# Diabetes Prediction using Machine Learning

**Neha Thakur[1], Virendra Singh[2]**

*Computer Science & Engineering, SAGAR Institute of Research & Technology, SAGE University, Indore*

*[1]nehathakurcs96@gmail.com,[2]virendra.cse@sageuniversity.in*

**Abstract**-Now days from health care industries large volume of data is generating. It is necessary to collect, store and process this data to discover knowledge from it and utilize it to take significant decisions.Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. Insulin, a hormone made by the Pancreas, helps glucose from food to get into your cells to be used for energy. Sometimes your body doesn't make enough or any insulin or doesn't use insulin well;glucose then stays in your blood and doesn't reach to your cells, which turns into diabetes.The objective of this research is to make use of various Machine Learning Algorithms,to predict the type2 diabetes. The Pima Indians Diabetes Datasets (PIDD) has been used to predict diabetes disease. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.This paper discusses the Machine Learning approach for the prediction of diabetes. A performance comparison between different Machine Learning Algorithms i.e. Predictive Modelling, Decision tree, Logistic regression, Gradient Boosting is done. The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity.

Keywords- **Machine learning, Predictive Modelling, Accuracy, Machine Learning, PIDD, Diabetes Prediction**

## I. INTRODUCTION

Diabetes may be a chronic, metabolic disease characterized by elevated levels of glucose (or blood sugar), which leads over time to serious damage to the center, blood vessels, eyes, kidneys and nerves. the foremost common is type2 diabetes, usually in adults, which occurs when the body becomes immune to insulin or doesn't make enough insulin. Type2 diabetes is accounting around 90% of all diabetes cases.

In 2019, approximately 463 million adults (20-79 years) were living with diabetes worldwide; by 2045 this may rise to 700 million[10]. The proportion of individuals with type2 diabetes is increasing in most countries.79% of adults with diabetes were living in low and middle income countries.1 out of 5 of peoples who are above 65 years old has diabetes.1 in 2 (232 million) people with diabetes were undiagnosed. Diabetes caused 4.2 million deaths. Diabetes caused a minimum of USD 760 billion dollars in health expenditure in 2019; 10% of total spending on adults. More than 1.1 million children and adolescents live with type1 diabetes. over 20 million live births (1 in 6 live births) are laid low with diabetes during pregnancy.374 million people are at increased risk of developing type2 diabetes.

Machine learning (ML) is an application of computing (AI) wherein the system looks at observations or data, like examples, direct experience, or instruction, figures out patterns in data, and predicts events within the future supported by the examples that we offer. Machine learning is more and more used across industries for various reasons: the vast amount of knowledge is being captured and made available digitally; the processing of huge amounts of knowledge has become cost-effective because of the increased computing power now available at affordable prices; and various open-source frameworks, toolkits, and libraries are available that may be accustomed build and execute ML applications[1].

Machine learning is a wonderful technology for predicting data within the medical field. Machine learning automatically predicts a large amount of knowledge in a very very short time with minimum interaction with patients. Machine learning[3] deals with the event of technologies that permit machines to be told. The challenge is to form algorithms that will take a gaggle of patterns (on a broader range, the present knowledge) and automatically make new inferences from the initial information, with or without human intervention.

## II. Types of diabetes & symptoms

The three main types of diabetes are:

A. Type 1 Diabetes
B. Type 2 Diabetes
C. Gestational Diabetes

### A. Type 1 diabetes

If you have got sort one diabetes disease, your exocrine gland doesn't build hypoglycemic agent or makes very little hypoglycemic agent. The hypoglycemic agent is additionally a secretion that helps aldohexose enter the cells in your body wherever it's going to be used for energy. While not a hypoglycemic agent, aldohexose can't get into cells and builds up inside the blood. High glucose is damaging to the body and causes several of the symptom and complications of polygenic disease. sort one polygenic disease could also be a smaller quantity common than sort a pair of polygenic disease or so 5-10% of individuals with the polygenic disease has sort one polygenic disease. Currently, no one is aware of a way to forestall sort one polygenic disease, however, it's typically managed by following your doctor's recommendations for living a healthy fashion, managing your aldohexose, obtaining regular health checkups, and obtaining polygenic disease self-management education, and support. Testing for sort one, polygenic disease an easy diagnostic assay can enable you to understand if you have got the polygenic disease. If you've gotten your glucose tested at a health honest or pharmacy, follow-up at a clinic or doctor's workplace to form certain the results square measure correct.

