** |

Table 2: Diabetes Type 2 Accuracy

**5.8.3 Type 3**

|  |  |
| --- | --- |
| Algorithm | Accuracy |
| SVM | **67%** |
| Naive Bayes | **58%** |

Table 3: Diabetes Type 3 Accuracy

**5.9 Age Group Symptom Analysis**

**5.9.1 Age Group 1**

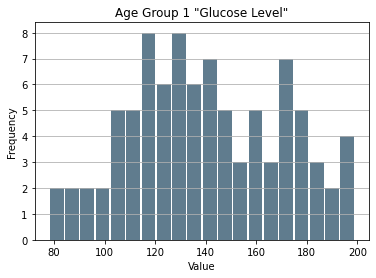
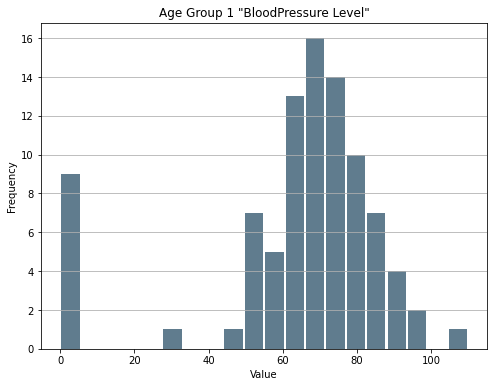
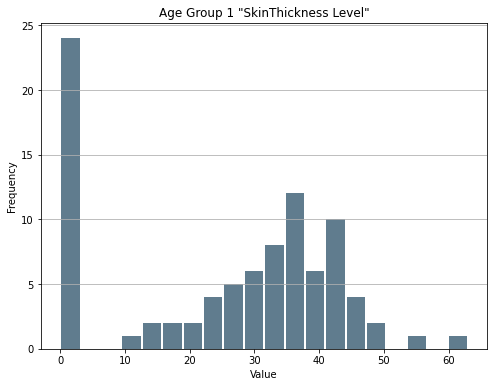
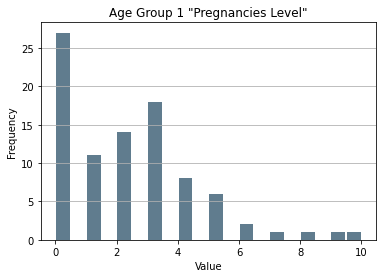
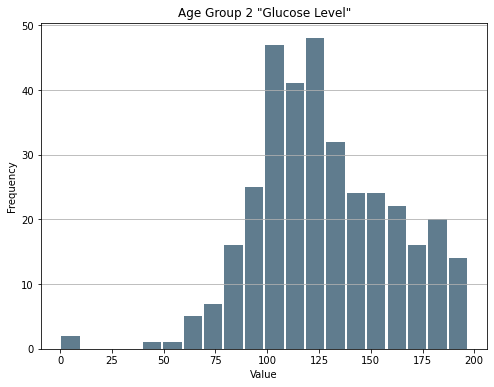
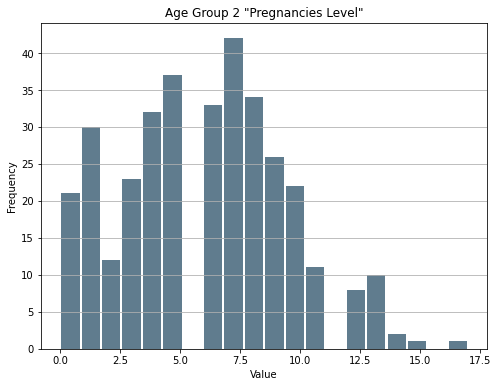
********

Fig 12: Age Group 1 all Symptoms

**5.9.2 Age Group 2**

****

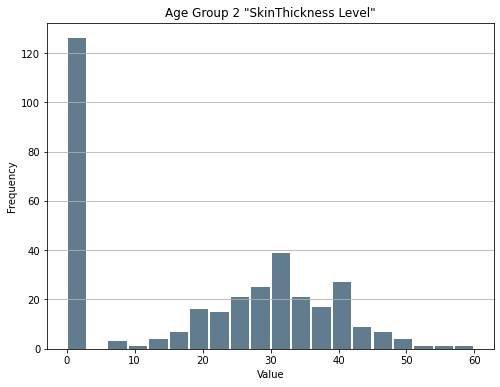
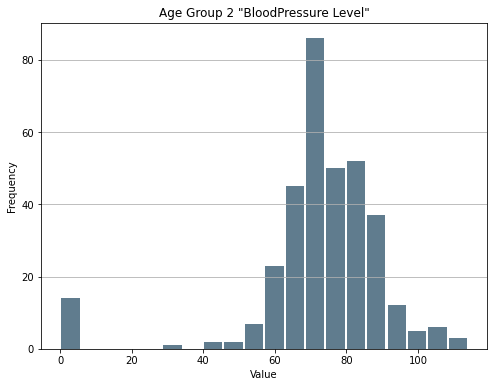
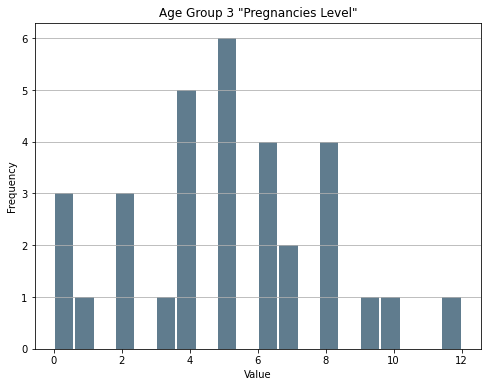
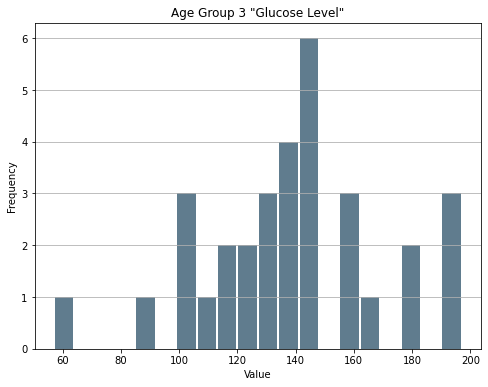
****

