Synopsis on

**“Diabties Prediction using ML and Data Science”**

**Group No.-** ----------------------------

**Project Guide-** -----------------------------

**Submitted By-** **Submitted To-**

Akshita Agarwal (1601410012) --------------------

Lavi Thukral (1601410060)

Priyanshi Kotnala (1601410081)



Department of Computer Science and Technology

**Shri Ram Murti Smarak College of Engineering and Technology, Bareilly**

**Table of Contents**

1. Title...…………………………………………………………………………………………………..3
2. Introduction……………………………………………………...………………………………....4
3. Objective…………………………………………………….......………………………………….6
4. Requirements……………………………………………………..………………………………..7
   1. Software
   2. Hardware
5. Methodology……………………………………………………..…………………………………8
6. Modules/Functionalities.…………………………………………………………………….10
7. Application…………………………………………………….....………………………………..11
8. Limitations…………………………………………………….....………………………………..12
9. References……………………………………………………......……………………………….13

**Title**

A proper project title describes the whole assignment in one sentence.  It helps the team to refer the Project with the assigned Name. Project titles makes you to understand the main goal of the Project work and deliverables. A project name can be the product name when it completed.

There are few types of Project Titles, titled based on the type of the Project.

* Encrypted Titles: These are the secrete titles, organizations do not want the developing team to know about the actual titles
* Temporary Tiles: Temporary titles are assigned to project if the actual title is not yet finalized
* Descriptive titles: These are the long titles and they represent the actual project in clear descriptive manner for easy understanding
* Sort Titles: Sort Project tiles are very creative and become the final product names in most of the time

When you start creating a new project, it is important to have a suitable Project Title for your assignment. It makes easy to identify the project in the Project Management Process. An ideal Project title will have the following characteristics.

* Represent the Goals of the Organization
* Team Objectives and Customer needs
* Sort name of the Project Work and Deliverables
* Easy to understand and pronounce
* Creative and Professional
* Easy to remember and recognize

Hence the title of our project is **“Diabties Prediction using Machine Learning and Data Science”**, which is descriptive and apparently reflects the technology used to buid this project.

**Introduction to the project**

A Diabetes is a common chronic disease and poses a great threat to human health. The characteristic of diabetes is that the blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects, or both (Lonappan et al., 2007). Diabetes can lead to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves (Krasteva et al., 2011).

Diabetes can be divided into two categories:

* Type 1 diabetes (T1D)
* Type 2 diabetes (T2D)

Patients with Type 1 diabetes are normally younger, mostly less than 30 years old. The typical clinical symptoms are increased thirst and frequent urination, high blood glucose levels. This type of diabetes cannot be cured effectively with oral medications alone and the patients are required insulin therapy.

Type 2 diabetes occurs more commonly in middle-aged and elderly people, which is often associated with the occurrence of obesity, hypertension, dyslipidemia, arteriosclerosis, and other diseases.

With the development of living standards, diabetes is increasingly common in people’s daily life. Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worthy studying. In medicine, the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels. The earlier diagnosis is obtained, the much easier we can control it.

Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors. For machine learning method, how to select the valid features and the correct classifier are the most important problems.

Recently, numerous algorithms are used to predict diabetes, including the traditional machine learning method such as:

* Support Vector Machine (SVM)
* Decision Tree (DT)
* Logistic Regression

and so on...

Polat and Günes (2007) distinguished diabetes from normal people by using principal component analysis **(PCA)** and neuro fuzzy inference. Yue et al. (2008) used quantum particle swarm optimization **(QPSO)** algorithm and weighted least squares support vector machine **(WLS-SVM)** to predict type 2 diabetes Duygu and Esin (2011) proposed a system to predict diabetes, called **LDA-MWSVM**. In this system, the authors used Linear Discriminant Analysis (LDA) to reduce the dimensions and extract the features. In order to deal with the high dimensional datasets, Razavian et al. (2015) built prediction models based on logistic regression for different onsets of type 2 diabetes prediction. Georga et al. (2013) focused on the glucose, and used support vector regression **(SVR)** to predict diabetes, which is as a multivariate regression problem. Moreover, more and more studies used ensemble methods to improve the accuracy. Ozcift and Gulten (2011) proposed a newly ensemble approach, namely **rotation forest**, which combines 30 machine learning methods. Han et al. (2015) proposed a machine learning method, which changed the SVM prediction rules.

With the development of living standards, diabetes is increasingly common in people’s daily life. Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worthy studying. In medicine, the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels (Iancu et al., 2008; Cox and Edelman, 2009; American Diabetes Association, 2012). The earlier diagnosis is obtained, the much easier we can control it. Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors (Lee and Kim, 2016; Alghamdi et al., 2017; Kavakiotis et al., 2017). For machine learning method, how to select the valid features and the correct classiﬁer are the most important problems.

Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world’s diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health. The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used. Machine learning methods are widely used in predicting diabetes, and they get preferable results. Decision tree is one of popular machine learning methods in medical field, which has grateful classification power. Random forest generates many decision trees. Neural network is a recently popular machine learning method, which has a better performance in many aspects.

