ABSTRACT

Research on diabetes prediction is a fast-growing area. Better care will come from early diabetes prediction. Numerous health problems are brought on by diabetes. Preventing, keeping an eye on, and spreading awareness about it are therefore essential. Heart disease, kidney issues, and vision problems can all be brought on by both Type 1 and Type 2 diabetes. In this study, we offer a data-mining-based diabetes prediction model. We use Random Forest and Bagging. Furthermore, the performance of the proposed mechanism is analyzed using the confusion matrix, and accuracy performance metrics.

INTRODUCTION

An increased blood glucose level is the hallmark of diabetes, a chronic illness. Over time, diabetes damages the kidneys, eyes, and heart. The task of detecting diabetes early is difficult.

TYPES OF DIABETES

TYPE 1 DIABETES -
In the past, the words "juvenile diabetes" and "insulin-induced diabetes mellitus" (IDDM) were used. It's unclear what caused it. Young people and those under 20 are affected by diabetes. Type 1 will damage and eventually destroy pancreatic cells. Patients with type 1 diabetes have been insulin-dependent their entire lives due to a lack of insulin production. Individuals diagnosed with type 1 diabetes ought to maintain a balanced diet and regular physical activity.

TYPE 2 DIABETES -
Type 2 diabetes is mostly caused by insulin resistance, which happens when cells do not react to insulin as they should. As the illness worsens, insulin deficiency could occur. There has been prior usage of the terms "adult-induced diabetes" and "non-insulin-based diabetes mellitus." The main culprits are inactivity and obesity. Usually, it starts when a child is four.

- A diabetes prediction machine learning model is useful for several reasons, primarily centered around early detection, preventive care, and more efficient healthcare management. Some of the major uses of such inventions include Early detection and Intervention, Personalized Risk Assessment, Preventive Healthcare Strategies, Resources Management, Reduced Healthcare Costs, Patient Empowerment, Population Health Management, Quality of Care Improvement, Research and Innovation.

RELATED WORK

RESEARCH PAPER 1

LINK - Diabetes prediction model using data mining techniques - ScienceDirect

→ This paper proposes a diabetes prediction model using four data mining techniques- Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naive Bayes. This helps in predicting the presence of diabetes.The predicted outcome i.e., Positive or negative has been calculated on the basis of the various parameters such as Glucose, Pregnancies, Blood Pressure, Skin Thickness, Diabetes Pedigree Function, BMI, Insulin, Age, etc.

	METHODOLOGY	CHALLENGES
RANDOM FOREST	Using the input X as the unit for each tree, RF is a tree-based collection classifier in which the randomly assigned vectors are distributed in a completely different way. RF method is an easy-to-use, scalable algorithm for combining tree predictors. It can process many kinds of binary, nominal, and numeric data.	RF models can be more complex, consisting of more trees which further makes it more prone to overfitting. RFs provide a categorical output rather than probabilistic outputs.
SVM	SVMs can be effectively used for diabetes prediction by leveraging their ability to classify data points into different classes based on a hyperplane that maximally separates them in the feature space.	SVM models can become complex, especially when dealing with high-dimensional datasets or using non-linear kernels.SVMs may not scale well to very large datasets, both in terms of the number of instances and features.
LOGISTIC REGRESSION	Logistic regression is a straightforward and interpretable model, making it suitable for diabetes prediction. It provides probabilities and predictions within the [0, 1] range, making it easy to interpret and apply in a clinical context.	Logistic regression is inherently linear, and it may struggle to capture complex, nonlinear relationships in the data. If the dataset is imbalanced, logistic regression may be biased towards the majority class.
NAIVE BAYES	It usually consists of direct acyclic graphics with one parent and only numerous children (relative to the observed nodes), with a crucial premise of autonomy between children in the context of their parents.	Naive Bayes can be sensitive to outliers. Naive Bayes traditionally assumes that features are discrete.The simplicity of Naive Bayes makes it sensitive to feature correlations.

RESULTS

After implementing the proposed mechanism, analyzing the model using various metrics like Sensitivity, Accuracy, Confusion Matrix is done. It is visible that Logistic Regression is high which helps us in getting more accurate results. The rate of computation of positive outcomes is high in logistic regression as compared to other techniques.

