**Chapter 1**

**Introduction**

**Chapter 1**

**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.

Diabetics 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.

**CHAPTER 2**

**RELATED WORK**

**CHAPTER 2**

**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.

Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K. and Jonkman, M., et al. [10] shows in their work, they were trying to find out the best performing algorithm for diabetes prediction. They used 4 algorithms those are Gradient Boosting (GB), Support Victor Machine (SVM), AdaBoost (AB), Random Forest (RF). Algorithms did good work to predict diabetics but not all of them gives better performance result. According to accuracy result only Random Forest (RF) give highest result which is 99.35%, On the other Hand SVM and AB had the lowest performance.

Selvi, R.T. and Muthulakshmi, I., et al. [11] shows in their work, to get a good performance they introduce a new method which is MRODC. Hadoop ecosystem differently used this method to run the whole system. On the other hand, K-means learning improved its process and involved GBT classification. As a result of using this method is consistently giving better results, precision is 99.23%.

Orabi, K.M., Kamal, Y.M. and Rabah, T.M., et al. [12] shows in their work, they worked on a system that can predict diabetes, to find out the age specifically people are suffering diabetes problem how to have diabetes or not have. They did this by machine learning algorithms support, they run their plan with Decision Tree. The system was designed well and it gives satisfactory results.

Kumar, D.A. and Govindasamy, R., et al. [13] They tried the predictions with other classification algorithms such as support vector machines (SVM), Naive Bayes, and, logistic regression, with cross-validation. Their work aimed to find out a good model from different machine learning algorithms that can be used to predict diabetes by using these processes on a dataset of diabetes patients. They compared the different algorithms based on the accuracy parameter, and they found that the support vector machine (SVM) was a good accuracy score among other algorithms.

Alkaragole, M.L.Z. and Kurnaz, A.P.S., et al. [14] the researcher in this thesis tried different data mining methods as a hybrid framework using the dataset. And they applied classification algorithms like decision tree, Naive Bayes, and support vector machines (SVM) to assess the Apache servers. They implemented their proposed method and they achieved an accuracy of 94%.

Woldemichael, F.G. and Menaria, S., et al. [15] they applied different data mining methods to achieve a proposed predictive model to compare the accuracy between different machine learning algorithms. And they found that the logistic regression accuracy score was 82.35% and that was the highest score compared to other classifiers such as K-nearest neighbor (K-NN), Naive Bayes, Decision tree, and support vector machine (SVM). Logistic regression performs better than others but, other algorithms also did good work.

**CHAPTER 3**

**METHODOLOGY**

**CHAPTER 3**

**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.

**3.1 Support Vector Machine (SVM) Algorithm**

**3.1.1 Advantages**

* Regularization Capabilities.
* Handles non-linear data efficiently
* Solves both Classification and Regression work
* SVM model is more stabile

**3.1.2 Disadvantages**

* Choosing an appropriate kernel function is difficult.
* Extensive memory requirements.
* Requires Features Scaling.
* Long training time.

**3.2 Random Forest Algorithm**

**3.2.1 Advantages**

* Solve both classification and regression problems.
* Works well in both categorical and continuous variable.
* Handle missing value.

**3.2.2 Disadvantages**

* Complexity.
* Long training period.

**3.3 KNN algorithm**

**3.3.1 Advantages**

* No training periods.
* New data can be added seamlessly.
* Easy to implement.

**3.3.2** **Disadvantages**

* Does not work well with large dataset.
* Does not work well with high dimensions.
* Need features scaling.

**3.4 Naïve Bayes Algorithm**

**3.4.1 Advantages**

* Less Training Period.
* Easy to implement.

**3.4.2 Disadvantages**

* In real life it is impossible to independent Prediction.