### B. Type 2 diabetes

In a person with sort a pair of polygenic disorder, the body is unable to use hypoglycemic agent properly and this condition is termed hypoglycemic agent resistance. The duct gland or duct gland 1st creates further hypoglycemic agents for this. However, over time, it doesn't build enough to stay glucose at a traditional level. though the precise trigger for this condition isn't familiar, it should be a result of a mixture of the sort a pair of polygenic disorder Causes. Some triggers are also genetically susceptible to the condition of Obesity and kind a pair of Diabetes: folks with a case history of avoirdupois are in danger of developing hypoglycemic agent resistance and polygenic disorder. folks that are rotund have exaggerated pressure on their body's ability to use hypoglycemic agents to regulate glucose levels. this will cause a sort of a pair of polygenic disorder. A lot of fat an individual has in his body, a lot of resistant his cells are. way factors conjointly play a significant role during this.

The exact cause of type 2 diabetes is unknown. Contributing factors may include:

- Genetics
- Lack of exercise
- Being overweight

There may also be other health factors and environmental reasons.

### C. Gestational diabetes

The gestational polygenic disorder may be a condition within which the blood glucose (sugar) level will increase in a very pregnant lady. People who don't have already have sugar conjointly suffer from it. This happens once the pregnant female body isn't manufacturing enough internal secretion referred to as hormone.

Low hormone levels will cause physiological state polygenic disorder the physiological state hereditary condition is hyperglycemia with blood glucose values on prime of ancient but below those diagnostic of a hereditary condition, occurring throughout gestation. Girls with the physiological state polygenic disorder are at associate hyperbolic risk of complications throughout gestation and delivery. They and their children area unit at hyperbolic risk of a sort a try of the hereditary condition at intervals the long run. Physiological state hereditary condition is diagnosed through ante partum screening, rather than through reportable symptoms.

## General symptoms of diabetes include:

- Excessive thirst and hunger
- Frequent urination
- Drowsiness or fatigue
- Dry, itchy skin
- Blurry vision
- Slow-healing wounds

## III. Glucose Monitoring Systems:-

Blood glucose observance reveals individual patterns of glucose changes, and helps within the coming up with meals, activities, and at what time of day to require medications. Also, testing permits for fast response to high glucose (hyperglycemia) or low glucose (hypoglycemia). For instance, information from polygenic disease management systems like aldohexose observance devices and hypoglycemic agent dose regimens square measure transmitted to the cloud. the Adviser will then access the additive information from the cloud and learn the patient's distinctive habits and wishes are employing a proprietary rule. The patterns derived from analysis square measure documented to supply automatic recommendations for hypoglycemic agent dosing.

Machine learning ways square measure wide utilized in predicting polygenic disease, and that they get desirable results. Recently, various algorithms square measure accustomed to predict polygenic disease, as well as the normal machine learning methodology like a support vector machine (SVM), call tree (DT), supply regression then on call tree is one in all well-liked machine learning ways in the medical field, that has grateful classification power. Random forest generates several call trees. The 3 main styles of polygenic disease square measure sort one polygenic disease (T1D), sort two DM (T2D), and physiological state DM (GDM). Since 2000, the International polygenic disease Federation (IDF) has reportable the national, regional, and world incidence of polygenic disease. In 2009, it was calculable that 285 million individuals had polygenic disease (T1D and T2D combined), increasing to 366 million in 2011, 382 million in 2013, 415 million in 2015, and 425 million in 2017.

The age- and sex-stratified polygenic disease prevalence was calculated for every country, accounting for polygenic disease prevalence variations in urban and rural areas. Urban to rural polygenic disease prevalence ratios were updated victimization the weighted average of the ratios in numerous information sources within the Israeli Defense Force|force} Regions and UN agency financial gain group, wherever the weights were the study scores calculated victimization the AHP classification system. Logistical regression was performed to come up with smoothened age- and sex-specific prevalence estimates for a 5-year age team for adults aged 20–79 years. The regression used age (as the center of every age-group) and therefore, the quadratic older than separate freelance variables for every sub-group (sex and urban/rural area) if on the market. The quadratic age term was utilized in the regression to permit a come by polygenic disease prevalence for the oldest age-groups to account for mortality.

