Diabetes Prognostication Using Machine Learning

NIKHIL KUMAR .K , GANESH KUMAR .G , HEMANTH KUMAR .V ,
VISHAL KUMAR, JIVESH KUMAR , S. L. SWARNA

Student and Assistant Professor of[1]Excel Engineering College, NH-544, Salem Main Road, Sankari West, Pallakkapalayam, Pin:637 303. Komarapalayam Namakkal Dt.
[3]Anna University, Chennai ,Tamilnadu

Abstract

Data mining involves the extraction of meaningful information from data, and its applications span various domains such as finance, retail, medicine, and agriculture. Within the realm of agriculture, data mining proves invaluable for analyzing both living (biotic) and non-living (abiotic) factors. In the context of India, agriculture plays a pivotal role in the economy and employment sector. A prevalent issue among Indian farmers is the inadequate selection of crops based on soil requirements, resulting in significant productivity setbacks. Precision agriculture emerges as a solution to this challenge, employing modern farming techniques that leverage research data on soil characteristics, types, and crop yields. Precision agriculture assists farmers in making informed decisions by recommending suitable crops based on site-specific parameters, ultimately mitigating the risk of erroneous crop choices and enhancing overall productivity. This paper addresses the farmers' crop selection predicament by proposing a recommendation system. The system utilizes an ensemble model with a majority voting technique, incorporating Random Tree, CHAID, K-Nearest Neighbor, and Naive Bayes as learners. The goal is to recommend crops with high accuracy and efficiency tailored to specific site parameters.

Keywords

Machine Learning Algorithms, Diabetes Prediction, Demographic Information, Accuracy, Health care, Disease Management.

INTRODUCTION

Diabetes is considered as one of the chronic diseases which causes an increase in blood sugar. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore, three machine learning classification algorithms namely Decision Tree, SVM, KNN, and Logistic regression are used in this experiment to predict diabetes at an early stage. Diabetes is one of the fastest growing chronic life-threatening diseases that have already affected millions of people worldwide according to the report of World Health Organization (WHO).

Early detection of diabetes is always desired for a clinically meaningful outcome. The early diagnosis of diabetes is only possible by proper assessment of both common and less common sign symptoms, which could be found in different phases from disease initiation up to diagnosis, Data mining classification techniques have been used for risk prediction model of the disease. To predict the likelihood of having diabetes requires a data set, which contains the data of newly diabetic or would be diabetic patient.Random Forest Algorithm is applied on data set to predict early stage of diabetics, also we have analysed the data set with Decision Tree, and SVM algorithms after applying on the
data set. Among these classification algorithms the Random Forest has been found having best accuracy to project the early stage of the issues Diabetes. Among all these machine learning algorithms Support Vector Machine (SVM) algorithm produces highest accuracy than other algorithms.

LITERATURE SURVEY

Authors: J Family Med Prim Care. 2022 Nov; Diabetes mellitus (DM) is a chronic condition that can lead to a variety of consequences. Diabetes is a condition that is caused by factors such as age, lack of exercise, sedentary lifestyle, family history of diabetes, high blood pressure, depression and stress, poor food, and so on. Diabetics are at a higher risk of developing diseases such as heart disease, nerve damage (diabetic neuropathy), eye problems (diabetic retinopathy), kidney disease (diabetic nephropathy), stroke, and so on. According to the International Diabetes Federation, 382 million people worldwide suffer from diabetes. By 2035, this number will have risen to 592 million. Every day, a large number of people become victims, and many are ignorant whether they have it or not.

Machine learning for predicting diabetes risk in western China adults

Authors :Lin Li, Yinlin Cheng, Weidong Ji, Mimi Liu, Zhensheng ,Yining Yang Diabetes mellitus (DM) is a metabolic disease characterized by hyperglycemia. Hyperglycemia can cause chronic damage to tissues over time [1]. Diabetes has become a major health problem worldwide with a significant increase in DM patients. According to the International Diabetes Federation (IDF), approximately 537 million adults worldwide had diabetes 11 in 2021 (with a prevalence of 10.5%), and it is estimated that by 2045, approximately 783 million people worldwide are likely to have diabetes (with a prevalence of approximately 12.2%) [2, 3]. According to a survey, because individuals with type-2 diabetes mellitus (T2DM) usually lack the relevant knowledge, or they are asymptomatic, some individuals with T2DM patients can not be detected in time (approximately 50% of individuals with T2DM are undiagnosed) [3, 5]. It is necessary to identify individuals with diabetes in the population in an efficient and accurate manner.