RESEARCH PAPER 2

LINK - Combining knowledge extension with convolution neural network for diabetes prediction - ScienceDirect

→ This Paper proposes a diabetes prediction model which involves combining knowledge extension and convolution neural networks (CNN). This model basically uses knowledge extension, Word2vec, entity recognition technique called BERT-BiLSTM-CRF, softmax function, CNN. This prediction model tackles the problem of large-scale annotated diabetes data and the lack of diabetes knowledge. The experimental results show that the KE-CNN model effectively improves the accuracy of the diabetes prediction compared to the benchmark model.

METHODOLOGY:

This diabetes prediction model deals with the numerical vector form of data. The process we follow in this prediction model are as follows:

- We extract the abnormal data feature of diabetes patients from the physical examination record. We fix the extracted abnormal feature using **Word2vec** which is used to create a distributed representation of words into numeric vectors and convert the text into vectors that capture semantics and relationships among words.
- We use an entity recognition technique named **BERT-BiLSTM-CRF** which is used to predict entity labels of unlabeled data. This technique uses a knowledge graph to expand the knowledge of the medical entities. It is then followed by pre-trained Chinese vectors to fix the abnormalities in the data. This step helps our model to be more precise about the latest semantic features.
- Now, we construct a semantic enhanced convolutional neural network. We have a dual channel input for both word and text embedding vectors. This dual channel input intensifies the feature expression of the KE-CNN model. SoftmaxFunction is used in the KE-CNN. This function converts a vector of

numbers into a vector of probabilities which results in obtaining relevant diabetes categories.

ADVANTAGES:

- Using Knowledge extension is beneficial as we only extract relevant information from data sources.
- This prediction model significantly reduces the amount of data required as the model uses softmax function.
- This model comprises dual channel input which handles word and text vectors which results in enhanced feature expression.

CHALLENGES:

- This model uses CNN in which training takes a long time and it is tend to be much slower
- Model may require high computational requirements.
- Model may have to deal with large scale annotated data which is resulting in usage of embedded and entity recognition techniques.

RESULTS:

The diabetes prediction model KE-CNN is to integrate knowledge extension and convolutional neural networks which is enhancing our diabetes data feature as we aim to make the data to be in semantic environment. Using embedded and entity recognition techniques, we handle the abnormalities and make features relevant and semantic possible. This model results in capturing fine grained semantic information which resulted in improving the diabetes diagnostic accuracy.

RESEARCH PAPER 3

LINK - A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures - ScienceDirect

→ This paper presents an effective framework for diabetes disease prediction using a customized hybrid model of artificial neural network (ANN) and genetic algorithms. The large amount of big data in medical research necessitates the use of frontier technologies such as machine learning, deep learning and cloud computing to fully utilize the big data and automate the computation processes in medical research, Here we use techniques like regularization, normalization to manage the skewness of data and achieve a prediction model with highest accuracy.

METHODOLOGY:

This hybrid prediction model uses an improvised technique of detecting the more visible patterns of relations between the variables. We deal with a diabetes patient's dataset which involves a heavy amount of skewness. This degree of skewness is the main factor which may affect our accuracy. So, to handle the degree of skewness in our data, we preprocess the data using **novel normalization.**This normalization makes adjustment of the values of features in our dataset to a common scale thereby resulting in removing the skewness of our dataset. Then,our proposed decision-making algorithm will correctly identify the degree of importance of each variable in influencing the output. This is then followed by the implementation of a **regularization** method that is custom-made for the prediction of diabetes. This customized regularization is used to prevent the overfitting and improve the model's generalization performance. This also enables our model to adjust how closely our model is trained to fit historical data which is an important factor for diabetes prediction.

ADVANTAGES:

- Skewness of data is handled using normalization which is resulting in usage of the features to the fullest.
- Historical data is also used which makes the predication more easy as it plays a vital role in prediction of disease.

CHALLENGES:

- As we are using regularization, it may lead to dimensionality reduction.
- Normalization may increase the complexity of the dataset, decrease the query performance and speed.
- Artificial neural networks have greater computational burden and may be prone to overfitting.

RESULTS:

This hybrid ML framework model using ANN and genetic algorithms with the implementation of regularization and normalization which truly address the context of diabetes, in which all the steps of our algorithms are dedicatedly customized for the context of diabetes. Positive values are favored over negative values as per the human understanding and knowledge of diabetes.