The global organization (UN) population estimates for 2019, for every one of the 211 countries and territories were accustomed to generate national estimates. The number of individuals with polygenic disease in every of the seven Israeli Defense Force|force} Regions and every UN agency financial gain cluster was calculated by aggregating the number of individuals with the polygenic disease for every country among the individual Israeli Defense Force Region and UN agency financial gain group. World estimates were calculated by aggregating the whole variety of individuals with the polygenic disease for every country, with population denominators obtained for every country and territory from the global organization Population Division (UNPD). UNPD doesn't give age- and sex-stratified population information for countries and territories with populations smaller than ninety,000. In these cases, age- and sex-specific regional level population information was accustomed to calculate age- and sex-specific population estimates for the little countries and territories.

## IV. Literature Survey

Diabetes a non-communicable malady is resulting in semi-permanent complications and heavy health issues. A report from the planet Health Organization [30] addresses polymeric disorder and its complications that impact individuals physically, financially, economically over families. The survey says regard one.2 million deaths because of the uncontrolled stage of health cause death. About 2.2 million deaths occurred because of the danger factors of polygenic disorder sort of vessel and different diseases.

Diabetes [31] is an associate degree disorder caused because of the extended level of sugar obsession within the blood. this paper mentioned numerous classifiers, call network is planned that uses the AdaBoost formula with call Stump as a base classifier for classification. Moreover, Support Vector Machine, Naive Bayes and call Tree to has in addition, dead as a base classifier for AdaBoost calculation for accuracy confirmation. The accuracy got for AdaBoost calculation with decision stump as a base classifier is eighty.72%, which is additional noteworthy contrasted therewith of Support Vector Machine, Naive Bayes, and call Tree.

Still, machine learning lends itself to some processes higher than others. Algorithms will offer immediate profit to disciplines with processes that area unit consistent or standardized. Also, those with massive image datasets, like radiology, cardiology, and pathology, area unit robust candidates. Machine learning is often trained to appear at pictures, establish abnormalities, and purpose to areas that require attention, so rising the accuracy these processes. Long-term, machine learning can profit the family practicing or medical specialist at the side. Machine learning offers AN objective opinion to enhance potency, responsibility, and accuracy.

At Health Catalyst, we tend to use a proprietary platform to research knowledge and loop it back in real-time to physicians to help in clinical deciding. At a similar time, a medico sees a patient and enters symptoms, data, and take a look at results into the EMR, there's machine learning behind the scenes watching everything that patient, and prompting the doctor with helpful info for

creating a designation, ordering a take a look at, or suggesting a preventive screening. Long-term, the capabilities can touch all aspects of drugs as we tend to get a lot of usable, higher integrated knowledge. We'll be able to incorporate larger sets of knowledge which will be analyzed and compared in real-time to produce all types of data for the supplier and patient

## V. Algorithms

Since there ar various algorithms for machine learning, it's impractical to use all of them for analysis. For this analysis paper, we have a tendency to ar about to be practice four of them call tree, provision regression, Gradient Boosting. Predictive analytics may be represented as a branch of advanced analytics that's utilized within the creating of predictions regarding unknown future events or activities that result in choices.
It is a discipline that 493 minimize varied techniques as well as 493 inimizi, data processing, and statistics, yet as computer science (AI) (such as machine learning) to guage historical and period of time knowledge and create predictions regarding the long run. These predictions provide a novel chance to examine into the long run and determine future trends in patient care each at a personal level and at a cohort scale. Predictive analytics is predicated on the logic that's drawn from theories developed by humans to suit a hypothesis (supervised learning). A collection of rules And processes ar developed into a formula that undertakes calculations and is understood as an formula. Prophetical Analytics may be supported unattended learning that doesn't have a guiding hypothesis and uses an formula to hunt patterns and structure in knowledge and cluster them into teams or insights. In unattended learning, the machine might not apprehend what it's trying to find however because it processes the info it starts to spot complicated processes and patterns that a personality's could ne'er have known and thus will add important worth to researchers trying to find one thing new. Each supervised and unattended prophetical modeling ar valid analytical tools to use during a all-around application ofthose technologies. Predictive analytics is increasing in its application and has been terribly helpful in varied industries together with producing, marketing, law, crime, fraud detection, and health care. The health care sector, with its several stakeholders, stands to be a key beneficiary of prophetical analytics, with the advanced technology being recognized as AN integral a part of health care service delivery. This paper can cross-check the assorted ethical and moral hazards that require to be navigated by government agencies, doctors, and first caregivers once investing the potential that prophetical analytics has

