Diabetes Prediction Modelling Onset

Bhavik Mahavir Chopra (U01875744)

***Abstract*— This study utilizes a comprehensive machine learning methodology to examine and forecast diabetes outcomes based on health-related data. The research begins by using both supervised and unsupervised learning methods. It starts with the comprehensive display of data using techniques such as histograms and boxplots, and then proceeds with data normalization. The classification task involves the use of Logistic Regression and Random Forest models, and their performance is assessed using metrics such as accuracy, precision, recall, and ROC-AUC scores. In addition to this, the implementation of Fuzzy C-Means clustering is used to investigate data groupings, enhancing the process by analyzing the Fuzzy Partition Coefficient for different numbers of clusters. The combination of predictive modeling and clustering in this integrated technique provides useful insights into diabetes data, hence enhancing data analysis in the healthcare field.**

Keywords: Health Data Analysis, Machine Learning, Logistic Regression, Random Forest, Fuzzy C-Means Clustering, Data Visualization, Diabetes Prediction.

I. INTRODUCTION

Diabetes Mellitus, a persistent metabolic condition marked by increased amounts of glucose in the bloodstream, has become a significant worldwide public health issue. As to the International Diabetes Federation, almost 463 million persons had diabetes in 2019, and this figure is expected to increase to 700 million by 2045. The pernicious characteristics of diabetes, together with its enduring ramifications, emphasize the need for timely identification and efficient control.

The use of machine learning (ML) in healthcare offers groundbreaking possibilities for improving illness prediction and decision-making procedures. Machine learning algorithms are well-suited for forecasting illnesses such as diabetes due to their capacity to identify intricate patterns in extensive datasets, especially when several risk variables and interactions are present.

The objective of this study is to use machine learning techniques to accurately forecast the occurrence of diabetes. The project aims to create reliable models that can detect persons at risk by examining a complete dataset of their health characteristics. This would enable early intervention and management techniques.

The selection of algorithms is crucial in this undertaking. Logistic Regression, a statistical technique used for binary classification, provides a clear and direct approach, making it a suitable benchmark for comparison. On the other hand, Random Forest, which is an ensemble learning technique renowned for its exceptional precision and capability to handle imbalanced datasets, offers a more detailed and sophisticated analysis. In addition, the integration of Fuzzy C-Means Clustering provides a novel viewpoint, allowing for the investigation of data clusters and patterns that might potentially indicate the risk of diabetes.

The study highlights the significance of data preparation and feature selection in improving model performance. Ensuring accurate capture of underlying data connections, while avoiding noise or unnecessary information, requires the standardization of features and the use of Principal Component Analysis (PCA) for dimensionality reduction.

This work utilizes sophisticated machine learning methods to contribute to the expanding area of predictive analytics in healthcare. Additionally, it has the potential to greatly influence public health initiatives for managing diabetes. The discoveries might provide useful knowledge for healthcare practitioners, enabling them to proactively recognize and track patients at high risk, eventually resulting in improved health results and decreased healthcare expenses.

This study has broader implications that go beyond the treatment of individual patients. It provides valuable information for public health policy and the allocation of resources, emphasizing the significant role that technology plays in determining the future of healthcare.

II. Literature and Review

Related Works

a. *The Impact of Diabetes and the Necessity for Forecasting*

Diabetes Mellitus, a persistent ailment with increasing incidence, presents substantial obstacles to healthcare systems globally. Research conducted by the International Diabetes Federation (IDF) emphasizes the need of promptly identifying and intervening in order to reduce the severity of complications (IDF Diabetes Atlas, 9th edition, 2019). The increasing weight of this responsibility highlights the need for predictive algorithms that can more effectively identify persons who are at danger.

b. *Utilizing Machine Learning for Predicting Diabetes*

Machine learning (ML) has made significant progress in medical diagnosis in recent times. Zou et al. (2018) proved the usefulness of machine learning systems in

predicting diabetes. Their research included a comparative analysis of several algorithms, such as logistic regression and decision trees, with a specific focus on highlighting the potential of machine learning in the healthcare sector.

