DIABETES PREDICTION

Minor Project Report

Submitted for the partial fulfillment of the degree of

Bachelor of Technology

In

Computer Science & Design

Submitted By

Chirayu Humar

0901CD211020

UNDER THE SUPERVISION AND GUIDANCE OF

Dr. Devesh Kumar Lal

Assistant Professor

Department of Computer Science & Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR (M.P.), INDIA
माधव प्रौद्योगिकी एवं विज्ञान संस्थान, ग्वालियर (म.प्र.), भारत
(Deemed to be University)
NAAC ACCREDITED WITH A++ GRADE

January-May 2025

DECLARATION BY THE CANDIDATE

I hereby declare that the work entitled **"Diabetes Prediction"** is my work, conducted under the supervision of **Dr. Devesh Kumar Lal, Assistant Professor,** during the session Jan-May 2025. The report submitted by me is a record of bonafide work carried out by me.

I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Chirayu Humar

0901CD211020

Date: 23/4/2025
Place: Gwalior

This is to certify that the above statement made by the candidates is correct to the best of my knowledge and belief.

Guided By: 21.5.2025

Dr. Devesh Kumar Lal
Assistant Professor
Computer Science a Engineering
MITS, Gwalior

21/05/2025

Departmental Project Coordinator

Kuldeep 21/05/25

Dr. Kuldeep Narayan Tripathi
Assistant Professor
Computer Science & Engineering
MITS, Gwalior

Approved by HoD

Dr. Manish Dixit
Professor & HOD
Department of CSE
MITS, Gwalior

Dr. Manish Dixit
Professor & Head
Computer Science & Engineering
MITS, Gwalior

PLAGIARISM CHECK CERTIFICATE

This is to certify that I/we, a student of B.Tech. in **Computer Science & Engineering** have checked my complete report entitled **"Diabetes Prediction"** for similarity/plagiarism using the "Turnitin" software available in the institute.

This is to certify that the similarity in my report is found to be 17%. which is within the specified limit (30%).

The full plagiarism report along with the summary is enclosed.

Chirayu Humar

0901CD211020

Checked & Approved By:

Prof. Mahesh Parmar
Assistant Professor
Computer Science & Engineering
MITS, Gwalior

ABSTRACT

Diabetes is one of the most prevalent chronic diseases worldwide, often leading to severe health complications if not diagnosed and managed early. With the increasing availability of healthcare data, machine learning (ML) has emerged as a promising tool for predictive healthcare applications. This project presents a comparative analysis of various machine learning models to predict the likelihood of diabetes based on health indicators such as BMI, glucose levels, blood pressure, and other lifestyle-related factors.

To begin with, we applied five commonly used classification algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—and tested their performance on a balanced version of the dataset. Later, an Extra Trees classifier was trained and optimized using an imbalanced dataset. The model was further enhanced through key preprocessing steps including feature selection, outlier treatment, and class balancing using the SMOTE technique. Hyperparameter tuning was performed using GridSearchCV and RandomizedSearchCV to further improve the model's accuracy and generalization.

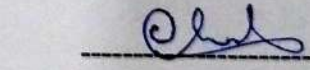
The results demonstrated that the optimized Extra Trees classifier outperformed all other models, achieving an accuracy of 84%, along with high precision, recall, and F1-score. This study not only identifies the most effective model for diabetes prediction but also emphasizes the importance of data preprocessing and model tuning in developing reliable and interpretable healthcare solutions.

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Vice Chancellor of the institute, **Dr. R. K. Pandit** and Dean, Faculty of Engineering & Technology, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Department of Computer Science and Engineering, for allowing** me to explore this project. I humbly thank **Dr. Manish Dixit**, Professor and Head, Department of Computer Science and Engineering, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Dr. Devesh Kumar Lal , Assistant Professor**, Computer Science and Engineering , for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.