With new technologies come back new risks. This paper can measure varied eventualities within the use of prognosticative analytics with a specific 493minimizing in service delivery among health care. A risk rising for prognosticative analytics includes the 493minimizing493ion of information that presents an incredible risk in terms of security and integrity of the information. Given the increasing quantity of information that's usually hold on within the cloud or otherwise accessible via the net, there's the persistent threat of hacking from people with malicious intent. There are moral problems to be thought-about, given the role the cloud technology plays in prognosticative analytics and therefore the overall outcome.3 during this article, we tend to 493minimizing in the moral problems and leave security of information and therefore the cloud to a different time

### ❖ Data Cleaning

The next part of the machine learning work flow is information cleanup. thought-about to be one in every of the crucial steps of the advancement, as a result of it will create or break the model. there's an adage in machine learning "Better information beats admirer algorithms", that suggests higher information offers you higher ensuing models .

There are many factors to contemplate within the information cleanup method.

- Duplicate or immaterial observations.
- Bad labeling of knowledge, same class occurring multiple times.
- Missing or null information points.
- Unexpected outliers.

**Missing or Null Data points**

We can notice any missing or null knowledge points of the information set (if there's any) victimisation the subsequent pandas operate.

**diabetes.isnull().sum()**

**diabetes.isna().sum()**

We can observe that there aren't any knowledge points missing within the knowledge set. If there have been any, we should always cope with them consequently.

- **Statistical report of Pima Indian Dataset**.

| Attribute No. | Attribute | Variable Type | Range |
|---|---|---|---|
| A1 | Pregnancy (No of times pregnant) | Integer | 0–17 |
| A2 | Plasma Glucose (mg/dL) | Real | 0–199 |
| A3 | Diastolic Blood Pressure (mm Hg) | Real | 0–122 |
| A4 | Triceps skinfold (mm) | Real | 0–99 |
| A5 | Serum Insulin (mu U/ml) | Real | 0–846 |
| A6 | Body mass index (kg/m$^2$) | Real | 0–67.1 |
| A7 | Diabetes Pedigree | Real | 0.078–2.42 |
| A8 | Age (years) | Integer | 21–81 |
| Class | | Binary | 1 = Tested Positive for diabetes<br>0 = Tested Negative for diabetes |

❖ **Phase 1 — Data Exploration**

When encountered with an information set, initial we must always analyze and "get to know" the info set. This step is important to acquaint with the info, to realize some understanding of the potential options and to ascertain if information improvement is required.

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 2 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 3 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 4 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

**Table 1 — Diabetes data set**

First, we are going to import the required libraries and import our information set to the Jupyter notebook. we are able to observe the mentioned columns within the information set

❖ **Data Visualization**

Another vital feature within the knowledge distribution is that the lopsidedness of every category. Data visualization helps to check however {the knowledge |the info |the information} skewers and conjointly what reasonably data correlation we've. The dataset distribution of every feature is shown below in figure 3.5. This is often a bar chart. A bar chart is a correct graphical illustration of the distribution of numerical data. It's AN estimate of the chance distribution of a never-ending variable. Histograms are an excellent way to get to understand your data. They permit you to simply see wherever an outsized and a bit quantity of the data are found. In short, the bar chart consists of an x-axis and a y-axis, wherever the y-axis shows however often the values on the x-axis occur within the data.

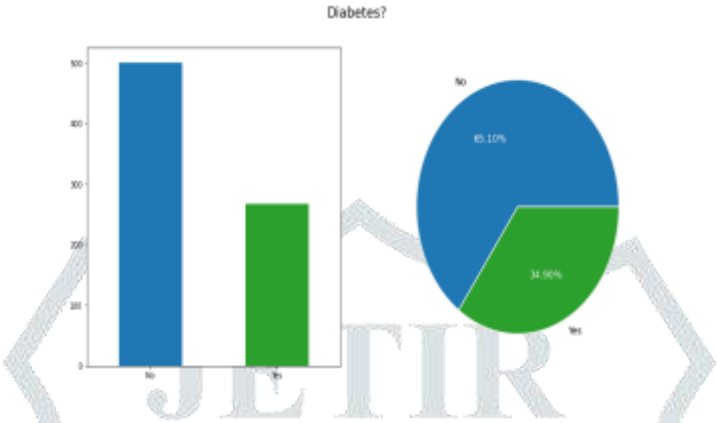


**Figure1- Data visualization result**

❖ **Violin Plots**

Variables at intervals a dataset will be connected for numerous reasons. It will be helpful in data analysis and modeling to higher perceive the relationships between variables. The applied math relationship between 2 variables is spoken as their correlation. A correlation may well be positive, which means each variable move within the same direction, or negative, which means that once one variable's worth will increase, the opposite variables' values decrease. Correlation also can be neural or zero, which means that the variables are unrelated

