EARLY DETECTION OF OUTBREAKING DIABETES USING MACHINE LEARNING ALGORITHM

Mrs. A. BENETAMARY, M.E., Assistant Professor,

Department of Computer Science and Engineering

Ms. S.SANTHOSHINI, B.E, Student of Computer Science Engineering

Ms. T.DIVYA, B.E, Student of Computer science and Engineering

St. Joseph College of Engineering, Sriperumbudur, Chennai.

ABSTRACT

The aim of our project is to improve the accuracy and timeliness of diabetes outbreak forecasts with a view to proactively addressing health problems, as well as allocating resources. Retrospective analysis and comparison of existing epidemiological data will be used to verify the results obtained from this model. In addition, this research examines the interpretation ability of a machine learning model and provides an insight on its major contributors and their influence on forecast results.

The methodology covers the collection and preparation of massive data sets, containing information about demographics, environmental factors as well as historical health records. In order to analysis the complicated relationship between data and find patterns that indicate possible outbreaks of diabetes, machine learning techniques are used in addition to decision trees, support vector machines as well as neuron networks. Effective strategies for preventing and mitigating possible outbreaks are needed, due to the growing prevalence of diabetes worldwide.

Our project looks at applying machine learning techniques to predict diabetes outbreaks through a wide variety of data sources, combined with more sophisticated analytical tools. To set up a comprehensive forecasting framework, the model is integrated with demographic, environment and health indicators. The models incorporate four distinct ML algorithms, including Support Vector Machines (SVM), Linear Regression, Decision Trees, and K-Nearest Neighbors (KNN).

Through the strategic application of these algorithms, we aim to extract actionable insights from the data, shedding light on the critical factors influencing diabetes prediction. Diseases Outbreak Prediction is an iterative and dynamic process.

The cases for diabetes are ever growing, with more constraints playing huge factors in causing diabetes. To maintain model relevance and accuracy, regular updates are essential to keep our models aligned with health studies and findings regarding diabetes.

KEY TERMS:

BMI - Body Mass Index, KNN – K Nearest Neighbor, LightGBM – Light Gradient Boosting Machine, ML – Machine Learning, SKLearn – Sci-Kit Learn, SMOTE – Synthetic Minority Oversampling Technique.

INTRODUCTION

Diabetes mellitus is a chronic metabolic disease that affects a lot of people worldwide. It is marked by elevated levels of blood sugar, potentially resulting in severe complications In like heart disease, stroke, and blindness. In order to prevent these complications and im prove overall health outcomes, early diagnosis and effective management of diabetes is essential.

There are a number of limitations in today's approaches to the diagnosis and management of diabetic disease. For example, traditional methods for detecting diabetes such as glucose tolerance test carried out orally are time consuming and invasive. Additionally, current treatment strategies do not adequately address the individual needs of patients with diabetes, and many patients still experience complications from the disease.

Our project involves identifying diabetes outbreak patterns in cities that are at risk of having a growing population of diabetes patients. As mentioned earlier, this disease affects millions of people worldwide. Our aim is to leverage machine learning algorithms to proactively identify potential outbreaks, allowing health organization and facilities to take notice of it and act accordingly.

Our project’s transformative potential for public health lies in the prediction of outbreaks of disease. We'll gain the ability to anticipate and plan for possible epidemics, which would allow us to react in time and with precision. Our project's significance lies in its contribution to a paradigm shift, where predictive analytics becomes a cornerstone in the arsenal against infectious diseases, guiding strategic planning and mitigating the devastating consequences of global health crises. This proactive approach not only saves lives but also optimizes resource allocation, minimizing the economic and societal impact of outbreaks. Disease outbreak prediction empowers healthcare systems, governments, and communities to respond swiftly and effectively, fostering a resilient and proactive stance against emerging health threats.