**3.5 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.

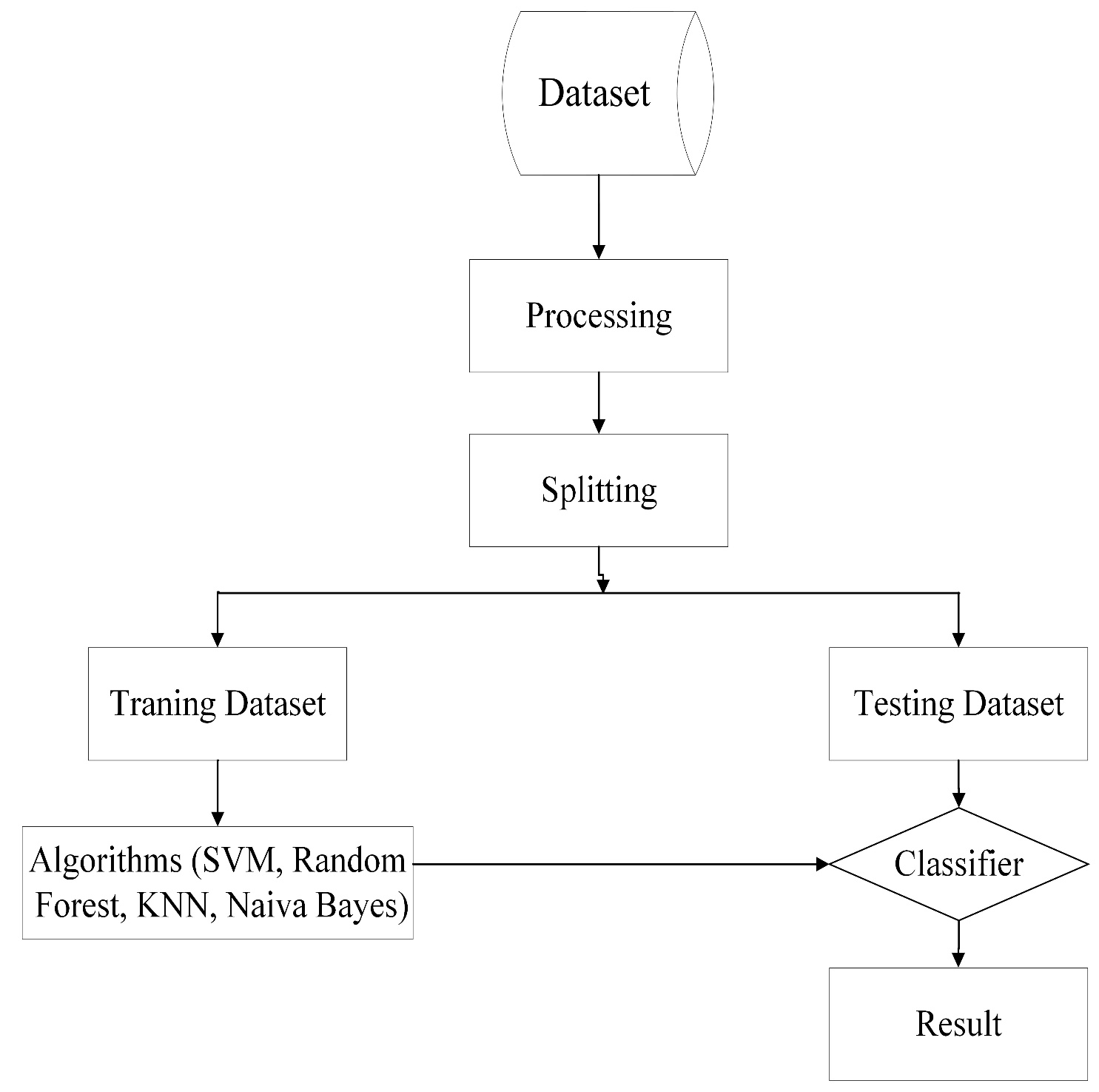


Fig 3.1: Proposed Methodology

**3.6 Flowchart**

**3.6.1 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.

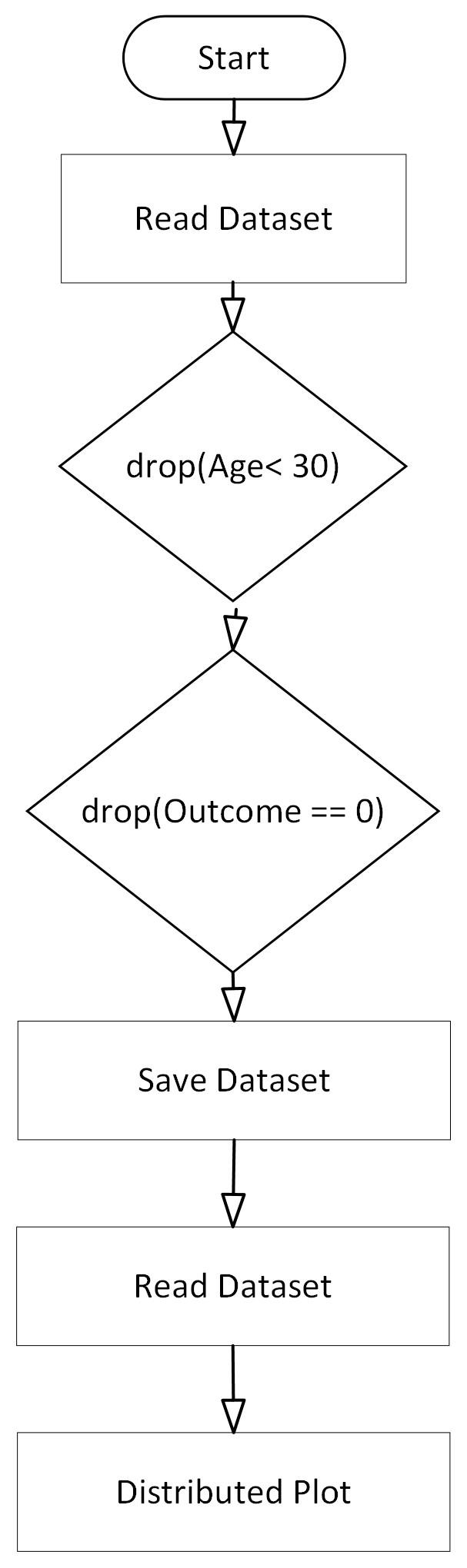


Fig 3.2: Age Group Flowchart

**3.6.2 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%.

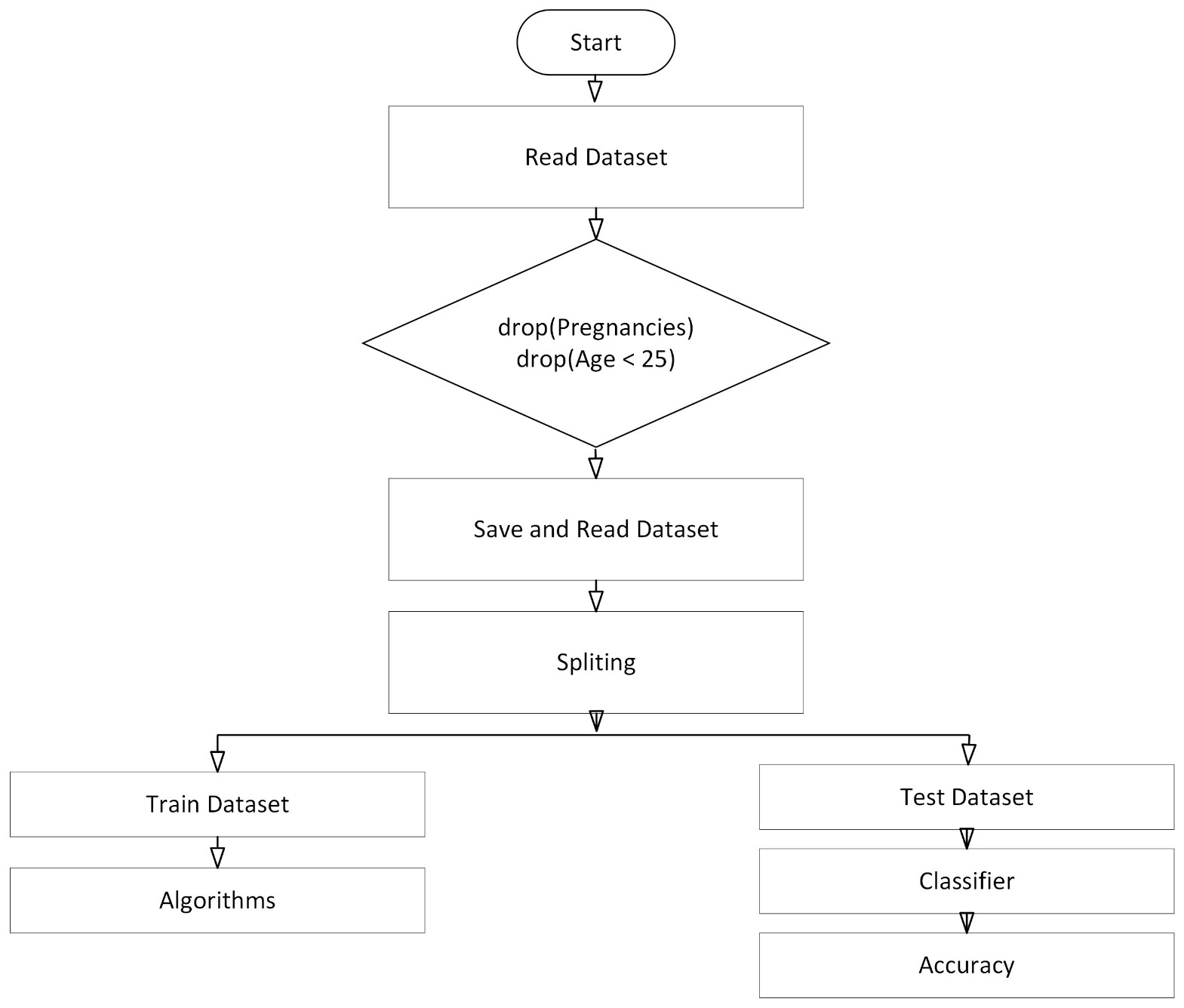


Fig 3.3. Diabetes Types Flowchart

**CHAPTER 4**

**PERFORMENCE EVALUATION**

**CHAPTER 4**

**PERFORMENCE EVALUATION**

The result section shows us the multiple results that we are made by the prediction processes and we use mathplotlib to make plot and visualize those images to show our works. Firstly, we generate predictions then types of diabetes, and also, we distributed the whole dataset and visualize diabetics positive people’s symptoms which depended on their ages.

We find out the accuracy of the people and plot them. Those plots and results give us a clear view of diabetes.

**4.1 Diabetes Prediction**

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.

**4.2 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.

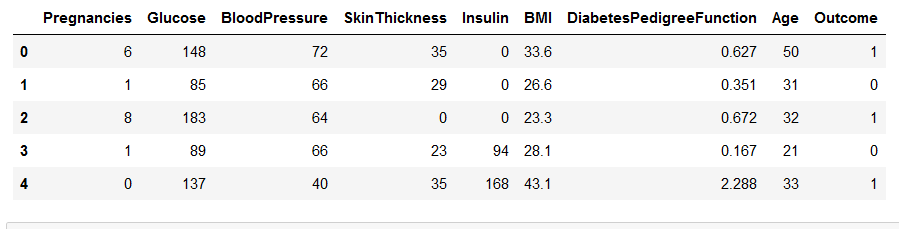


Fig 4.1: Dataset Information

**4.3 Result Analysis**

**4.3.1 Mean and Standard Deviation of Dataset**

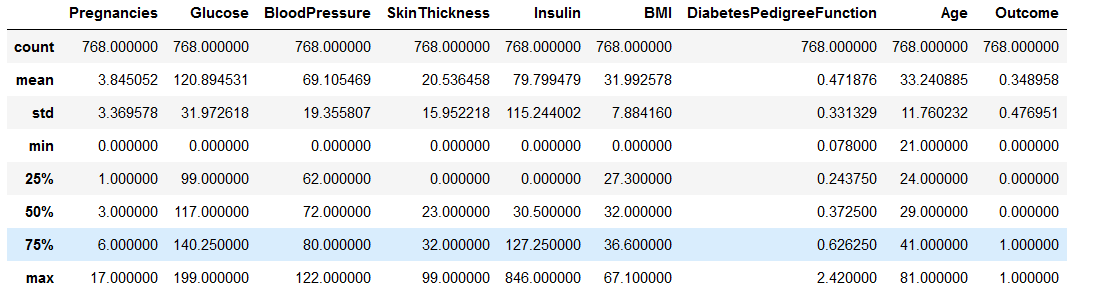


Fig 4.2: Mean and Standared Deviation of Dataset

**4.3.2 Balanced Dataset**

Here 0 represent negavive value and 1 represent Negative value. Tolat dataset has 768 values and among them 500 negative and 268 positive.