PROPOSED MODEL

PROBLEM STATEMENT

Doctors rely on standard information to treat patients. Examines are summed up after a certain number of cases have been considered when standard information is insufficient. However, this interaction takes time; however, if AI is used, the examples can be recognized earlier. A massive amount of information is required to use AI. Depending on the infection, there is a very limited amount of information available. Furthermore, the number of tests with no illnesses is greater than the number of tests with the infection. Several researchers have proposed various techniques for predicting diabetes using various ML classification techniques, but each has advantages and disadvantages. Some techniques are less accurate than others, but the elapsed time increases as well. So, to overcome these kinds of issues, it is intended to propose a mechanism to predict diabetes outcomes with a high accuracy rate.

OUTCOMES

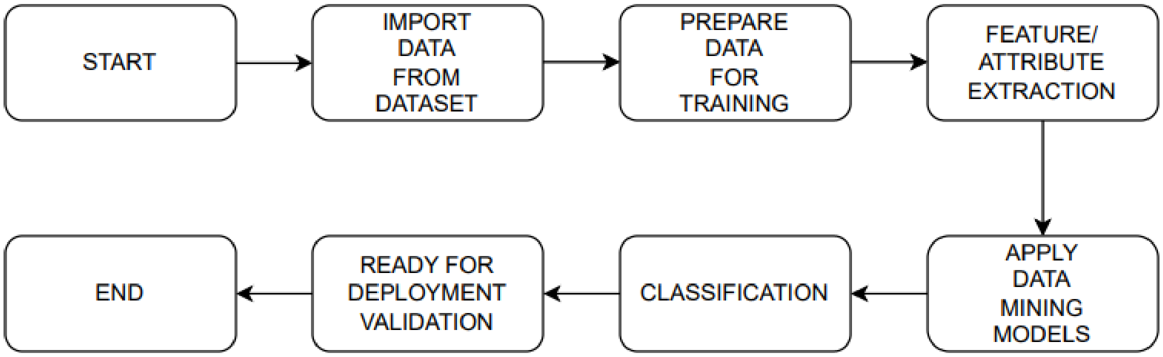
- We present a model for the classification of diabetes disease.
- We discuss different data mining techniques.
- The proposed mechanism is trained using Python for Random Forest, Decision trees, Bagging, and analyzed with a real dataset, which is collected from Kaggle.

PROPOSED FRAMEWORK

This section presents the proposed mechanism. The following figure shows the suggested model. A diabetes dataset is used for processing in the suggested mechanism. The dataset includes characteristics related to diabetes, including BP, cholesterol, skin thickness, BMI, sugar. Next, use the trained model to extract features from the dataset. Next, classify diabetes into two states—1 or 2—using the data mining models Bagging and Random Forest.

And then we interpret the results in the context of diabetes prediction and it is also very important to understand how the selected features contribute to the model's predictions.Once satisfied with the model's performance, deploy it for making predictions on new,unseen data.

Regular evaluation and refinement are essential for maintaining the model's effectiveness over time. Additionally, after training, the model is validated to retrieve the accuracy, sensitivity, and confusion matrix performance of the models.


START
IMPORT DATA FROM DATASET
PREPARE DATA FOR TRAINING
FEATURE/ ATTRIBUTE EXTRACTION
APPLY DATA MINING MODELS
CLASSIFICATION
READY FOR DEPLOYMENT VALIDATION
END

IMPLEMENTATION STRATEGY

The information base is the source of the datasets. The information will be pre-handled in stage two, which includes cleaning, combining, and altering the information. When compared to other calculations, we can obtain greater precision by using RF calculation. The data was obtained via Kaggle. In order to provide a good model, it was collected, and in keeping with this, the data was entered as instructional tests and thoroughly examined. The arrangement of relevant data that is put together using an exploitation question measure is known as data variety. With several highly specific categories, the data is divided in an extremely specific manner.

- DATA PREPROCESSING - A significant development in the theory of information disclosure is information pre-processing. Information about wheezy, irregular, and missing values can be found in most medical service data sets.
- DATA CLEANING - Finding and changing (or removing) bad or incorrect records from a record set, table, or information is known as data cleaning. It involves identifying parts of the information that are missing, inaccurate, erroneous, or distracting and replacing, adjusting, or removing the coarse or messy information.
- FEATURE SELECTION AND EXTRACTION - Feature selection is the process of selecting a subset of relevant features from the original set of features in a dataset. The goal is to choose features that contribute the most to the predictive performance of a model while disregarding irrelevant or redundant features. Feature extraction involves transforming the original set of features into a new set of features with reduced dimensionality while retaining as much of the relevant information as possible.