LITERATURE SURVEY

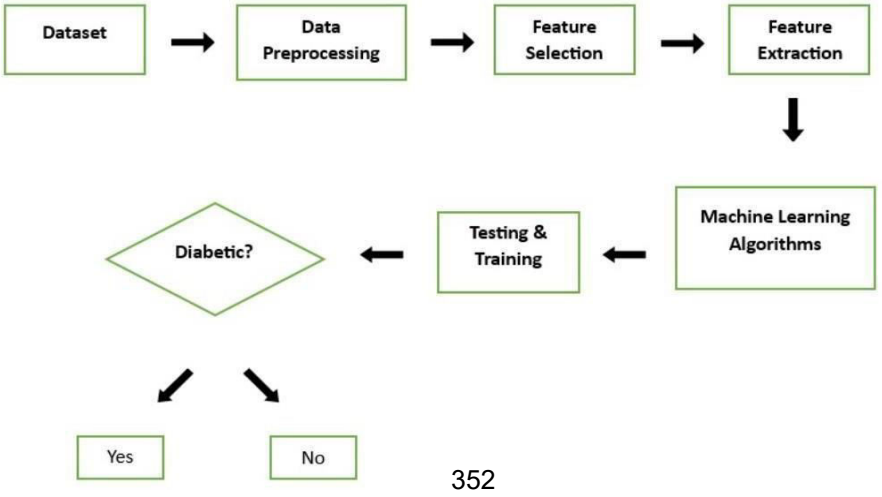
Ayan Mir et al. focused on diabetes prediction in their paper. Diabetes datasets from Pima Indians are used. On the Weka interface, classification algorithms such as Naive Bayes, SVM, Random Forest, and Simple CART are used. As a result, SVM outperforms the others in terms of accuracy.

Veenavijayan.V et al. Choosing appropriate classification algorithms clearly improves the system's accuracy and efficiency. The primary goal of this study is to compare the benefits of various pre-processing techniques for diabetes prediction decision support systems based on Support Vector Machine (SVM), Naive Bayes classifier, and Decision Tree. Principal Component Analysis and Discretization are the pre-processing methods used in this study.

The variation in accuracy was evaluated with and without pre- processing techniques. In this study, the Weka tool is used. The dataset was obtained from the machine learning repository at the University of California, Irvine (UCI).

Another study used the K-nearest neighbour (KNN) algorithm to diagnose diabetes, and the authors discovered that as the value of k increased, the accuracy and error rates improved. Other data mining techniques compared in the study included kernel density, Automatically Defined Groups, bagging algorithms, and support vector machines. KNN was found to be one of the most powerful and widely used algorithms, producing more precise and effective results.

SYSTEM DESIGN



The article presents a comprehensive outline of a system designed for predicting diabetes outbreaks. This system encompasses multiple stages, beginning with data retrieval and culminating in model training, with the objective of offering businesses valuable insights into and foresight regarding diabetic prediction. Let's delve into the architectural elements of this system.

Central to this system is the procedure of extracting data from a database. The author obtains data from Kaggle, emphasizing the importance of effective data management for analytical purposes. The ability to establish a connection to the database and query it forms the fundamental infrastructure for subsequent analysis and model development.

Once the data is successfully extracted, the article highlights the importance of data preprocessing. This crucial step involves the selection of pertinent columns for analysis while discarding those of lesser relevance. Additionally, the article uncovers a noteworthy insight concerning zero values in certain columns, which are attributed to newly acquired patients details. Understanding and managing these intricacies are critical for ensuring the accuracy of the predictive model.

The next step is Feature Selection. The extracted data is now optimized to work on certain features, which will help in building the model. After this stage, various machine learning algorithms are then deployed to build models. In this case, we use Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Decision tree, and Gradient Boosting. The data is split into training and testing data, for the algorithms to operate on. Through various evaluations from the data, we can conclude whether the patient is diabetic or not.