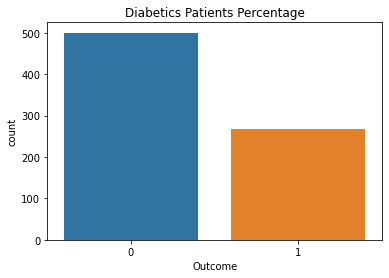


Fig 4.3: Check Balanced Data

**4.3.3 Patients Percentage**

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

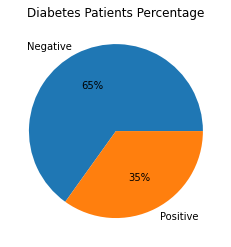


Fig 4.4: Diabetes Patients Percentage

**4.3.4 Simple Age Distribution on Diabetes**

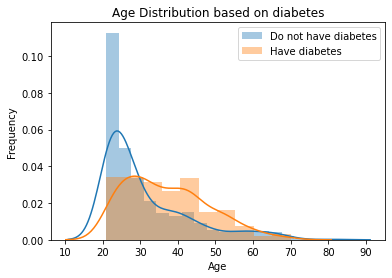
****

Fig 4.5: Simple age Distribution on Diabetes

This graph showsdiabetics patients percentage is high in between near 25 to 32 ages people. Younger people are much affected than the oldest people.

**4.3.5 Heat Map**

****

Fig 4.6: Heat Map Plot

**4.3.6 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.

**4.3.6.1 Pregnancies**

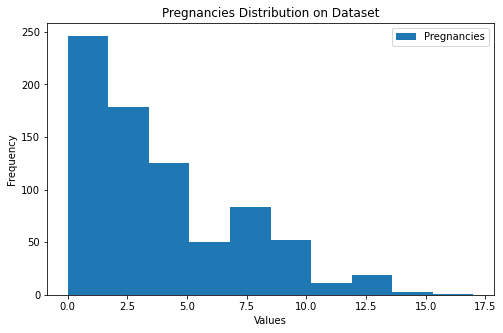
****

Fig 4.7: Pregnancies Distribution on Dataset

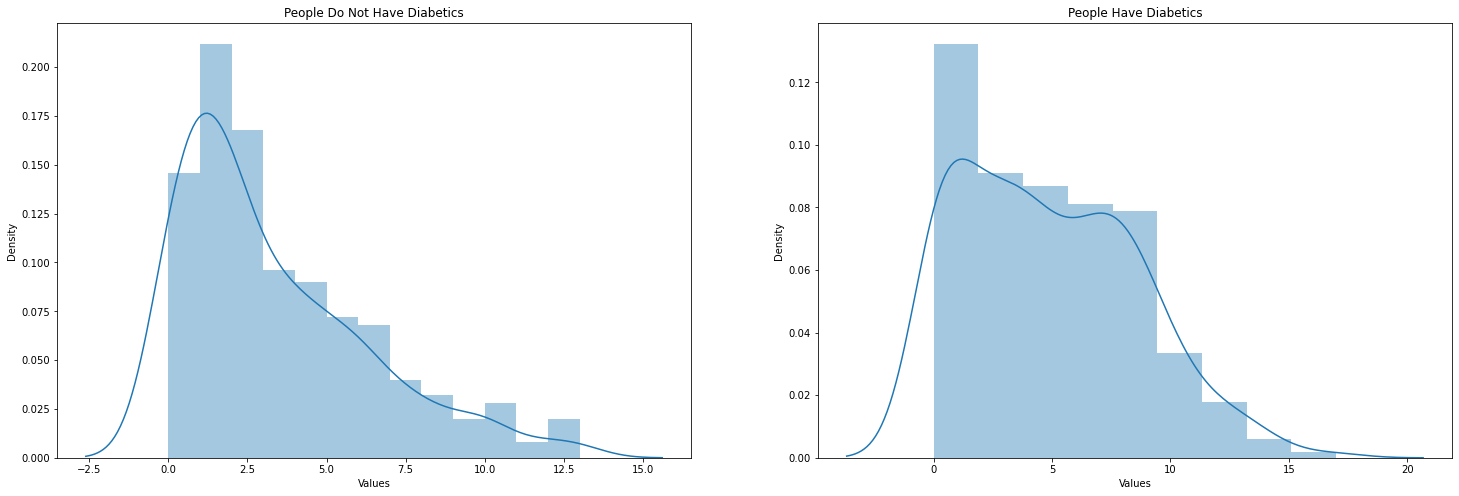


Fig 4.8: Pregnancies Distribution on positive and negative Diabetes.

**4.3.6.2 Glucose**

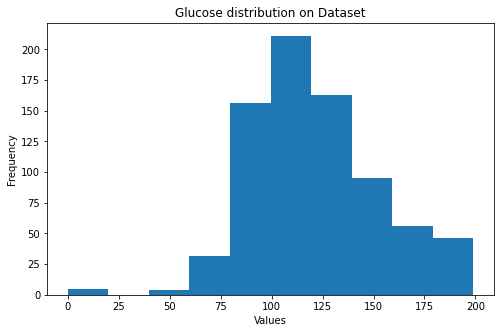
****

Fig 4.9: Glucose Distribution on Dataset

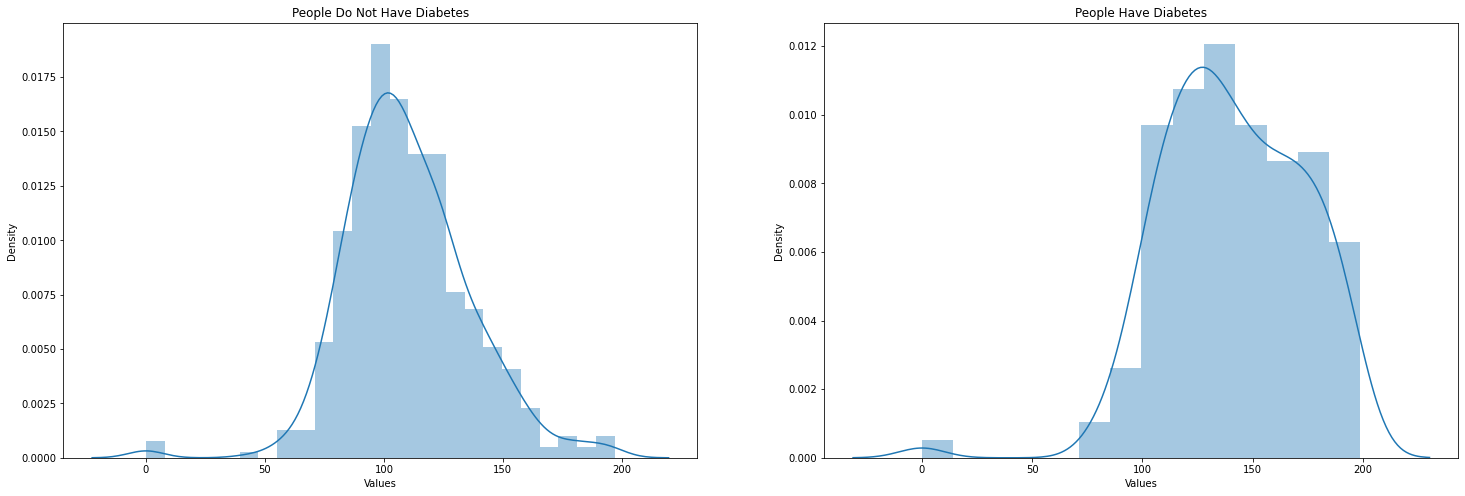


Fig 4.10: Glucose Distribution on positive and negative Diabetes.