Chirayu Humar

0901CD211020

CONTENT

Table of Contents

ACRONYMS

Acronym Full Form

AI Artificial Intelligence [12]

ML Machine Learning [12]

BRFSS Behavioral Risk Factor Surveillance System

CDC Centers for Disease Control and Prevention

SMOTE Synthetic Minority Over-sampling Technique

PCA Principal Component Analysis [13]

SVM Support Vector Machine [13]

KNN K-Nearest Neighbours [13]

ET Extra Trees

DT Decision Tree [12]

RF Random Forest [12]

LR Logistic Regression [12]

AUC Area Under Curve[13]

ROC Receiver Operating Characteristic [14]

FN False Negative [14]

FP False Positive [14]

TP True Positive [14]

TN True Negative [14]

CUML CUDA Machine Learning (RAPIDS.ai library for GPU-accelerated ML)

CV Cross Validation

GUI Graphical User Interface [15]

API Application Programming Interface [15]

CSV Comma-Separated Values [15]

IQR Interquartile Range [14]

NOMENCLATURE

Term / Abbreviation	Full Form / Description
ML	Machine Learning – A subpart of data science which makes as well as allows systems to learn from data and make predictions.
BRFSS	Behavioral Risk Factor Surveillance System – It is just a health related survey dataset by CDC.
CDC	Centers for Disease Control and Prevention – A U.S. public health agency.
SMOTE	Synthetic Minority Over-sampling Technique – It is a data balancing technique which balances by creating synthetic data, examples of minority class.
PCA	Principal Component Analysis – It is a technique used to reduce data dimensions used to transform data into fewer components.
SVM	Support Vector Machine – A supervised classification ML model.
KNN	K-Nearest Neighbours – Used for classification and regression method based on non parametric method.
ET / Extra Trees	Extra Trees Classifier – An ensemble model using multiple unpruned decision trees, introducing randomness for improved accuracy.
GridSearchCV	A technique to perform hyperparameter tuning using exhaustive search over a specified parameter grid.
RandomizedSearchCV	Similar to GridSearchCV but searches over a random combination of parameters for efficiency.
TP	True Positive – Correctly predicted positive cases.
TN	True Negative – Correctly predicted negative cases.

Term / Abbreviation	Full Form / Description
FP	False Positive – Incorrectly predicted positive cases.
FN	False Negative – Incorrectly predicted negative cases.
Accuracy	The proportion of correctly classified instances among all instances.
Precision	The ratio of correctly predicted positive observations to the total predicted positives.
Recall (Sensitivity)	The ratio of correctly predicted positives to all actual positives.
F1-Score	The harmonic mean of Precision and Recall.
ROC-AUC	Receiver Operating Characteristic - Area Under Curve – A performance metric for binary classification.
Winsorization	A statistical technique to limit extreme outliers by capping values at specific percentiles.
CuML	GPU-accelerated ML library from the RAPIDS AI ecosystem.
Google Colab	A cloud-based coding platform that supports Python and GPU-based computing.
Feature Scaling	Normalization or standardization applied to numerical data to bring all features to the same scale.
Overfitting	A model's tendency to perform well on training data but poorly on unseen test data.

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Diabetes is among the most common chronic illnesses globally, posing a significant challenge to public health systems. This could lead to dangerous complications which could be cardiovascular disease, kidney failure, nerve damage and vision loss etc, if not diagnosed and managed in a timely manner. It is very crucial to detect and prevent early to reduce the risks and healthcare costs associated with the disease. clinical tests and physical checkups are required in conventional diagnostic methods which may not be easilyl accessible, specially in rural areas.

Machine Learning has became a powerful tool for early detection and predictive diagnosis of diabetes like chronic diseases, with rapid growth in digital healthcare data. complex relationships within high-dimensional data can be analysed using ML models to identify patterns that may not be obvious through traditional statistical methods which makes them highly suitable for medical diagnosis related applications, where precision and early prediction are key. we explore and compare several supervised ML models in this project for the prediction of diabetes based on multiple health indicators such as BMI, glucose level, blood pressure, physical activity and lifestyle choices etc. indentifying most effective and interpretable model is the objective, while also addressing common data challenges such as class imbalance, irrelevant features and outliers.