IMPLEMENTATION
SNAPSHOTS

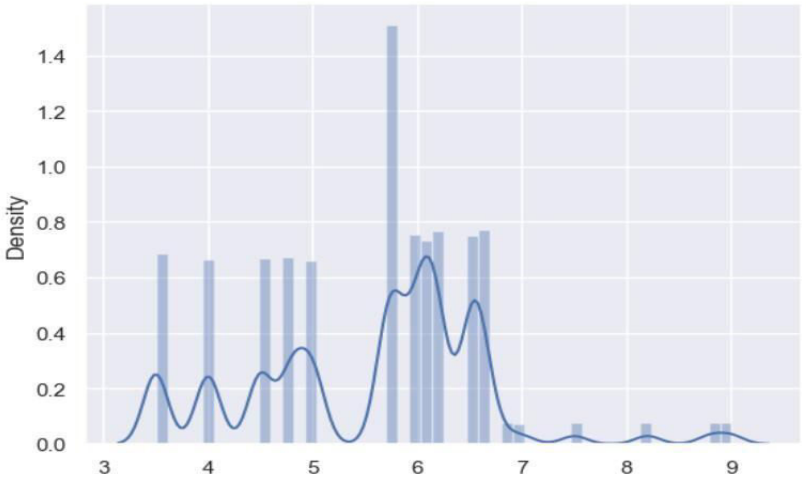
```
In [8]: df.describe() # Statistical values of all columns
```

Out[8]:

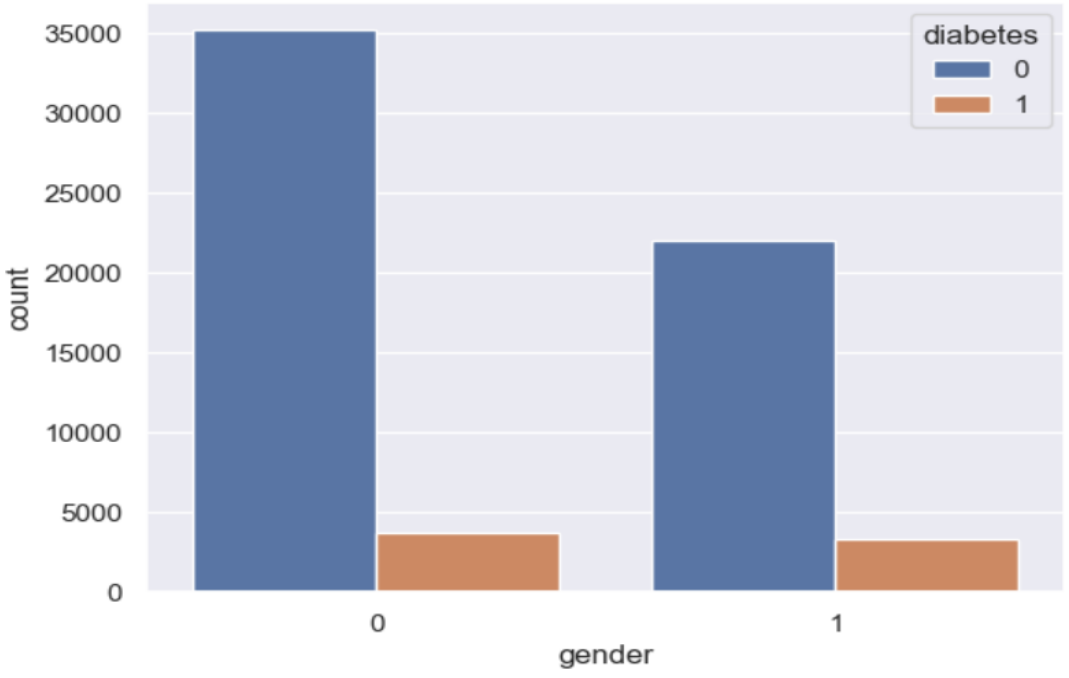
	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060	0.085000
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136	0.278883
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000	0.000000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000	0.000000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000	0.000000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000	0.000000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000	1.000000