**4.3.6.3 Blood Pressure**

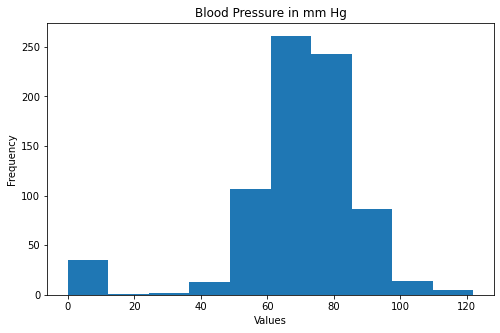
****

Fig 4.11: Blood Pressure Distribution on Dataset

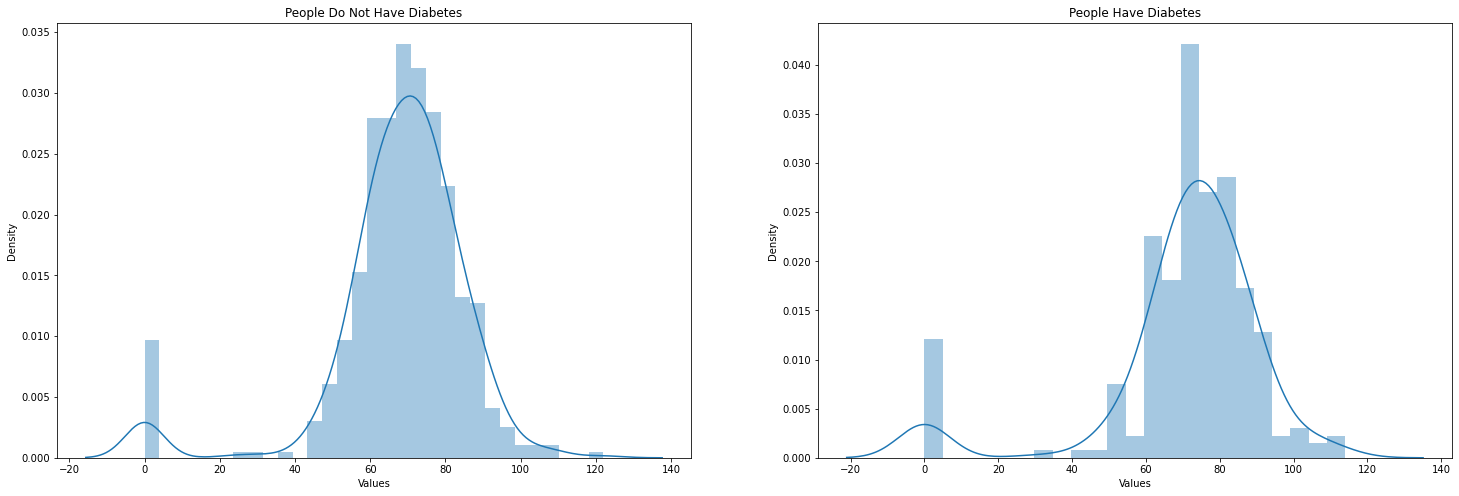


Fig 4.12: Blood Pressure Distribution on positive and negative Diabetes.

**4.3.6.4 Skin Thickness**

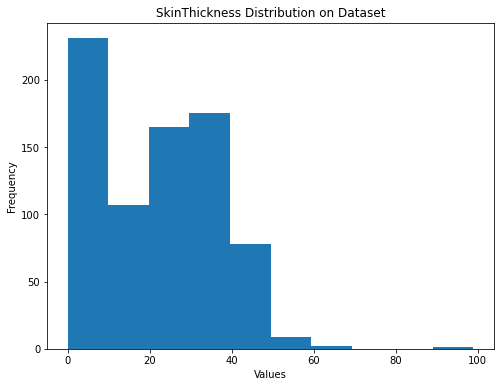


Fig 4.13: Skin Thickness Distribution on Dataset

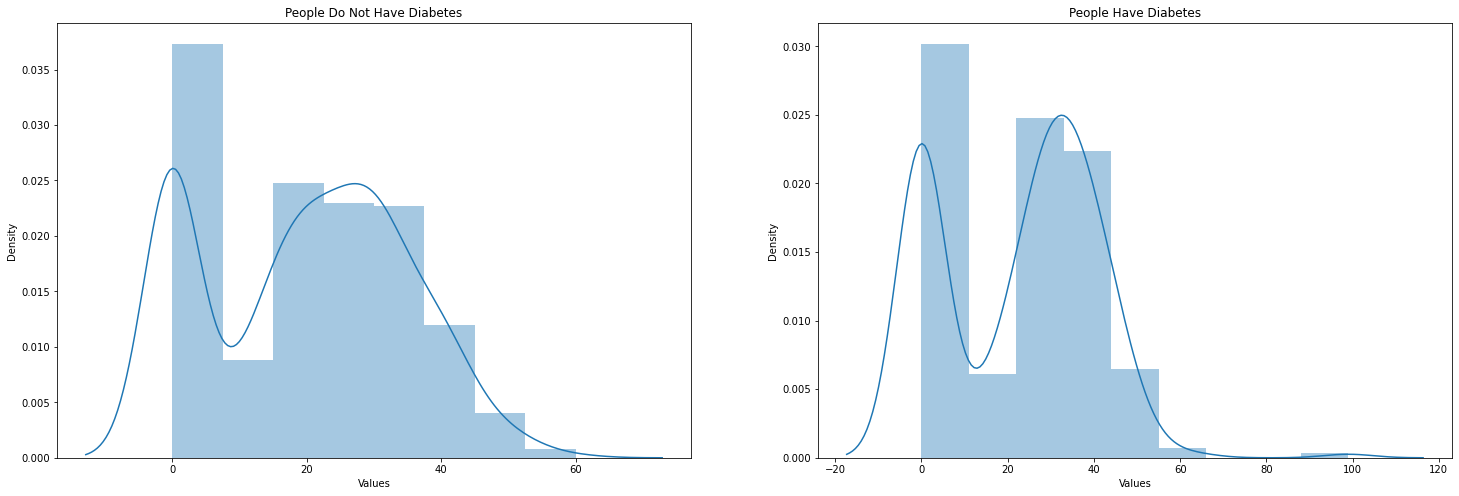


Fig 4.13: Skin Thickness Distribution on positive and negative Diabetes.