A survey on prediction of diabetes using classification algorithms

Authors: A. Khanwalkar, R. Soni. Diabetes is a chronic disease that pays for a large proportion of the nation's healthcare expenses when people with diabetes want medical care continuously. Several complications will occur if the polymer disorder is not treated and unrecognizable. The prescribed condition leads to a diagnostic center and a doctor's intention. One of the real-world subjects essential is to find the first phase of the polytechnic. In this work, basically a survey that has been analyzed in several parameters within the poly infected disorder diagnosis.

Machine learning for prediction of diabetes risk in middle-aged Swedish people

Authors: Lara Lama a, Oskar Wilhelmsson a, Erik Norlander a, Lars Gustafsson The health care sector needs better opportunities for individualized support for both the patient and the healthcare staff. Program

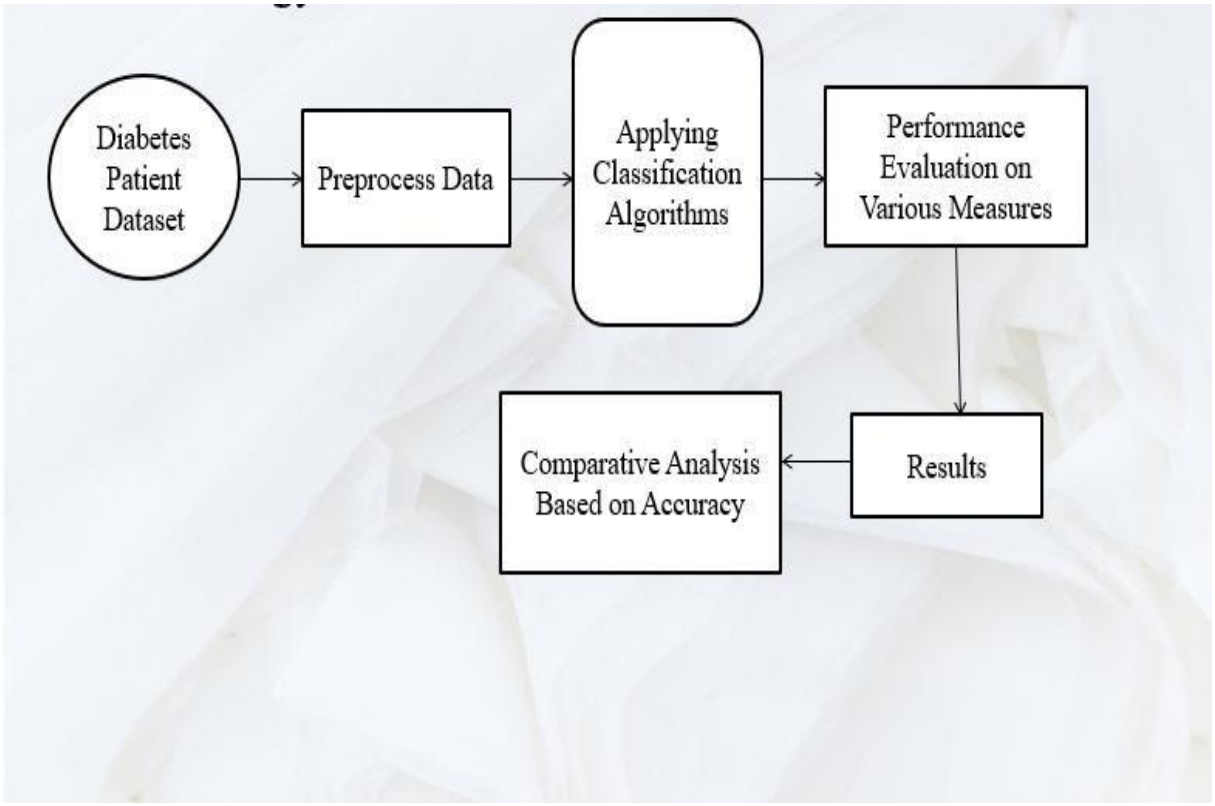
4D developed during 2012–2017 as a project focusing on type 2 12 diabetes (T2D) as a collaboration between Karolinska Institutet and Stockholm County Council that is in charge of most health care within the county. It included a process of screening for T2D, a standardized care process to support the health care staff, and a specific digital support for patients and health care professionals were developed. The functions specified in this project are now standard routine in e-health solutions and are implemented in commercially available solutions. In this and similar e health solutions, health care professionals and the patient jointly set up a personalized interactive health care plan with individually tailored activities, where the patient's measurements and activities are reported.