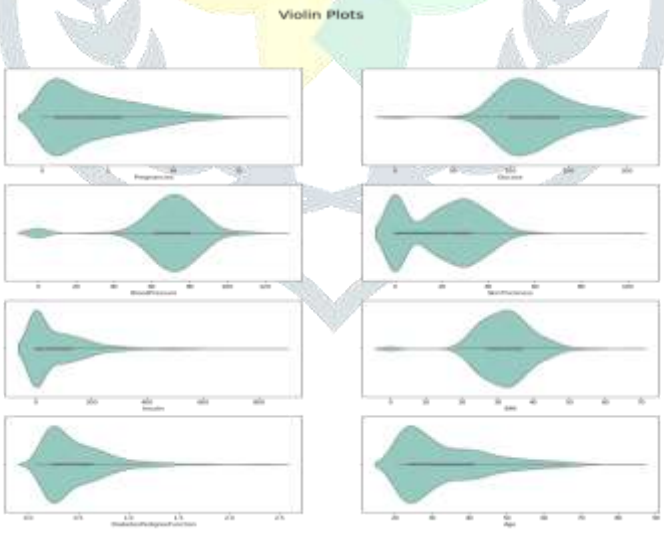


**Figure 1- Violin plot**

Dist Plot helps U.S. to flexibly plot a univariate distribution of observations. It is used essentially for univariant set of observations and visualizes it through a bar chart i.e. only 1 observation and thence we decide one specific column of the dataset

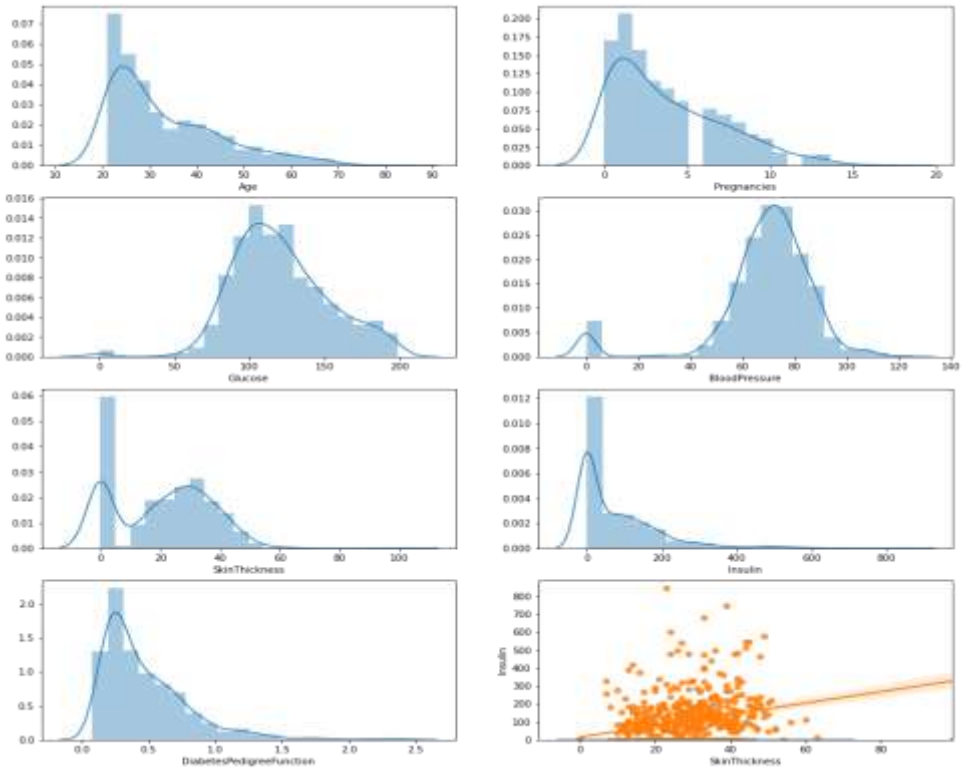


**Figure2 Dist plot**

- **Correlation between features**

Variables at intervals a dataset will be connected for numerous reasons. It will be helpful in data analysis and modeling to higher perceive the relationships between variables. The applied math relationship between 2 variables is spoken as their correlation. A correlation may well be positive, which means each variable move within the same direction, or negative, which means that once one variable's worth will increase, the opposite variables' values decrease. Correlation also can be neural or zero, which means that the variables are unrelated