METHODOLOGY

- ATTRIBUTES IN THE DATA SET - High Bp, High Cholesterol, Cholesterol Check, BMI, Smoker, Stroke,Heart Disease, Physical Activities, Fruits, Veggies, Heavy Alcohol, Any Healthcare, NoDocbc cost, Gen Health, Mental Health, Physical Health, Diff Walk, Age, Education.

- WHAT IS RANDOM FOREST AND WHY RANDOM FOREST ?

  Random Forest is an ensemble learning method used for both classification and regression tasks in machine learning. It belongs to the family of bagging algorithms, which aim to improve the accuracy and robustness of a model by combining the predictions of multiple individual models. Random Forest is particularly known for its flexibility, simplicity, and effectiveness in handling complex datasets.

  For each tree in the Random Forest, a random subset of the training data is selected with replacement (bootstrapping). This introduces diversity among the trees.At each node of each tree, a random subset of features is considered for splitting. This randomness helps prevent individual trees from becoming too specialized and overfitting to the training data. Random forest converts low bias, high variance model into low bias, low variance model by training multiple decision trees at same time.

RESULTS

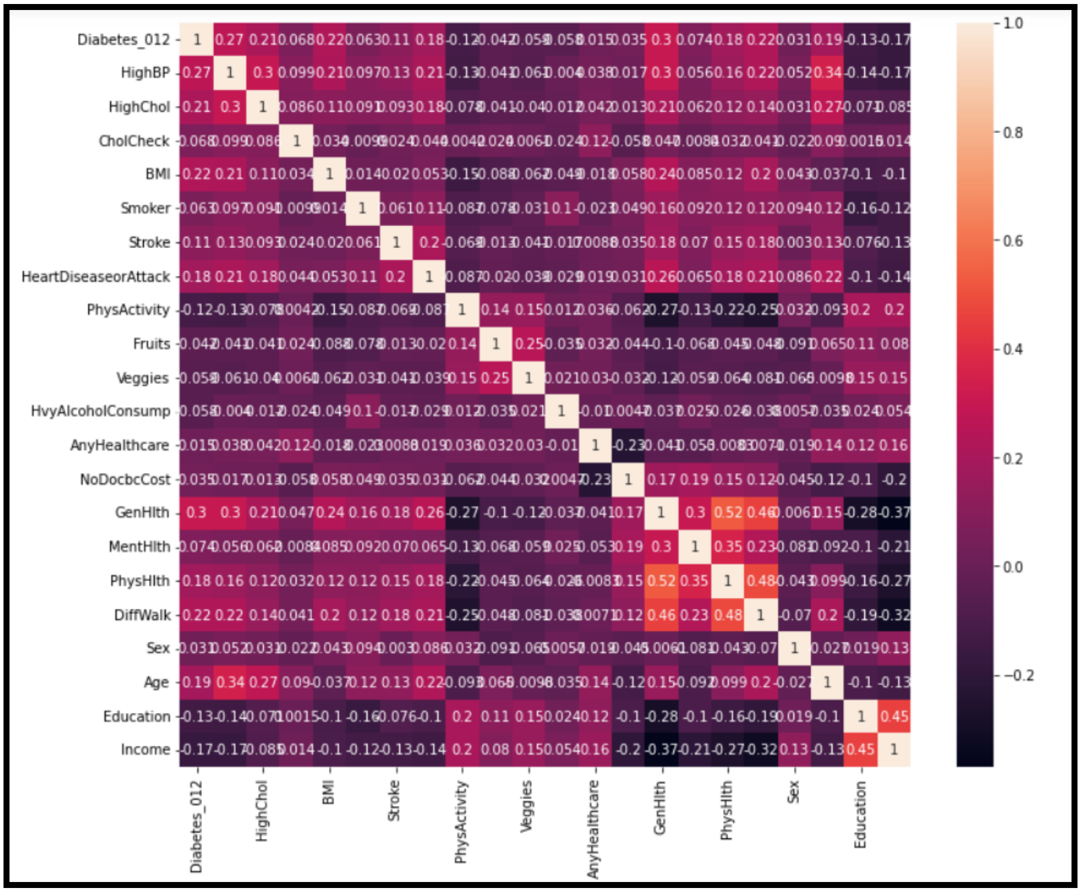
The execution of a programme is fast in Python. It can provide inbuilt library files to run the programs.