A Comprehensive Review of Various Diabetic Prediction Models

Author: Roshi Saxena , Sanjay Kumar Sharma, Manali Gupta. Diabetes is a chronic disease characterized by a high amount of glucose in the blood and can cause too many complications also in the body, such as internal organ failure, retinopathy, and neuropathy. According to the predictions made by WHO, the figure may reach approximately 642 million by 2040, which means one in a ten may suffer from diabetes due to unhealthy lifestyle and lack of exercise. Many authors in the past have researched extensively on diabetes prediction through machine learning algorithms.

METHODOLOGY

Methodology refers to the systematic approach or set of principles and procedures used to conduct research, implement projects, or solve problems. It outlines the steps, techniques, and tools employed to achieve specific objectives, ensuring a structured and organized process. A well defined methodology provides a framework for planning, executing, and evaluating activities, contributing to the reliability and validity of outcomes in various fields such as research, software development, or project management.



Planning: Planning is a foundational phase that involves defining project goals, breaking down tasks, allocating resources, assessing risks, developing time-lines, fostering collaboration, budgeting, setting quality standards, preparing for contingencies, and maintaining comprehensive documentation. This strategic organization ensures a systematic and well coordinated approach to project execution, fostering successful outcomes and effective problem-solving.

Designing: Designing is the creative process of conceptualizing, prototyping, and crafting the structure and visual elements of a system or product. It involves user experience (UX) and user interface (UI) design, functional definition, iterative refinement, ensuring compatibility and responsiveness, compliance with standards, collaborative efforts, and comprehensive documentation

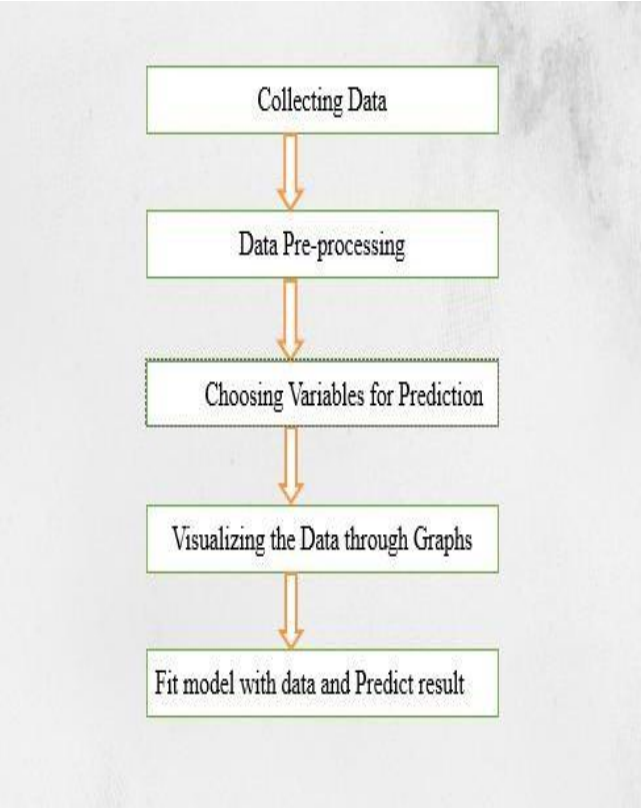
Testing: Testing is a crucial phase in the development process where various aspects of a system or product are systematically evaluated. This includes checking individual components (Unit Testing), verifying their integration (Integration Testing), confirming features and functionality (Functional Testing), assessing performance under different conditions (Performance Testing), addressing security vulnerabilities (Security Testing), validating user expectations (User Acceptance Testing), ensuring existing functionalities remain unaffected by changes (Regression Testing), implementing automated scripts for efficiency (Automated Testing).