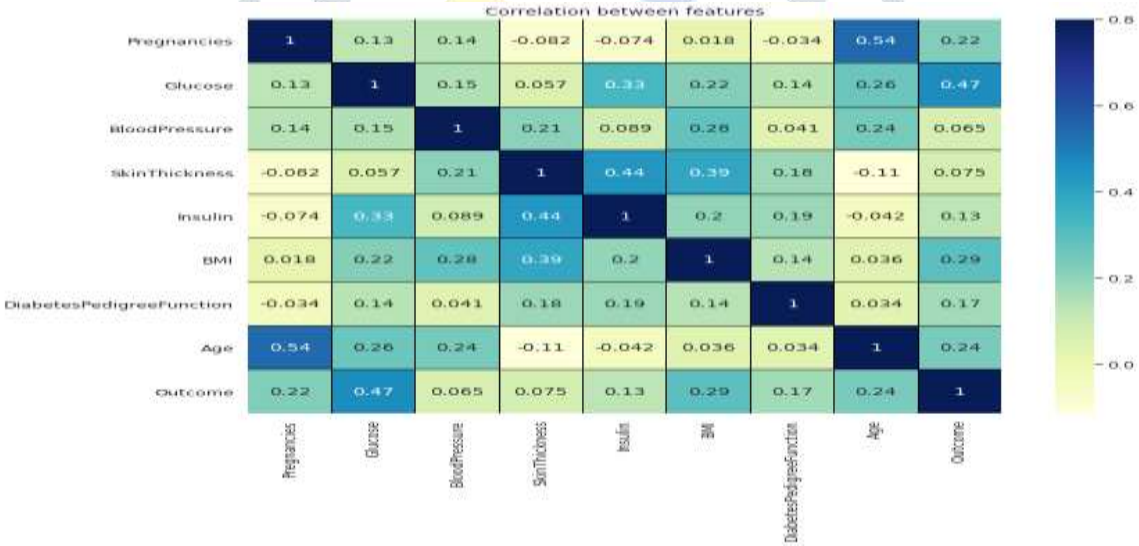


**Table 2- Correlation between feature**

- **Training Accuracy of Decision Tree**

The decision tree builds regression or classification models within the style of a tree structure. It breaks down a dataset into smaller and smaller subsets whereas at identical times an associated decision tree is incrementally developed. the ultimate result's a tree with decision nodes and leaf nodes. a choice node (e.g., Outlook) has 2 or additional branches (e.g., Sunny, Overcast, and Rainy), every representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a choice on the numerical target. The uppermost decision node in an exceedingly tree that corresponds to the most effective predictor referred to as the root node. Decision trees will handle each categorical and numerical information

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

| Play Golf | |
|---|---|
| Yes | No |
| 9 | 5 |

**Entropy(PlayGolf)** = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

Where **S → Current state, and Pi → Probability of an event *i* of state S or Percentage of class *i* in a node of state S**

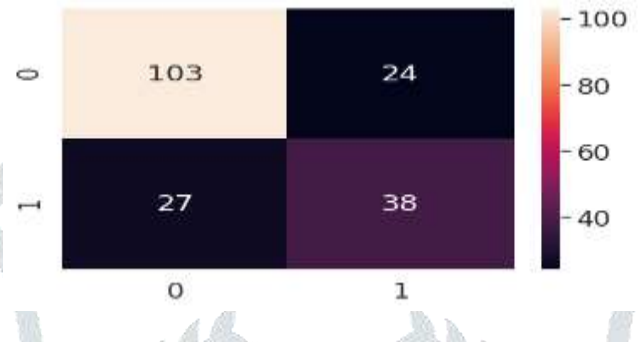


**Figure 3- Training accuracy of decision tree**

**Training Accuracy of Logistic Regression**

Logistic regression is that the acceptable multivariate analysis to conduct once the variable quantity is binary. Like all regression analyses, the logistical regression could be a prophetical analysis. logistical regression is employed to explain information and to clarify the connection between one dependent binary variable and one or additional nominal, ordinal, interval, or ratio-level freelance variables.