**4.3.6.5 Insulin**

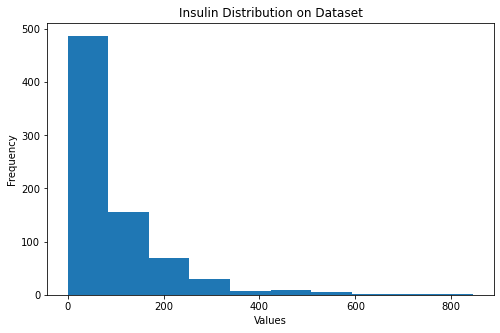


Fig 4.14: Insulin Distribution on Dataset

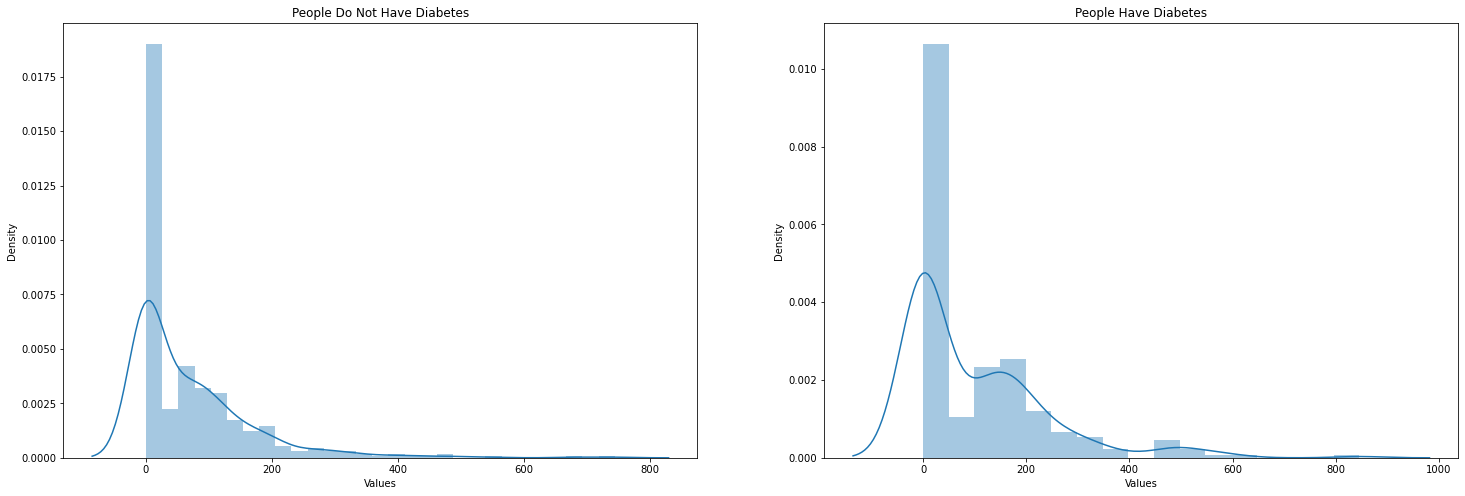


Fig 4.15: Insulin Distribution on positive and negative Diabetes.

**4.3.6.6 BMI**

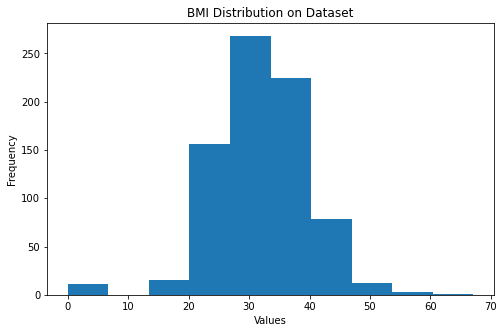


Fig 4.16: BMI Distribution on Dataset

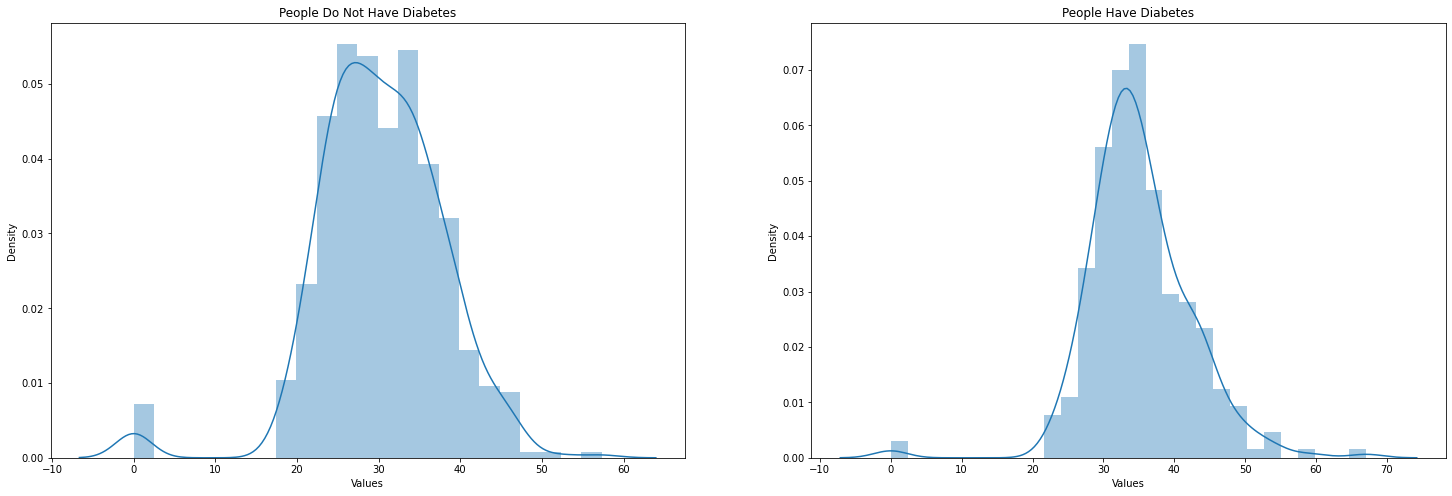


Fig 4.17: BMI Distribution on positive and negative Diabetes.

**4.3.6.7 Diabetes Pedigree Function**

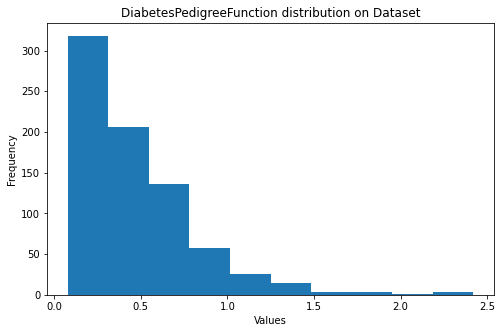


Fig 4.18: Diabetes Pedigree Function Distribution on Dataset

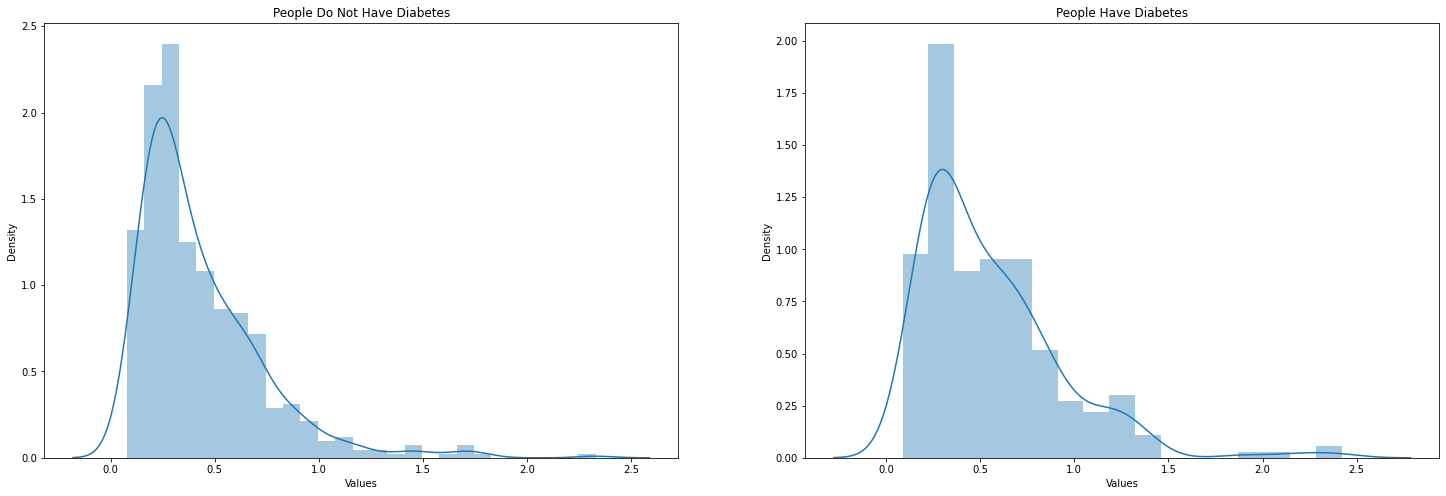


Fig 4.19: Diabetes Pedigree Function Distribution on positive and negative Diabetes.