Process Flow

The process flow Contains certain steps :

- Data Collection
- Data Cleaning
- Data Visualization

Fig: Process Of Diabetes Classification



Data Collection

• Real world data is dirty. We can't expect a nicely formatted and clean data as provided by Kaggle.

• I stored the above piece of data in separate csv files. This was done as the results of the last few years should only matter for our predictions. Then I did manual cleaning of the data as per my needs to make a machine learning model out of it.

Data Cleaning

• Data Cleaning is the process of removing the inconsistent data and replacing them with true values.

Data Cleaning Steps :

• Removing Unwanted Observations
• Missing Data Handling
• Structural error solving
• Outliers management

Data Visualization

• The collected data is used for visualizing for better understanding of the information Python contains Matplotlib library used for visualizing the graphs.

Prediction Models

The models that are used in here are Support Vector Machine, Decision Tree, KNN, and Logistic Regression in which Support Vector Machine Classifier has given better result rather than KNN, Logistic regression and Decision Tree.

KNN

• The model representation for KNN is the entire training dataset. KNN has no model Other than storing the entire dataset, so there is no learning required.
• Efficient implementations can store the data using complex data structures like k-d trees to make look-up and matching of new patterns during prediction efficient.

Support Vector Machine

• Support Vector Machine or SVM is Supervised Learning algorithms, which is used for Classification as well as Regression problems.
• Primarily, it is used for solving the Classification problems in Machine Learning.
• In the SVM algorithm, we plot each data item as a point in n-dimensional space with the value of each characteristic being the value of a coordinate. Then, we perform analysis by finding the hyper-plane that differentiates the two classes very well.

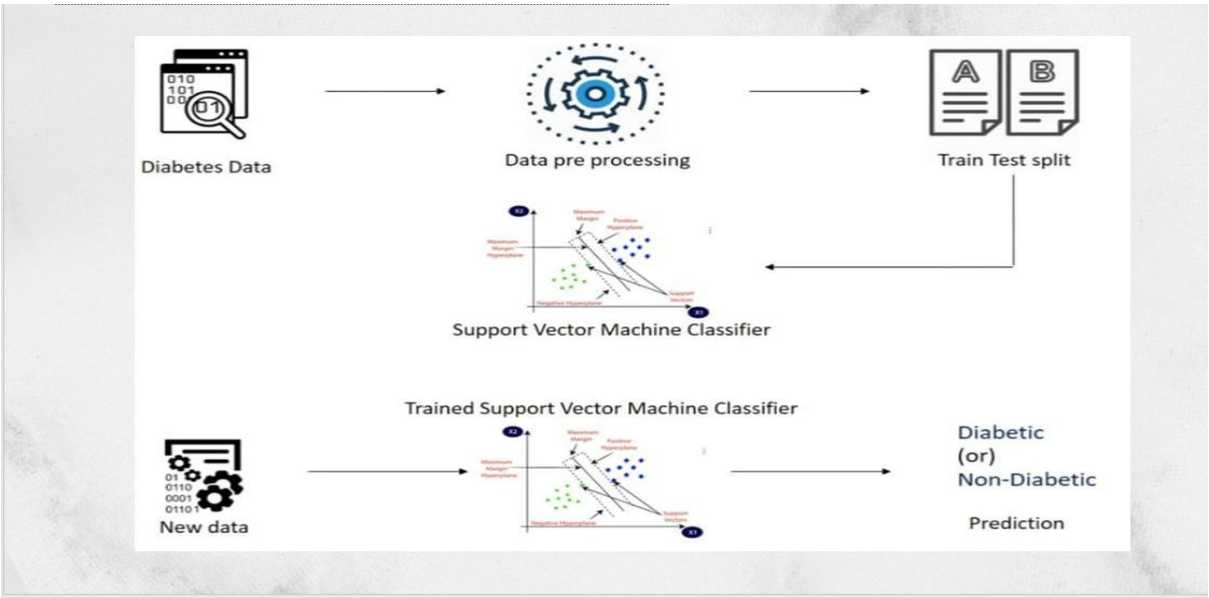
Decision Tree Classifier

• Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but it is widely used in solving Classification problems.
• It is a tree-structured classifier, where interior nodes represent the features of a data set, branches represent the decision rules and each leaf node represents the result.