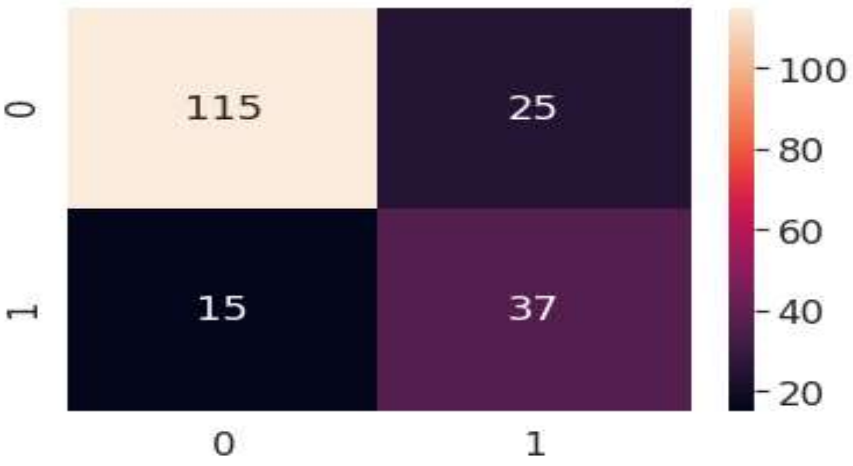


**Figure 4- Training accuracy of logistic regression**

❖ **Training Accuracy of Gradient Boosting**

Gradient boosting may be a machine learning technique for regression and classification problems, that produces a prediction model at intervals the design of associate ensemble of weak prediction models, typically call trees. It builds the model in a {very} very stage-wise fashion like different boosting ways that do, associate degreed it generalizes them by allowing improvement of a arbitrary differentiable loss to work

Step1- we assume an alpha(t)

Step2- get a weak classifier h(t)

Step3-update the population distribution for the next step

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Where,

$$Z_t = \sum_{i=1}^{m} D_t(i)\exp(-\alpha_t y_i h_t(x_i))$$

Step4- Use the new population distribution to again find the next learner

Step5- Iterate step1- step4 until hypothesis is found which can improve further

Step6- Take a weighted average of the frontier using all the learners used till now. Weights are simply the alpha values.

$$\alpha_t = \tfrac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

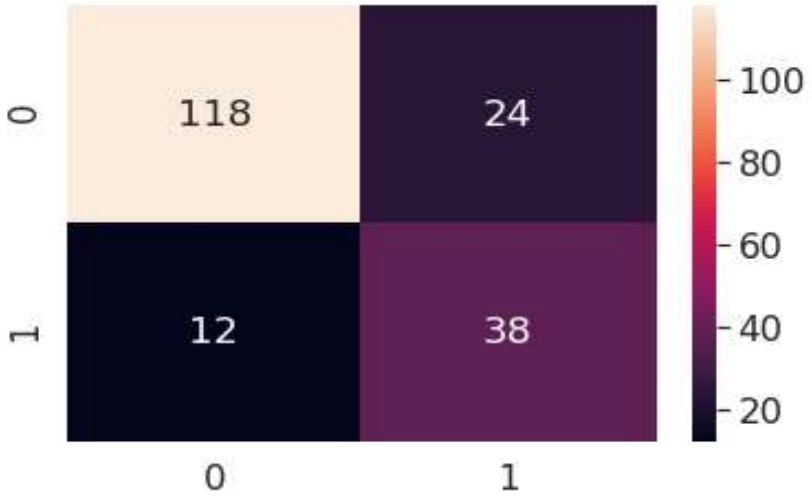


**Figure 5- training accuracy of gradient boosting**

The results obtained are mentioned victimization the various prophetical approaches and so these results are compared between each other. the standard of the prediction models is assessed by mathematical metrics that quantify the error between expected symptom events and therefore the actual ones. Specifically, predictions were evaluated with regard to the accuracy, specificity, and sensitivity. Accuracy is the proportion of testing set examples that are properly classified by the model. Sensitivity refers to the proportion of instances with symptoms and is assessed in and of itself. On the opposite hand, specificity refers to the flexibility of the model to properly diagnose nonhypoglycemic events in and of itself.