**4.3.6.8 Age**

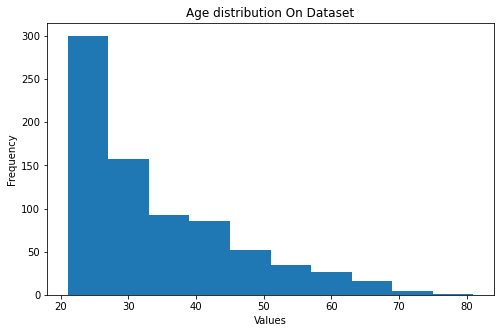
****

Fig 4.20: Age Distribution on Dataset

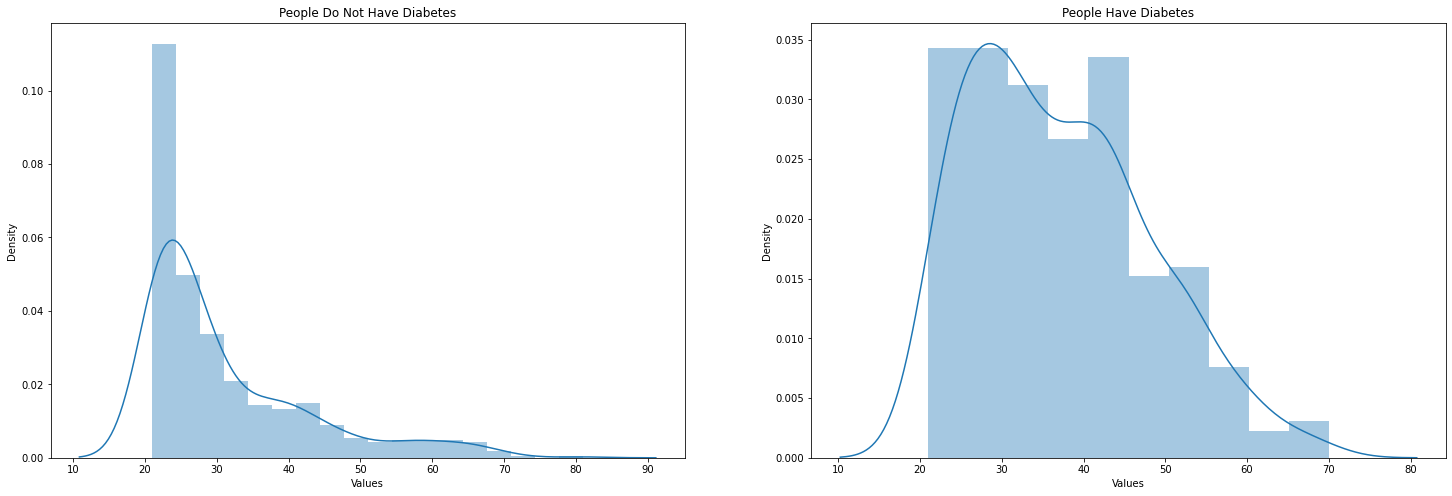


Fig 4.21: Age Distribution on Positive and Negative Diabetes.

**4.3.7 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.

**4.3.7.1 SVM Algorithm**

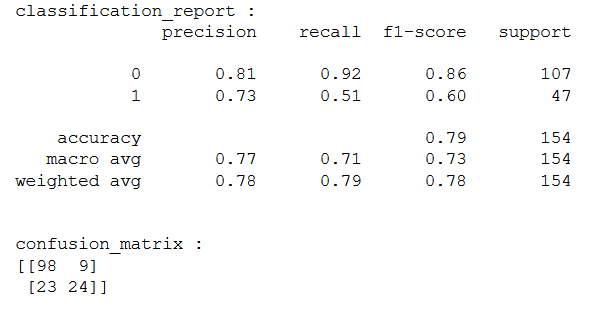


Fig 4.22: Confusion Matrix for SVM Classification.

**4.3.7.2 Random Forest Algorithm**

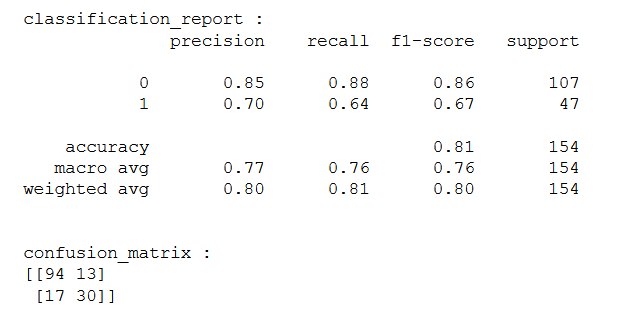


Fig 4.23: Confusion Matrix for Random Forest Classification.

**4.3.7.3 KNN Algorithm**

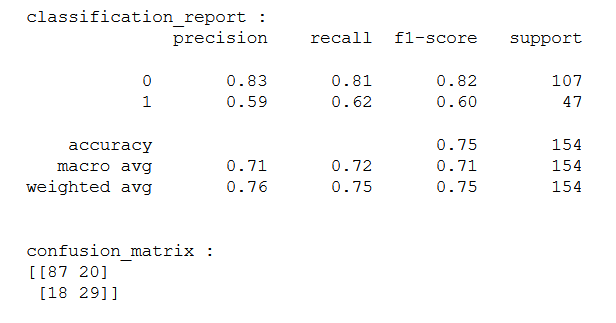


Fig 4.24: Confusion Matrix for KNN Classification.

**4.3.7.4 Naïve Beyas Algorithm**

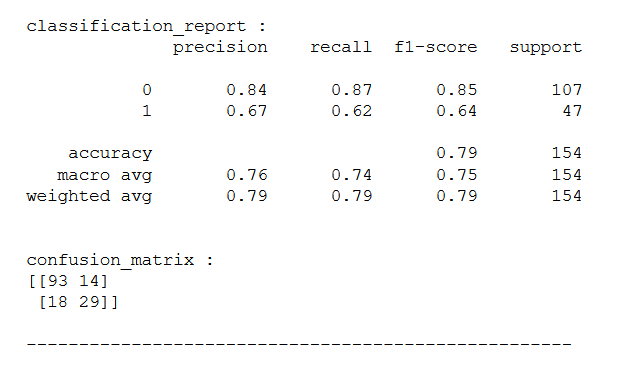


Fig 4.25: Confusion Matrix Naïve Bayes Classification.