Statistical Data

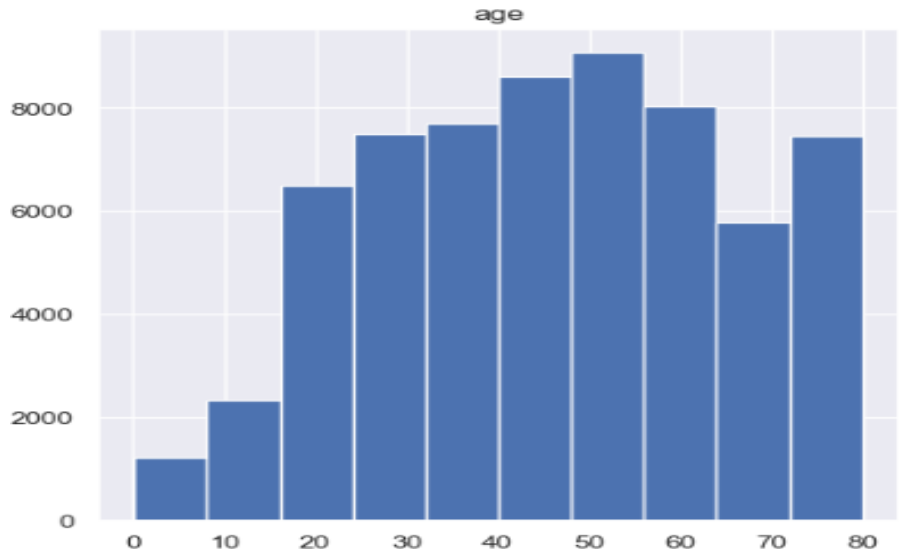
```
In [22]: sns.distplot(x=df['HbA1c_level']) # density plot for hemoglobin level column
Out[22]: <Axes: ylabel='Density'>
```



Frequency of hemoglobin levels



Bar Chart displaying genders with having diabetes



Distribution of patient age

LOGISTIC REGRESSION

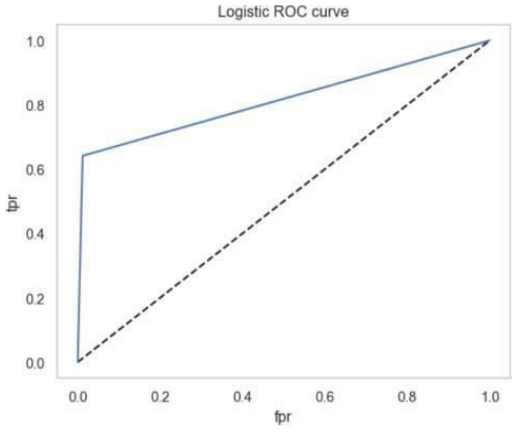
Classification Report

	precision	recall	f1-score	support
0	0.96	0.99	0.97	17080
1	0.87	0.64	0.74	2172
accuracy			0.95	19252
macro avg	0.91	0.81	0.86	19252
weighted avg	0.95	0.95	0.95	19252

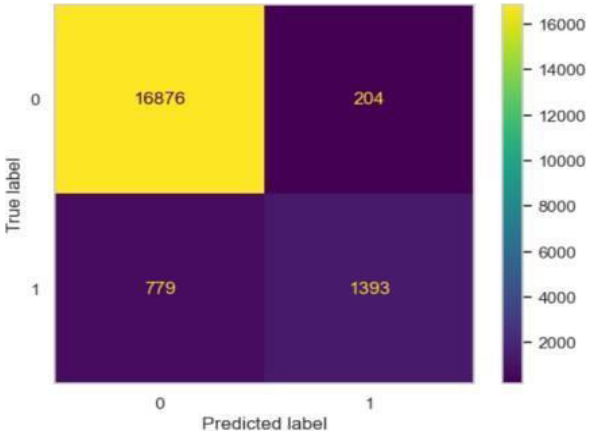
Accuracy

```
In [74]: # Make predictions on the test data
         y_pred = logReg.predict(X_test)

In [75]: # Calculate the accuracy of the model
         accuracy = metrics.accuracy_score(y_test, y_pred)
         print(f"Accuracy: {accuracy*100}")

Accuracy: 94.89403698317058
```

ROC Curve



Confusion Matrix



CONCLUSION

This dataset comprises randomly collected information from an Electronic Health Record database.
It includes a total of 100000 rows, each representing an individual patient,
with details recorded across 9 columns. The dataset encompasses various attributes,
such as gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, glucose level, diabetes.