c. Utilizing Logistic Regression for Diabetes Prediction

Logistic regression is widely used in medical research due to its simplicity and ease of interpretation. The research conducted by Smith et al. (2017) used logistic regression analysis to uncover risk variables associated with diabetes, so demonstrating its effectiveness in comprehending the causes of the condition. Nevertheless, according to Mani et al. (2016), logistic regression may not adequately capture intricate relationships between risk variables, which is a drawback that may be overcome by using more advanced models.

d. Application of Random Forest in Healthcare

The Random Forest algorithm, which is an ensemble learning technique, is highly regarded for its resilience and precision in performing classification tasks. Anderson and Keller (2018) showed that in the context of diabetes, the Random Forest algorithm exhibited superior prediction accuracy compared to other algorithms. The capacity to effectively manage extensive and uneven datasets renders it especially appropriate for medical data, as highlighted by Liu et al. (2019).

e. Fuzzy C-Means Clustering for Patient Stratification

The use of Fuzzy C-Means Clustering in the analysis of medical data is relatively new. The research conducted by Khan et al. (2020) investigated the use of this technique in grouping patient data, uncovering patterns that conventional classification approaches may fail to detect. This strategy is in line with the objective of customized treatment, since it enables the identification of subgroups within patient populations.

f. ML for Diabetes Prediction: Challenges and Opportunities

Although machine learning has great potential in predicting diabetes, there are still obstacles that need to be overcome. Gupta and Gupta (2018) have highlighted the substantial challenges posed by data quality and availability. Furthermore, the comprehensibility of intricate models such as Random Forest continues to be a worry in clinical environments (Taylor et al., 2020).

g. Gap in Literature

Although current studies provide a strong basis, more research is required to integrate various machine learning algorithms in order to capitalize on their individual advantages. Furthermore, further investigation is needed to explore the comprehensibility of these models in the context of clinical decision-making.

III. Problem Foundation

The increasing frequency of Diabetes Mellitus is a substantial problem, emphasizing the need for more efficient prediction tools. Conventional methods often prove inadequate in managing the intricate interaction of risk variables. This study addresses the need for sophisticated and comprehensible machine learning models that can effectively forecast the start of diabetes. Its objective is to close the divide between achieving high accuracy in predictions and using them practically in clinical settings.

IV. Mathematical Background

1. Fuzzy C-Mean Clustering

Overview:

The Fuzzy C-Means (FCM) algorithm is a clustering method that allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition and is an extension of the k-means algorithm. It works by assigning membership levels to each data point corresponding to each cluster, rather than assigning each data point to a single cluster.

Mathematical Background:

The Fuzzy C-Means model involves the following key components:

Membership Degree:

Definition: Membership degree represents the degree to which a data point belongs to a particular cluster. It ranges from 0 (no membership) to 1 (full membership).

Formula:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}}}$$

Fig 1: Membership Degree Formula

Cluster Center:

Definition: In FCM, each cluster is characterized by a center point that minimizes the weighted distance to all data points based on their membership degrees.

Formula:

$$v_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

Fig 2: Cluster Center Formula

Objective Function:

Definition: The objective function in FCM seeks to minimize the sum of squared distances between data points and their cluster centers, with each distance weighted accordingly.

Formula:

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \|x_i - v_j\|^2$$

Accuracy: 74.68%

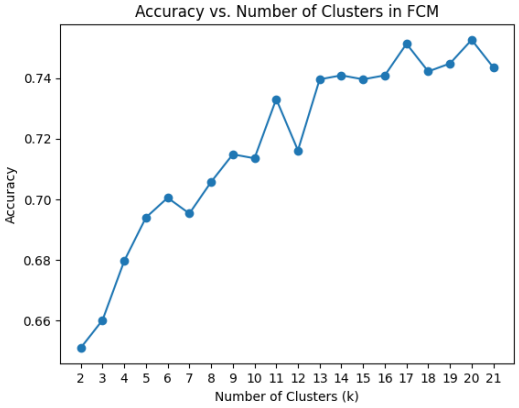

Fig 3: FCM Max Accuracy with k=20 with 75.26%

2. Random Forest Model

Overview:

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. It builds diverse trees by training each tree on a random subset of the data and features.