So in this study, we used **Naïve Bayes, Random Forest, Logistic Regression** for prediction of diabties.

**Objective**

Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world’s diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes.

There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health. The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used. Machine learning methods are widely used in predicting diabetes, and they get preferable results.

Diabetes is one of the deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patientsneed to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain.

But with the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to predict whether the patient has diabetes or not.

Furthermore, predicting the disease early leads to treating the patients before it becomes critical. Data mining has the ability to extract hidden knowledge from a huge amount of diabetes-related data. Because of that, it has a significant role in diabetes research, now more than ever. The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on following classification methods namely, Logistic Regression, Naïve Bayes and Random Forest.

**Requirements**

**Software:**

* A JSON parsing library such as Jsonpickle
* Machine Learning/Data Science libraries such as numpy and sci-kit learn
* Jupiter Notebook
* Pima Dataset API
* VMware

**Hardware:**

The between CPUs and GPUs favour the latter because of the large amount of cores of GPUs offsetting the 2–3x faster speed of CPU clocks – ~3500 (GPU) vs ~16 (CPU). The GPU cores are a streamlined version of the more complex CPU cores, but having so many of them enables GPUs to have a higher level of parallelism and thus better performance. However for our requirement CPU can also perform the task. CPU such as i7–7500U can train an average of ~115 examples/second will work just fine.

* Quad core Intel Core i7 Skylake or higher (Dual core is not the best for this kind of work, but manageable)
* 16GB of RAM (8GB is okay but not for the performance you may want and or expect)
* M.2 PCIe or regular PCIe SSD with at least 256GB of storage, though 512GB is best for performance. The faster you can load and save your applications, the better the system will perform. (SATA III will get in the way of the system’s performace)
* Premium graphics cards, so things with GTX 980 or 980Ms would be the best for a laptop, and 1080s or 1070s would be the best for the desktop setup.

**Methodology**

Classification is one of the most important decision making techniques in many real world problem. In this work, the main objective is to classify the data as diabetic or non-diabetic and improve the classification accuracy. For many classification problem, the higher number of samples chosen but it doesn’t leads to higher classification accuracy. In many cases, the performance of algorithm is high in the context of speed but the accuracy of data classification is low. The main objective of our model is to achieve high accuracy. Classification accuracy can be increase if we use much of the data set for training and few data sets for testing. This survey has analyzed various classification techniques for classification of diabetic and non-diabetic data.

The algorithms don’t work alone; we have developed an ensemble method which uses votes given by the other algorithms to produce the final result. The system accepts final result, only when more than two models give same prediction outputs. It gives the majorities decision. This voting mechanism eliminates the algorithm dependent misclassifications. It also helps to get a more accurate prediction of the disease.

**Naïve Bayes Algorithm**

This is based on Bayes rule of conditional probability. It uses all the attributes contained in the data and analyzes them individually as though they are equally important and independent of each other. The build process for Naive Bayes is parallelized. It overcomes various limitations like the omission of complex iterative estimations of the parameter because it can be applied to a large dataset in real time. We have used a 70:30 percentage split technique to build the model using this algorithm. 70 percent of the data set has been used to train the data and other 30 percent of the data set have been used to test the model.

**Decision Tree J48 Algorithm**

Decision-Tree is a tree structure which has the form of a flowchart. It can be used as a method for classification and prediction with a representation using nodes and internodes. Root and internal nodes are the test cases. Leaf nodes considered as class variables. In order to classify a new item, it creates a decision tree based on the attribute values of the available training data set. Every node of the tree is generated by calculating the highest information gain for all attributes. If any attribute gives an unambiguous end result, the branch of that attribute will be terminated and then the target value is assigned to it.

**The Random Forest Classifier**

Random forest, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

**Logistic Regression**

In statistics Logistic regression is a regression model where the dependent variable is categorical, namely binary dependent variable-that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed using logistic regression. Many other medical scales used to assess severity of a patient have been developed using logistic regression. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription. In economics it can be used to predict the likelihood of a person's choosing to be in the labour force, and a business application is about to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing.

**Dataset Used**

Data has been obtained from Pima Indians Diabetes Database and the National Institute of Diabetes and Digestive and Kidney Diseases.

**Procedure:**

* Load previous datasets to the system.
* Data pre-processing is performed.
* Following operations are performed on the dataset after that.

        a. Replace Missing values.

        b. Normalization of values.

* User input data to the system in order to diagnose whether he has the disease or not.
* Build three models using J48 Decision Tree, Naïve Bayes, and Logistic Regression Algorithms and train the data set.
* Test the dataset using three models.
* Get the evaluation results.

**Modules/Functionalities**

**Scikit-learn**

This is another popular Python library for machine learning. In fact, scikit-learn is the primary library for machine learning. It has algorithms and modules for pre-processing, cross-validation, and other such purposes. Some of the algorithms deal with regression, decision trees, ensemble modeling, and non-supervised learning algorithms like clustering.