**4.4 Types**

**4.4.1 Type 1**

Table 4.1: Diabetes Type 1 Accuracy

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

**4.4.2 Type 2**

Table 4.2: Diabetes Type 2 Accuracy

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

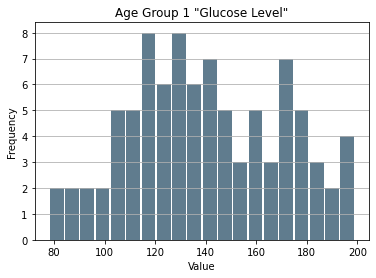
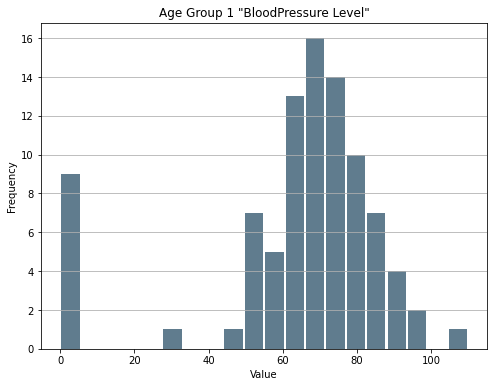
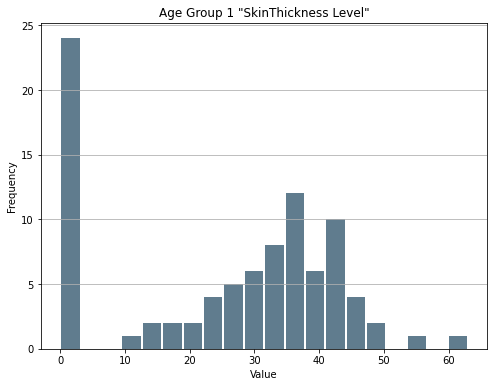
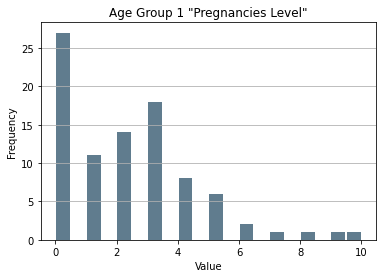
**4.4.3 Type 3**

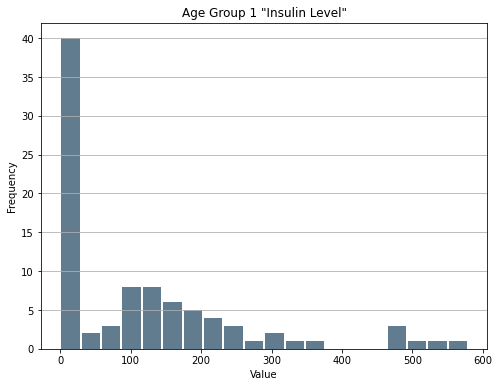
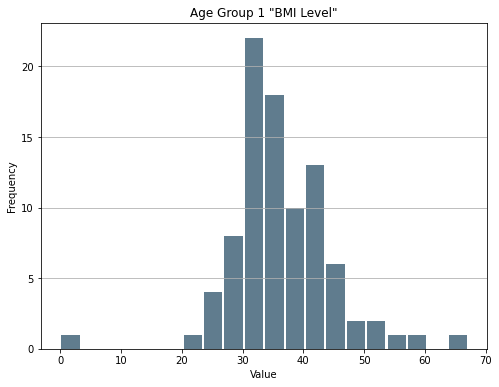
Table 4.3: Diabetes Type 3 Accuracy

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

**4.5 Age Group Symptom Analysis**

**4.5.1 Age Group 1**

********

****

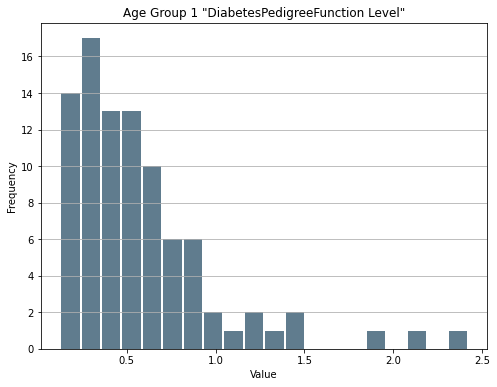
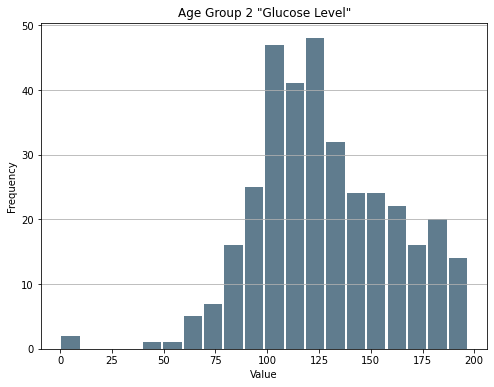
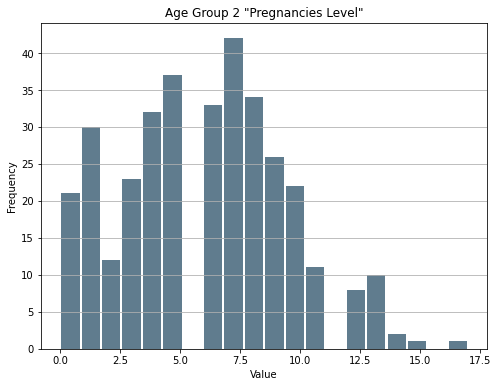
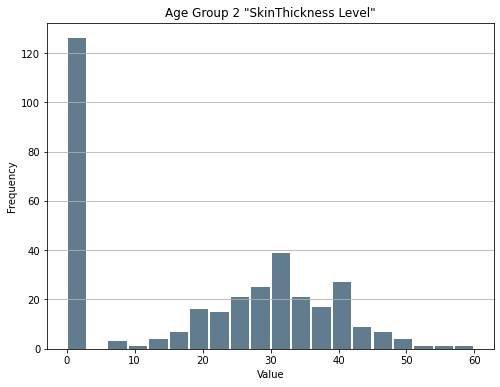
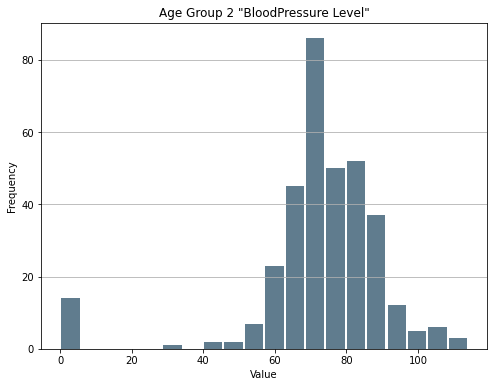
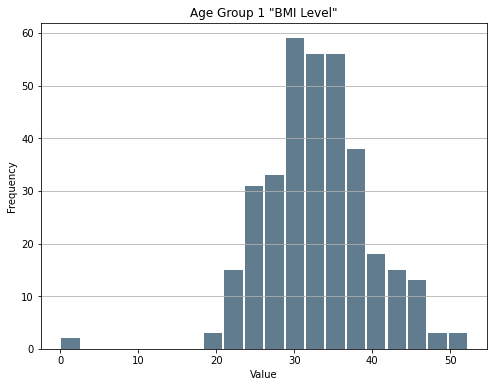
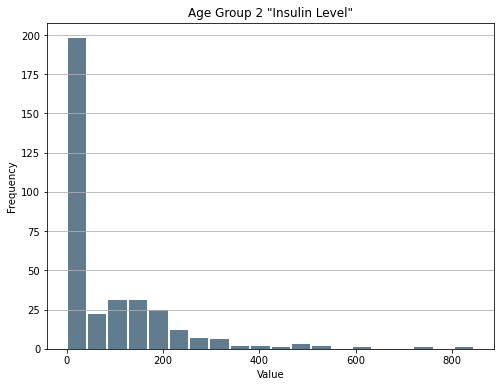