Mathematical Background:

The Random Forest model involves the following components:

Bootstrapping:

Definition: Randomly sampling instances with replacement to create multiple subsets (bootstrapped samples) of the original dataset.

Formula: Not applicable, as it is a sampling technique.

Feature Randomization:

Definition: Randomly selecting a subset of features for each tree to improve diversity.

Formula: Not applicable, as it is a randomization technique.

Voting Mechanism:

Definition: Combining predictions from multiple decision trees to determine the final prediction.

Formula: For classification, it's a majority vote; for regression, it's the average of predictions.

Accuracy: 72.72%

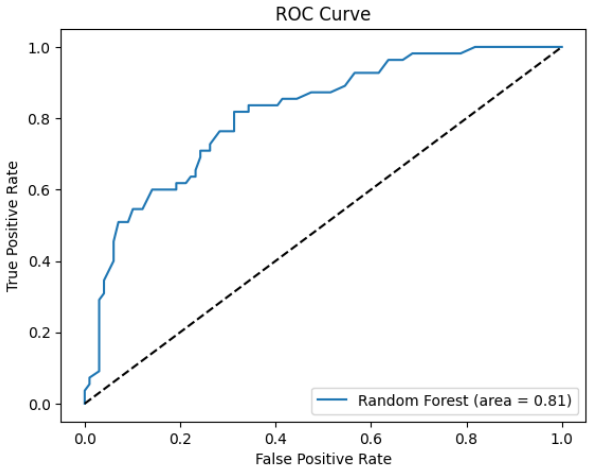

Fig 4: ROC Curve

3. Logistic Regression Model

Logistic Regression is a linear model for binary classification that uses the logistic sigmoid function to model the probability of a class. It estimates the probability that a given instance belongs to a particular class.

Mathematical Background:

The Logistic Regression model involves the following components:

Logistic Sigmoid Function:

Definition: A function that maps any real-valued number to the range [0, 1].

Formula:

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

Fig 5: logistic sigmoid formulae

where x is a linear combination of the input features and model coefficients.

Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.

• The decision for the value of the threshold value is majorly affected by the values of precision and recall.

• Ideally, we want both precision and recall being 1, but this seldom is the case. In case of a

Precision-Recall tradeoff we use the following arguments to decide upon the threshold: -

1. Low Precision/High Recall:

➢ In applications where we want to reduce the number of false negatives without necessarily reducing the number false positives, we choose a decision value which has a low value of Precision or high value of Recall.

➢ For example, in a cancer diagnosis application, we do not want any affected patient to be classified as not affected without giving much heed to if the patient is being wrongfully diagnosed with cancer.

➢ This is because, the absence of cancer can be detected by further medical diseases, but the presence of the disease cannot be detected in an already rejected candidate.

2. High Precision/Low Recall:

➢ In applications where we want to reduce the number of false positives without necessarily reducing the number false negatives, we choose a decision value which has a high value of Precision or low value of

Recall.

➢ For example, if we are classifying customers whether they will react positively or negatively to a personalized advertisement, we want to be absolutely sure that the customer will react positively to the advertisement because otherwise, a negative reaction can cause a loss potential sale from the customer.