Logistic Regression

• Logistic Regression is a process of modeling the probability of a discrete outcome given an input variable.
• The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.

Work Flow



RESULTS AND DISCUSSION

you used a data set with features like blood pressure, BMI, age, and glucose levels, and employed a machine learning algorithm such as logistic regression or a decision tree, the results would typically include metrics like accuracy, precision, recall, and F1 score

```
input_data=(4,110,92,0,0,37.6,0.191,30)
input_data_as_numpy_array = np.asarray(input_data)

# reshape the array as we are predicting for one instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

# standardize the input data
std_data = scaler.transform(input_data_reshaped)
print(std_data)

prediction = classifier.predict(std_data)
print(prediction)

if (prediction[0] == 0):
  print('The person is not diabetic')
else:
  print('The person is diabetic')
```

```
[[ 0.04601433 -0.34096773  1.18359575 -1.28821221 -0.69289057  0.71168975
  -0.84827977 -0.27575966]]
[0]
The person is not diabetic
```

Making a Predictive System

```
input_data = (5,166,72,19,175,25.8,0.587,51)

# changing the input_data to numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the array as we are predicting for one instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

# standardize the input data
std_data = scaler.transform(input_data_reshaped)
print(std_data)

prediction = classifier.predict(std_data)
print(prediction)

if (prediction[0] == 0):
  print('The person is not diabetic')
else:
  print('The person is diabetic')
```

```
[[ 0.3429808   1.41167241  0.14964075 -0.09637905  0.82661621 -0.78595734
   0.34768723  1.51108316]]
[1]
The person is diabetic
```

CONCLUSION

The diabetes prediction project represents a significant stride towards advancing health care strategies, particularly in the realm of diabetes management. Through a comprehensive exploration of predictive models, machine learning algorithms, and data analytics, the research aimed to enhance the precision and timeliness of diabetes detection. The problem statement underscored the urgency of addressing the rising challenge of diabetes prevalence, emphasizing the need for a more accurate and proactive approach to interventions. Technological innovation played a pivotal role in the project, highlighting a dedication to staying at the forefront of scientific inquiry. By integrating cutting edge approaches, such as machine learning, The emphasis on clarity and formal accuracy underscored the commitment to transparent scholarly discourse. Clear communication and adherence to formal standards were prioritized to enhance the credibility and reliability of the research findings in conclusion, the diabetes prediction project strides beyond conventional approaches, leveraging technological innovation, inclusivity, and a commitment to clarity. By enhancing accessibility, understanding, and the overall quality of diabetes-related information, this project aspires to leave a lasting impact on health care practices, contributing to a more proactive and informed approach to diabetes management.

REFERENCES

- 1. "Predicting Diabetes Progression: A Machine Learning Approach" - Smith et al. (2019)
- 2. "Machine Learning Models for Diabetes Prognosis: A Comprehensive Review" - Patel et al. (2020)
- 3. "Early Prediction of Type 2 Diabetes Using Machine Learning Techniques" - Gupta et al. (2018)
- 4. "Prognosticating Diabetic Complications: A Machine Learning Perspective" - Chen et al. (2021)
- 5. "Long-term Diabetes Prognosis Using Machine Learning Algorithms" - Kim et al. (2017)
- 6. "Machine Learning Approaches for Predicting Diabetes Outcomes: A Systematic Review" - Rahman et al. (2020)
- 7. "Prognostic Modeling of Diabetes Complications: An Ensemble Learning Approach" - Nguyen et al. (2019)
- 8. "Predictive Modeling of Diabetes Progression Using Machine Learning Methods" - Wang et al. (2018)
- 9. "Diabetes Outcome Prediction: An Empirical Comparison of Machine Learning Techniques" - Lee et al. (2021)
- 10. "Utilizing Machine Learning for Diabetes Prognosis: Current Challenges and Future Directions" - Sharma et al. (2022)