Fig 4.26: Age Group 1 all Symptoms

**4.5.2 Age Group 2**

****

****

****

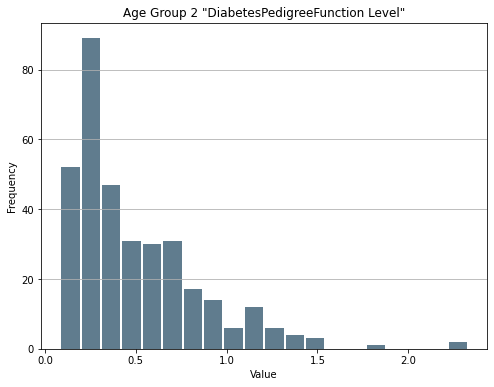
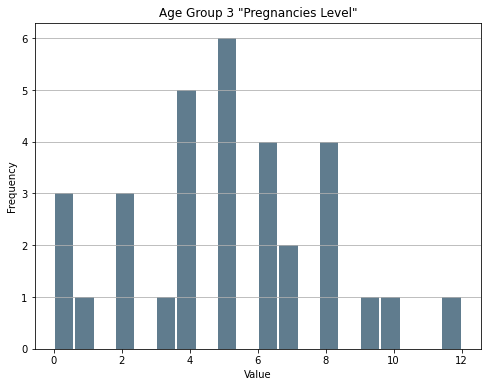
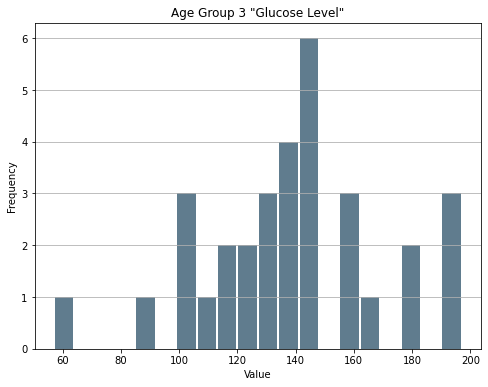
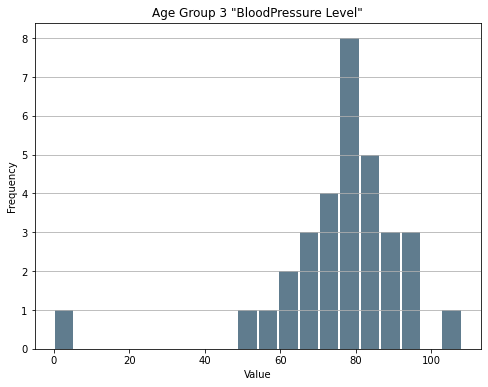
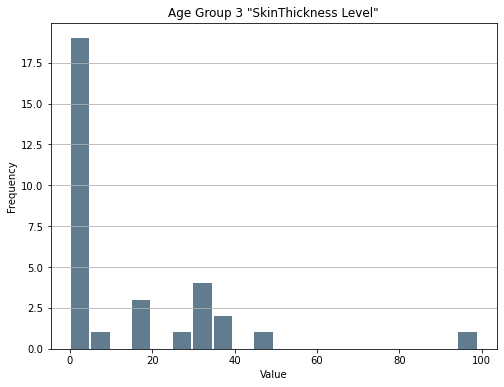
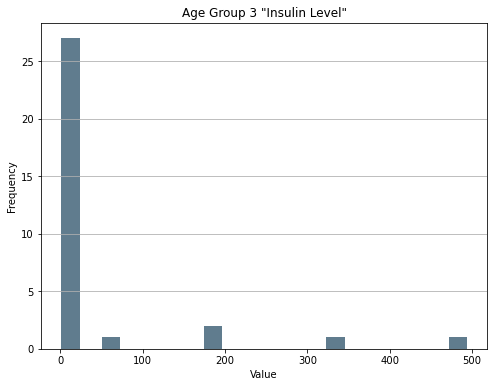
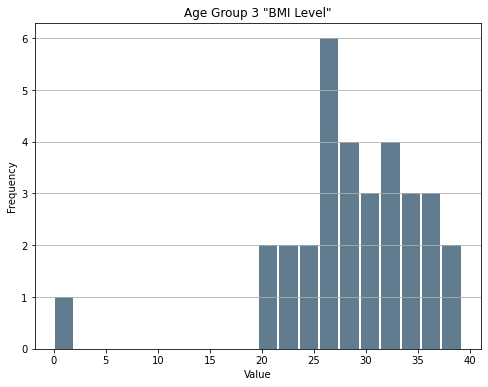
****

Fig 4.27: Age Group 2 all Symptoms

**4.5.3 Age Group 3**

** **

** **

** **

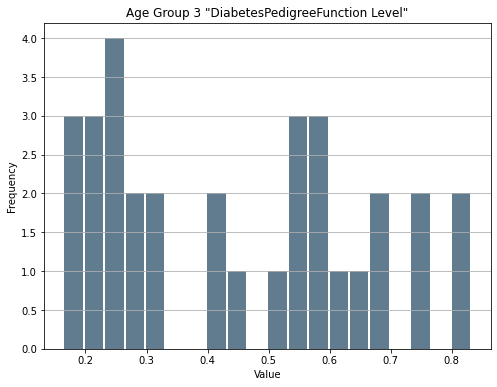
****

Fig 4.28: Age Group 3 all Symptoms

**CHAPTER 5**

**DISCUSSION**

**CHAPTER 5**

**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.

**CHAPTER 6**

**CONCLUSION**

**CHAPTER 6**

**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%.

**References:**

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.
10. Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K. and Jonkman, M., 2021. A Comparative Study of Different Machine Learning Tools in Detecting Diabetes. Procedia Computer Science, 192, pp.467-477
11. Selvi, R.T. and Muthulakshmi, I., 2021. Modelling the map reduce based optimal gradient boosted tree classification algorithm for diabetes mellitus diagnosis system. Journal of Ambient Intelligence and Humanized Computing, 12(2), pp.1717-1730.
12. Orabi, K.M., Kamal, Y.M. and Rabah, T.M., 2016, July. Early predictive system for diabetes mellitus disease. In Industrial Conference on Data Mining (pp. 420-427). Springer, Cham.
13. Kumar, D.A. and Govindasamy, R., 2015. Performance and evaluation of classification data mining techniques in diabetes. International Journal of Computer Science and Information Technologies, 6(2), pp.1312-1319
14. Alkaragole, M.L.Z. and Kurnaz, A.P.S., 2019. COMPARISON OF DATA MINING TECHNIQUES FOR PREDICTING DIABETES OR PREDIABETES BY RISK FACTORS.
15. Woldemichael, F.G. and Menaria, S., 2018, May. Prediction of diabetes using data mining techniques. In 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 414-418). IEEE.
16. Rao, L.G., Rao, A.M., Rao, S.P., Rao, R., Bhavsar, J.H., Liu, W.Y., Igyarto, Z. and Martinsen, B.J., 2021. Outcomes After Coronary Orbital Atherectomy at Centers Without On-Site Surgical Backup: Diabetics Versus Non-Diabetics and Impact of Access Site. Cardiovascular Revascularization Medicine, 30, pp.20-25.
17. Kumari, A. and Pandey, B.K., 2021. Assessment of the intraocular pressure of diabetics and non-diabetics: a comparative study. European Journal of Molecular & Clinical Medicine, 7(7), pp.4835-4840.
18. Youn, S.Y., Choi, M.H., Kim, D.H., Lee, Y.J., Huisman, H., Johnson, E., Penzkofer, T., Shabunin, I., Winkel, D.J., Xing, P. and Szolar, D., 2021. Detection and PI-RADS classification of focal lesions in prostate MRI: Performance comparison between a deep learning-based algorithm (DLA) and radiologists with various levels of experience. European Journal of Radiology, 142, p.109894.
19. Dino, H.I. and Abdulrazzaq, M.B., 2020. A Comparison of Four Classification Algorithms for Facial Expression Recognition. Polytechnic Journal, 10(1), pp.74-80.
20. Ebrahimi, F., Asemi, A., Nezarat, A. and Ko, A., 2021. Developing a mathematical model of the co-author recommender system using graph mining techniques and big data applications. Journal of Big Data, 8(1), pp.1-15.
21. Prabhusundhar, P., Poornima, N.V. and Srinivasan, B., 2021. Optimized Feed Forward Neural Network For Classification Of Diabetes In Big Data Environment. European Journal of Molecular & Clinical Medicine, 8(2), pp.2510-2528.
22. Baecker, L., Dafflon, J., Da Costa, P.F., Garcia‐Dias, R., Vieira, S., Scarpazza, C., Calhoun, V.D., Sato, J.R., Mechelli, A. and Pinaya, W.H., 2021. Brain age prediction: A comparison between machine learning models using region‐and voxel‐based morphometric data. Human brain mapping, 42(8), pp.2332-2346.