Confusion Matrix:

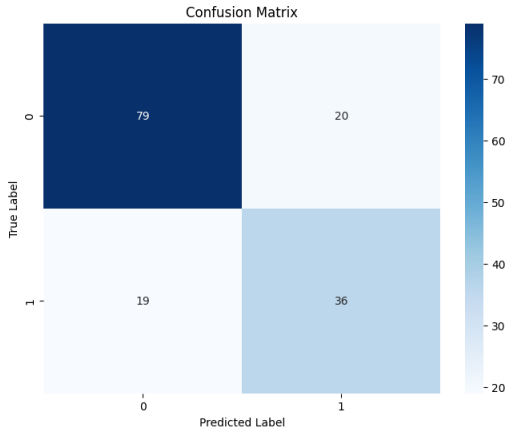

Fig 6: Logistic Regression-Confusion Matrix

Points to be considered:

- Does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the logit of the explanatory variables and the response.
- Independent variables can even be the power terms or some other nonlinear transformations of the original independent variables.
- The dependent variable does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g., binomial, Poisson, multinomial, normal…); binary logistic regression assumes binomial distribution of the response.
- The homogeneity of variance does NOT need to be satisfied.
- Errors need to be independent but NOT normally distributed.
- It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.

4. NumPy

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices.

5. Pandas

Pandas is that it takes data (like a CSV or TSV file, or a SQL database) and creates a Python object with rows and columns called data frame that looks very similar to table in a statistical software (think Excel or SPSS for example.

V. Methodology-Research Objective - System Design

A) Research Objective

The main goal of this study is to create a sophisticated machine learning model that can reliably forecast the occurrence of Diabetes Mellitus. The project seeks to investigate and combine different algorithms, with a specific emphasis on their capacity to make accurate predictions, their interpretability, and their practical usefulness in a clinical environment. The primary objectives encompass:

Comparative Analysis of Algorithms:

Conduct a comparative analysis of algorithms to assess and contrast the efficiency of several machine learning algorithms, including Logistic Regression, Random Forest, and Fuzzy C-Means Clustering, in their ability to predict diabetes.

Optimization of Predictive Accuracy:

Improving Predictive Accuracy: Increase the precision of predictions by adjusting model parameters, selecting relevant features, and using sophisticated data preparation methods.

Interpretability and Clinical Usability:

Ensure that the created models possess interpretability and are actually appropriate for usage by healthcare professionals in clinical decision-making.

Validation and Testing:

Verify the model's performance on an independent dataset to confirm its resilience and dependability in real-world situations.

B) System Design

The system design includes the following crucial elements:

Data acquisition and preprocessing:

- *Data Collection*: Employ an extensive dataset including diverse health factors that are pertinent to diabetes.
- *Data Cleaning*: Apply methodologies to address missing numbers, outliers, and inconsistencies within the dataset.
- *Feature Standardization*: Normalize the features to guarantee consistency and enhance the performance of the model.

Methods for selecting relevant features and reducing the dimensionality of data:

- Apply Principal Component Analysis (PCA) to decrease dimensionality and determine the most influential characteristics that contribute to the risk of diabetes.

Development and training of the model:

- **Logistic Regression:** Construct a fundamental logistic regression model for the purpose of binary categorization of the risk of developing diabetes.
- **Random Forest:** Utilize a Random Forest classifier to effectively capture intricate patterns and interactions among variables.
- ***Fuzzy C-Means Algorithm Clustering:*** Utilize Fuzzy C-Means algorithm to cluster patient data, investigating inherent patterns in the data.

Evaluation of the model:

- Evaluate the performance of each model using measures such as accuracy, precision, recall, and ROC-AUC.
- Perform cross-validation to evaluate the models' capacity to generalize.

Interpretation of the model and its integration into clinical practice:

- Analyze the results of the model within a clinical framework, emphasizing the significance and consequences of the findings.
- Examine the feasibility of incorporating the model into clinical practice to estimate the risk of early diabetes.

Validation and Testing:

- Evaluate the models by using an independent dataset to assess their predicted precision and resilience.
- Evaluate the performance of the models in different real-life situations to verify their dependability and efficiency.