FUTURE ENHANCEMENTS

Early Detection and Prevention Strategies: Focus on developing models that not only predict the onset of diabetes but also provide actionable insights for early intervention and preventive measures. This could involve integrating with healthcare systems to deliver personalized recommendations.

User friendly Interfaces: Develop user-friendly interfaces for healthcare professionals and patients to interact with the prediction models. This can facilitate easy interpretation of results and support informed decision-making.

Collaboration with Healthcare providers: Encourage teamwork among data experts, researchers, and healthcare providers to make sure the predictive models match up with medical guidelines and are really useful in real healthcare situations. This teamwork helps turn research discoveries into practical tools for healthcare

REFERENCES:

1. Report of theexpert committee on the diagnosis and classification of diabetes mellitus.Diabetes Care.1997;1183-97.

2. Norris SL,Lua J,Amith SJ,Schmid CH,Engelagu MM.Self-management education for adults with type2 diabetes:A meta-analysis of the effect on glycemic control.Diabetes Care.2002;25:1159-71.

3. Shaw JE,Sicree RA Zimmet PZ.Global estimates of the prevalence of diabetes for 2010 and 2030.Diabetes Res Clin Pract. 2010;87:4-14.

4. Anjana Rm,Pradeep R,Deepa M,Datta M,Sudha V,Unnikrishnan R,etal.Prevelence of diabetes and prediabetes(impaired fasting glucose and/or impaired glucose tolerance) in urban and rural india:Phase I results of the Indian council of Medical Research India Diabetes(ICMRINDIAB)study.Diabetologia.2011;54:3022-7.

5. Ramachandran A,Snehalatha C,Salini J,Vijay V.Use of glimepiride and insulin sensitizers in the treatment of type 2 diabetes-a study in Indians.J Assoc Physicians India.2004;52:459-63.

6. Wagai GA,Romshoo GJ.Adiposity contributes to poor glycemic control in people with diabetes mellitus,a randomized case study;in South Kashmir,India.J Family Med Prim Care.2020:4623-6.

7. AACE/ACE Position Statement on the prevention.Diagnosis and treatment of obesity(1998Revision)Endoc Practice. 1998;4:297-330.

8. Birjais R,Mourya AK,Chauhan R,kaur H.Prediction and diagnodis of future diabetes risk:A machine learning approach. SN Appl Sci.2019;1:1-8.

9. Sadhu A,Jadil A.Early-stage diabetes risk prediction:A comparative analysis of classification algorithms.Int Adv Res J Sci Eng Techno;l(IARJSET)2021;8:193-201.

10. Xue J,Min F,Ma F.Research on diabetes prediction method based on machine learning.J Phys Confser.2020;1684:1-6.

AUTHOR 1

Mrs. A. BENETAMARY, M.E., Department of Computer Science and Engineering at St. Joseph College of Engineering, Sriperumbudur, Chennai, TamilNadu.

AUTHOR 2

Ms. S,SANTHOSHINI, B.E., Student of Computer Science and Engineering at St. Joseph College of Engineering, Sriperumbudur, Chennai, TamilNadu. I had attended many Workshops and Seminars in the area of Python and Machine Learning.

AUTHOR 3

Ms. T.DIVYA, B.E., Student of Computer Science and Engineering at St. Joseph College of Engineering, Sriperumbudur, Chennai, TamilNadu. I had attended many Workshops, Seminars in Python, Machine Learning.