Fig 13: Age Group 2 all Symptoms

**5.9.3 Age Group 3**

** **

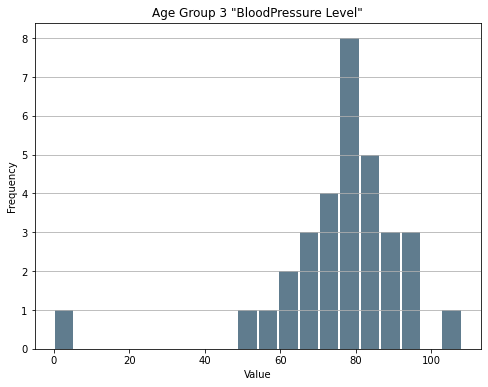
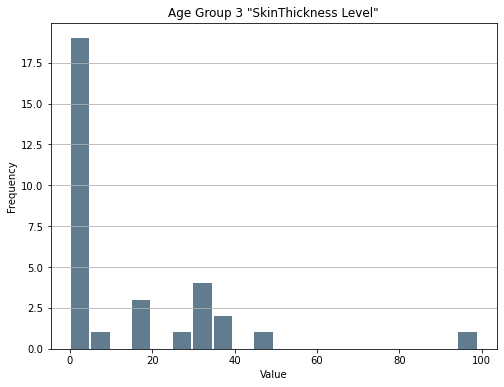
** **

Fig 14: Age Group 3 all Symptoms

**6. Discussion**

We generating a system process, which can give us a better solution, or way that we can use to make a system, which can give us peoples diabetes results as well as how they feel. If they have any symptoms after having diabetic disease, how its look or feel. In addition, we need a large dataset that will help us to get correct results because a small amount of data can’t give us the proper result. We have used four classification algorithms that help us to generate this way to work. However, after the accuracy that we make with the help of ML classification, we find out that almost every algorithm gives us the nearest result from each other. But we choose an algorithm which is Random Forest and its accuracy is 80%. We apply SVM and Naïve Bayes classification to find out diabetes types accuracy. Analysis age group symptoms on the basis of age and how have diabetes.

**7. Conclusion**

This work's aim is to make a process system that will give good accuracy to not only diabetes but also types of diabetes. A large dataset collection is not easy, that is why we use the existing dataset and which has all attributes we needed for our work. The type of diabetes depends on the age of the patients. The first type of production will separate age depending on age group and then find the accuracy we apply SVM and Naïve Bayes classification algorithms. Analysis of the patient's symptoms after separating the dataset according to age and how have diabetes give us an understanding of their conditions. Large Dataset gives proper results. For diabetes prediction, Random Forest gives better regulation 80%.

**8. Reference**

1. Maniruzzaman, M., Rahman, M., Ahammed, B. and Abedin, M., 2020. Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8(1). DOI: <https://doi.org/10.1016/j.icte.2021.02.004>
2. Rahman, M., Arman Nabid, R. and Hossain, M., 2020. Disease Symptom Analysis Based Department Selection Using Machine Learning for Medical Treatment. 2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS),. DOI: https://doi.org/10.48161/qaj.v1n2a55
3. Ram, Anant, and Honey Vishwakarma. "Diabetes Prediction using Machine learning and Data Mining Methods." In IOP Conference Series: Materials Science and Engineering, vol. 1116, no. 1, p. 012135. IOP Publishing, 2021. DOI: https://doi.org/10.1088/1757-899X/1116/1/012135
4. Nagi, A., Javed Awan, M., Javed, R. and Ayesha, N., 2021. A Comparison of Two-Stage Classifier Algorithm with Ensemble Techniques On Detection of Diabetic Retinopathy. 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA),. DOI: https://doi.org/10.1109/CAIDA51941.2021.9425129
5. Alcalá-Rmz, V., Galván-Tejada, C., García-Hernández, A., Valladares-Salgado, A., Cruz, M., Galván-Tejada, J., Celaya-Padilla, J., Luna-Garcia, H. and Gamboa-Rosales, H., 2021. Identification of People with Diabetes Treatment through Lipids Profile Using Machine Learning Algorithms. Healthcare, 9(4), p.422. DOI: <https://doi.org/10.3390/healthcare9040422>
6. Suzanne, M., 2009. Insulin resistance and Alzheimer’s disease. *BMB reports*, *42*(8), p.475.
7. Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., Deng, K., Yan, D., Tang, H. and Lin, H., 2021. Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. *Information Fusion*, *75*, pp.140-149.
8. He, Y., Lakhani, C.M., Rasooly, D., Manrai, A.K., Tzoulaki, I. and Patel, C.J., 2021. Comparisons of polyexposure, polygenic, and clinical risk scores in risk prediction of type 2 diabetes. *Diabetes care*, *44*(4), pp.935-943.
9. Østergaard, H.B., van der Leeuw, J., Visseren, F.L. and Westerink, J., 2021. Comment on Vistisen et al. A validated prediction model for end-stage kidney disease in type 1 diabetes. Diabetes Care 2021; 44: 901–907. *Diabetes Care*, *44*(6), pp.e139-e139.