- **Accuracy**
- **Specificity**
- **Sensitivity**

Where actuality positives (TP) and true negatives (TN) are correct classifications. Specifically, TP refers to the number of instances that were expected as symptoms, once they are in and of itself. Likewise, Volunteer State represents a variety of instances that were expected as non-hypoglycemia, once they are in and of itself. A false positive (FP) is once the end result is incorrectly expected as positive (hypoglycemia), however, once it's really negative (non-hypoglycemia). Contrary, a false negative (FN) is once the seventy outcomes is incorrectly expected as negative (non-hypoglycemia), however, once it's really positive (hypoglycemia), so decreasing FN is that the focal goal. Obviously, we tend to aim to realize the very best sensitivity, as we tend to don't wish to miss several symptom events. However, the next sensitivity can typically lead to a lower specificity, and contrariwise. thus it's a trade-off between sensitivity and specificity, wherever terribly low specificity can find yourself in several false alarms (1 – specificity); thus, this can cut back the reliableness of the model to patients

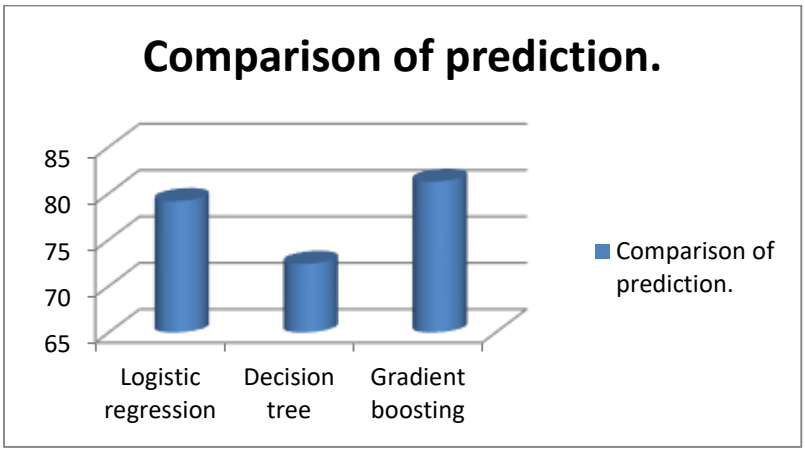


**Figure6- compression between algorithms**

## VI. Conclusion

We have delineated a machine learning approach to predicting illness} as in early-stage diabetes is an incredibly dangerous disease that causes premature death. Study shows that it's potential for a few folks to reverse it. Through diet changes and weight loss, you'll be able to reach and hold traditional glucose levels while not medication. Therefore it becomes nearly obligatory to use machine learning to predict illness} disease. During this paper, the author foresaw diabetes by using new Machine Learning Algorithms and decide. The initial symptoms of diabetes. The approach is compared with already existing solutions i.e. prognosticative Modeling, decision Tree, logistic Regression, and Gradient Boosting. The results of our methodology show higher performance in terms of potency and accuracy. we've got applied several ml algorithms on the diabetes dataset and therefore the performance of these algorithms is analyzed. The accuracy of logistical regression 79.16%, a decision tree was 72.39% lowest accuracy and gradient boosting was 81.25% highest accuracy.

## VII. References

[1]. Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908.

[2]. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003

[3]. MehrbakhshNilashi, Othman Bin Ibrahim, Abbas Mardani, Ali Ahani and Ahmad Jusoh, 2016.A softcomputing approach for diabetes disease classification, Health Informatics Journal University Technologic Malaysia, Malaysia.

[4]. Ioannis Kavakiotis.2017 Machine Learning and Data Mining Methods in Diabetes Research

[5]. DeeptiSisodiaa, Dilip Singh Sisodia ICCIDS 2018. Prediction of Diabetes using Classification Algorithms International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

[6]. Marina Basina, MD on oct.4,2018, Helthline

[7]. Blog Machine Learning 6 May 2020 expert system team. https://expertsystem.com/machine-learning-definition/

[8]. Nagesh Singh Chauhanis a Data Science enthusiast. 2020 KDnuggets.

[9]. Mackenzie Mitchell in towards data science nov. 14, 2019, selecting the correct Predictive Modelling Technique.

[10]. World health organization,15 may 2020, Diabetes,

[11]. Pima Indian Diabetes Database, URL: www.ics.uci.edu/~mlearn/MLRepository.html.

[12]. Tavishsrivastava ,September 11,2015. Learn Gradient Boosting Algorithm for better predictions

[13]. Amatul Zehra1 , Tuty Asmawaty1 , M.A M. Aznan2, A comparative study on the pre-processing and mining of Pima Indian Diabetes Dataset

[14]. Hang Lai, Huaxiong Huang, 2019, Predictive models for diabetes mellitus using machine learning techniques, BMC