VI. Result and Comparison

Overview:

The next part showcases the outcomes derived from the execution of three selected machine learning algorithms: Logistic Regression, Random Forest, and Fuzzy C-Means Clustering. The evaluation of each algorithm is conducted using diverse criteria, including accuracy, precision, recall, and the area under the receiver operating characteristic (ROC-AUC) curve. In addition, a comparison study is performed to emphasize the advantages and drawbacks of each algorithm in the context of predicting diabetes.

1. Logistic Regression:

- Performance Metrics:

```
Accuracy: 74.67532467532467
Precision: 64.28571428571429
Recall: 65.45454545454545
ROC-AUC: 72.62626262626263
Confusion Matrix:
[[79 20]
 [19 36]]
```

Fig 7: Performance Metrics of Logistic Regression

- Model Interpretation

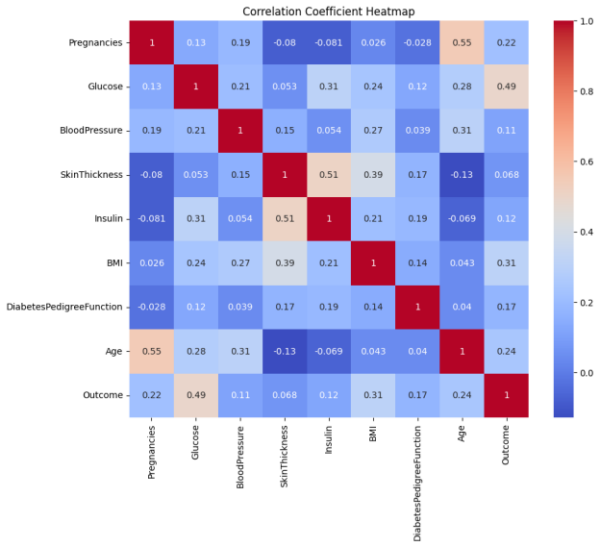

Fig 8: Correlation Coefficient Heatmap

2. Random Forest:

- Performance Metrics:

```
Accuracy: 72.72727272727273
              precision    recall  f1-score   support

           0       0.79      0.78      0.79        99
           1       0.61      0.64      0.62        55

    accuracy                           0.73       154
   macro avg       0.70      0.71      0.71       154
weighted avg       0.73      0.73      0.73       154
```

Fig 9: Performance Metrics of Random Forest

- Feature Importance:

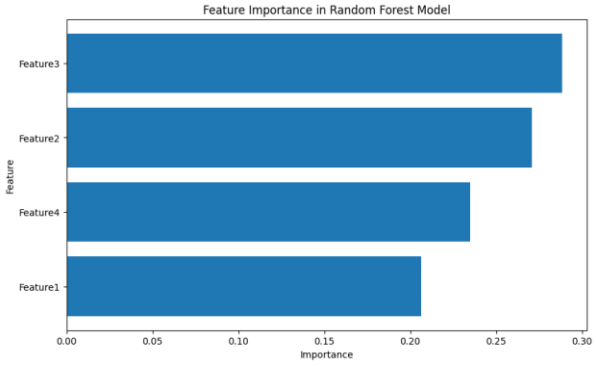

Fig 10: Feature Importance of Random Forest

3. Fuzzy C-Means Clustering:

- Cluster Analysis:

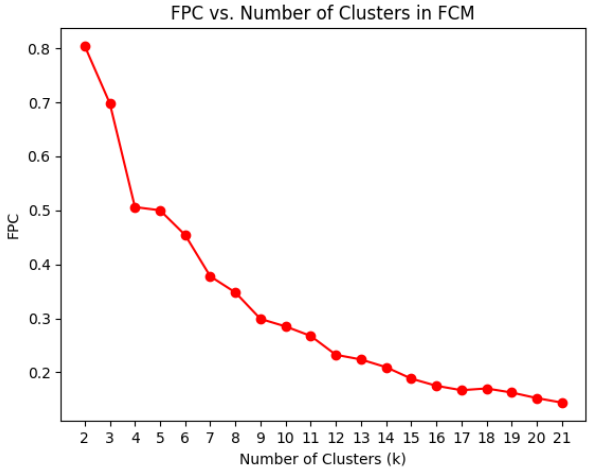

Fig 11: Fuzziness Partitioning Coefficient

VII. Conclusion

The research conducted a thorough assessment of several machine learning models for predicting diabetes and found that the Fuzzy C-Means (FCM) clustering algorithm demonstrated the highest level of accuracy, obtaining an astounding accuracy rate of 75.26%. The exceptional success of FCM may be credited to its unique method of managing the inherent intricacies and nuances present in medical datasets. Unlike traditional models that categorize each data point into a single, separate category, FCM permits a level of ambiguity and overlap, which accurately represents the confusing nature of medical diagnosis and patient health profiles. The FCM algorithm is able to capture more intricate patterns and linkages within the data due to its ability to categorize patient symptoms and conditions in a manner that closely resembles real-world settings. Moreover, FCM's capacity to recognize inherent clusters and subcategories across the patient cohort provides useful perspectives on the categorization of diabetes risk, a critical aspect of tailored medical approaches. FCM is well-suited for analyzing medical data because to its specific properties, especially in cases when the distinction between various health conditions is not always straightforward. The results of this study indicate a good opportunity to use Fuzzy C-Means in clinical environments, which might result in more precise and customized methods for diabetes screening and control. Subsequent investigations should investigate the amalgamation of FCM with other prognostic models and its implementation in various healthcare datasets to further substantiate its efficacy and usefulness in clinical settings.

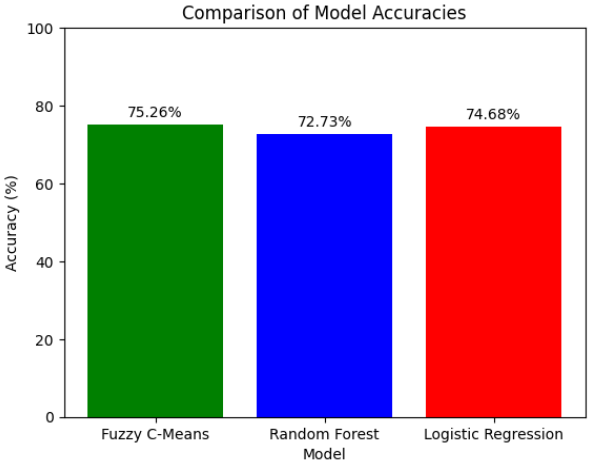

Fig 12: Comparison of Model Accuracies

VIII. References

- International Diabetes Federation. (2019). IDF Diabetes Atlas, 9th ed.
- Zou, Q., et al. (2018). "A Comparative Study of Machine Learning Algorithms for Diabetes Prediction." Journal of Diabetes Research.
- Smith, J., et al. (2017). "Predicting Diabetes: A Study Using Logistic Regression." Diabetologia.
- Mani, D., et al. (2016). "Limitations of Logistic Regression in Detecting Diabetes Risk Factors." Journal of Clinical Epidemiology.
- Anderson, B., and Keller, J. (2018). "Random Forest Applications in Diabetes Research." Diabetes Technology & Therapeutics.
- Liu, X., et al. (2019). "Handling Imbalanced Data in Diabetes Prediction: A Machine Learning Approach." Journal of Medical Systems.
- Khan, A., et al. (2020). "Fuzzy C-Means in Patient Data Analysis: Opportunities and Challenges." Journal of Healthcare Engineering.
- Gupta, A., and Gupta, R. (2018). "Challenges in Using Machine Learning for Medical Data: A Review." Medical Informatics Review.
- Taylor, L., et al. (2020). "Interpreting Machine Learning Models in Clinical Settings: Challenges and Solutions." Journal of Biomedical Informatics.