**NumPy**

NumPy is another wonderful Python library for machine learning and heavy computation. NumPy facilitates easy and efficient numeric computation. It has many other libraries built on top of it like Pandas.

**Pandas**

This is one of the Python libraries which is built on top of NumPy. It comes in handy with data structures and exploratory analysis. Another important feature it offers is DataFrame, a 2-dimensional data structure with columns of potentially different types.

**Matplotlib**

It is a low-level library for creating two-dimensional diagrams and graphs. With its help, you can build diverse charts, from histograms and scatterplots to non-Cartesian coordinates graphs. Moreover, many popular plotting libraries are designed to work in conjunction with matplotlib.

**Seaborn**

Like Matplotlib, it's also a good library for plotting but with Seaborn, it is easier than ever to plot common data visualizations. It is built on top of Matplotlib and offers a more pleasant, high-level wrapper. You should learn effective data visualization.

**SciPy**

It is a very popular library among Machine Learning enthusiasts as it contains different modules for optimization, linear algebra, integration and statistics. There is a difference between the SciPy library and the SciPy stack. The SciPy is one of the core packages that make up the SciPy stack. SciPy is also very useful for image manipulation.

**Applications**

Machine Learning is concerned with the development of algorithms and techniques that allows the computers to learn and gain intelligence based on the past experience. It is a branch of Artificial Intelligence (AI) and is closely related to statistics. By learning it means that the system is able to identify and understand the input data, so that it can make decisions and predictions based on it.

* Now days, machine learning algorithms are used for automatic analysis of high dimensional biomedical data.
* Diagnosis of liver disease, skin lesions, cancer classification, risk assessment for cardiovascular disease and analysis of genetic and genomic data are some of the examples of biomedical application of ML.
* For liver disease diagnosis, Hashemi et al. (2012) has successfully implemented SVM algorithm.
* In order to diagnose major depressive disorder (MDD) based on EEG dataset, Mumtaz et al. (2017) have used classification models such as support vector machine (SVM), logistic regression(LR) and Naïve Bayesian (NB)

Diabetes is a very common metabolic disease. Usually onset of type 2 diabetes happens in middle age and sometimes in old age. But nowadays incidences of this disease are reported in children as well. There are several factors for developing diabetes like genetic susceptibility, body weight, food habit and sedentary lifestyle.

Undiagnosed diabetes may result in very high blood sugar level referred as hyperglycemia which can lead to complication like diabetic retinopathy, nephropathy, neuropathy, cardiac stroke and foot ulcer.

So, early detection of diabetes is very important to improve quality of life of patients and enhancement of their life expectancy.

**Limitations**

Astounding technological breakthroughs in the field of Artificial Intelligence (AI) and its sub-field Machine Learning (ML) have been made in the last couple of years. Machines can now be trained to behave like humans enabling them to mimic complex cognitive functions like informed decision-making, deductive reasoning, and inferences. Robots behaving like humans is no longer science fiction, but a reality in multiple industry practices today. As a matter of fact, human society is gradually becoming more reliant on smart machines to solve day to day challenges and make decisions.

A good example of a simple use case for machine learning that has completely permeated our day-to-day lives is spam filters, which intrinsically determine whether a message is junk based on how closely it matches emails with a similar tag.

In all the hype surrounding these game-changing technologies, the reality that often times gets lost amidst both the fears and the headline victories like Cortana, Alexa, Google Duplex, Waymo, and AlphaGo, is that AI technologies have several limitations that will still need a substantial amount of effort to overcome.

Here are some of the limitations:

* Machine Learning Algorithms Require Massive Stores of Training Data.
* Labeling Training Data Is a Tedious Process.
* Machines Cannot Explain Themselves.
* Despite the multiple breakthroughs in deep learning and neural networks, AI models still lack the ability to generalize conditions that vary from the ones they encountered in training.
* As AI and machine learning algorithms are deployed, there will likely be more instances in which potential bias finds its way into algorithms and data sets. In some instances, models that are seemingly performing well maybe actually picking up noise in the data. As much as transparency is important, unbiased decision making builds trust.

**References**

* Komi, Zhai. 2017. Application of Data Mining Methods in Diabetes Prediction
* Alan Siper, Roger Farley and Craig Lombardo, “Machine Learning and Data Mining Methods in Diabetes Research”, Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 6th, 2005.
* Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research.
* Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
* Udemy online tutorials on Machine Learning with python and R.
* Stack Overflow community for debugging.

**Group Members-**

Akshita Agarwal (1601410012) ----------------------

Lavi Thukral (1601410060) ----------------------

Priyanshi Kotnala (1601410081) ----------------------

**Project Guide-** **Project Incharge-**

------------------- -----------------------

**Head of Department-**

LS Maurya



Department of Computer Science and Technology

**Shri Ram Murti Smarak College of Engineering and Technology, Bareilly**