METRICS USED :

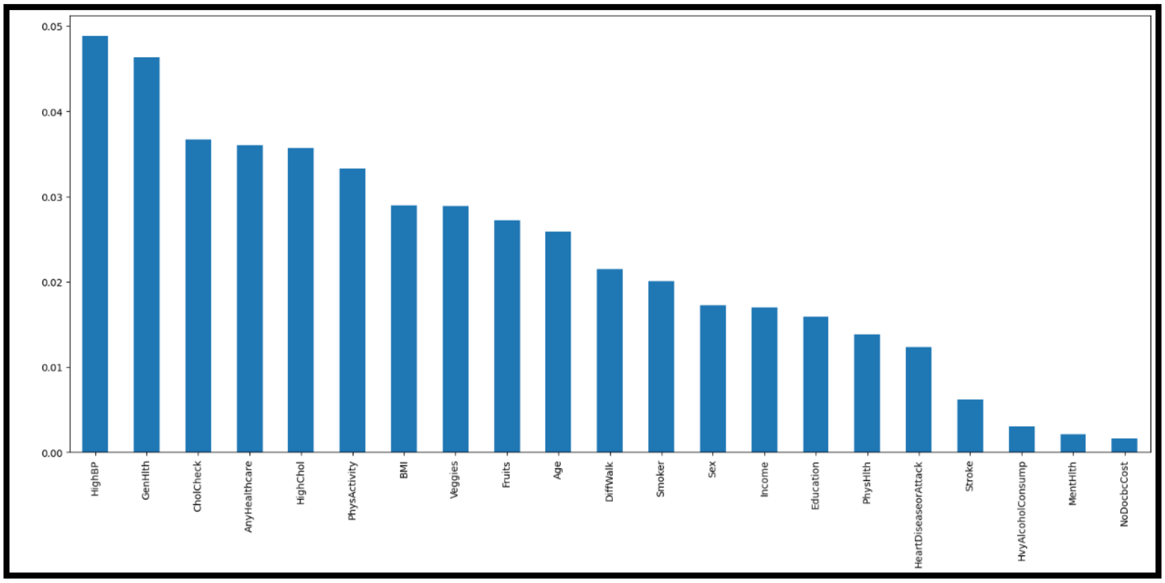
- *ACCURACY* - It shows the overall number of positive (+ve) outcomes in comparison to the total number of negative (-ve) outcomes in the entire dataset.

  $$Accuracy = TP + TN/\ (TP + FP + FN + TN)$$

- *CONFUSION MATRIX* - It depicts the prediction values of data in terms of TP, TN, FN, FP i.e., true + ve, true –ve, false + ve and false -ve. Based on these parameters the sensitivity and accuracy of techniques has been computed.
- *CLASSIFICATION REPORT* - A classification report is a summary of the performance metrics for a classification model. It provides a comprehensive view of how well the model is performing in terms of various evaluation metrics, particularly in the context of predicting categories or classes.



CORRELATION HEATMAP OF THE ATTRIBUTES



INFORMATION GAIN OF THE ATTRIBUTES

CONCLUSION AND FUTURE WORK

The development of a diabetes prediction model using Random Forest has proven to be a promising approach for identifying individuals at risk of diabetes. The Random Forest model, leveraging its ensemble of decision trees, demonstrated robustness, accuracy, and the ability to capture complex relationships within the dataset. Through rigorous evaluation and feature importance analysis, the model showcased its effectiveness in distinguishing between individuals with type 1 and type 2 diabetes.

- Further, we can develop a system for real-time monitoring of individuals' health metrics and integrate the predictive model to generate timely alerts and recommendations. This proactive approach can empower individuals to take preventive actions.
- We can also collaborate with healthcare professionals to validate the model's predictions in clinical settings. Gather feedback from practitioners to refine the model and ensure its practical utility in healthcare decision-making.
- We can further investigate the integration of multiple predictive models, possibly combining Random Forest with other machine learning algorithms. Ensemble methods can harness the strengths of different models, potentially leading to a more robust and accurate prediction system

We wish to express our sincere gratitude to our Professor Dr. Bharti for providing her valuable guidance, comments and suggestions throughout the course of the project.

REFERENCES

- Diabetes prediction model using data mining techniques - ScienceDirect
- Combining knowledge extension with convolution neural network for diabetes prediction - ScienceDirect
- A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures - ScienceDirect
- https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data

THE END