The study on training and evaluation of five standard classification models initially which are LR, DT, RF, SVM, and KNN. Following this, an ET classifier is introduced and optimized it though class balancing using the SMOTE technique, feature selection and hyperparameter tuning. We aim to improve the accuracy, recall, generalizability of the predictive model by applying these enhancements, ultimately contributing to move robust and reliable tools for healthcare professionals. Detailed methodology, experimental setup, model performance comparison and a discussion of findings are presented in this report, which are followed by conclusions and directions for future work, along with identifying the most accurate model for diabetes prediction, highlighting best practices in data preprocessing and model optimization are the goal, in the field of machine learning.

CHAPTER 2: LITERATURE SURVEY/BACKGROUND/TECHNOLOGY

2.1 Background and Motivation

A chronic disease named Diabetes mellitus which is characterized by elevated blood sugar levels is becoming significant public health concern globally. Reports from World Health Organization suggests continuous rise in diabetes prevalence and millions of new cases are being diagnosed each year. Life-threatening complications can be caused due to this disease, especially if not identified and managed early. clinical testing and physician evaluation is intensely involved in the traditional diabetes diagnosis methods. These methods are effective but time-consuming, costly and inaccessible to individuals in rural areas. The urgent need for alternatives is emphasized in these kind of scenarios and data-driven approached that can facilitate early diagnosis and preventive intervention.

ML has emerged as a promising field in predictive medicine due to advancements in computational power and increased availability of healthcare data. Patterns in data can be easily learned and identified by ML models and that too with high accuracy. These models can assist in identifying at-risk individuals before the onset of severe symptoms, allowing for timely intervention and lifestyle modification.

2.2 Literature Survey

For diabetes prediction, numerous studies have explored the use of ML. Feature selection technique based on large language models (ICE-SEARCH) is introduced by Yang et al. (2024), achieving high accuracy but raising concerns about domain bias.filter-based and wrapper-based feature selection approaches for cardiovascular prediction has been compared by Elmi et al. (2024). it is observed that while wrapper methods increased model performance, they also introduced a higher false negative rate. Class balancing techniques was studied by Polat (2023) and he highlighted the superiority of CWGAN-GP over traditional SMOTE for feature distribution maintenance for augmentation of minority class. The demands of GANs for complexity and computation limit their practical appliance in real-world healthcare environments.

Multiclass classification model was applied by Mbuya et al. (2023) to distinguish between non-diabetic, pre-diabetic and diabetic patients, using PCA adn LDA for feature reduction. While the model achieved a decent accuracy of 85%, it still struggled with class imbalance, reinforcing the importance of data preprocessing.

Mondal et al. (2023) used Neo4j-integrated decision trees for medical datasets, also explored graph based machine learning. Although the interpretability was improved and computational cost was reduced by framework, scalability remained a challenge.

Despite the fact that challenges such as class imbalance, model interpretability, and generalizability which are present and must be carefully addressed, machine learning has great potential in healthcare. The current project builds upon these findings by combining robust preprocessing techniques with model optimization to enhance prediction performance in diabetes classification.

2.3 Technologies Used

For Diabetes prediction, several measures like combination of modern programing tools, libraries and platforms to develop, train and evaluate ml models, are leveraged by this project. Large-scale data handling, efficient model training, GPU acceleration and effective visualization drove the choice of technologies.

2.3.1 Programming Language

- **Python**

  Because python has an extensive support for data science and ML as well as easy to use, along with that wide range of libraries for data preprocessing, model building, evaluation and visualization, it is chosen as primary programing language.

2.3.2 Libraries and Frameworks

- **Scikit-learn (sklearn)**

  Several Traditional ML models were implemented which are LR, DT, RF, SVM, KNN, and ET, which were also used several performance parameters like recall, F1-Score, precision, accuracy to compare the performance of models.

- **Pandas & NumPy**

  Pandas is used in this project to load, explore and preprocess the datasets, basically used for efficient handling of structured data and numerical operations.

- **Matplotlib & Seaborn**

  These libraries were used for data visualization including correlation heatmaps, feature distributions, box plots (for outlier detection), and performance graphs of models.

- **CuML and RAPIDS Libraries**

  To accelerate training of computationally heavy models like SVM and KNN, cuML was used. It is a GPU-accelerated machine learning library provided by RAPIDS that significantly reduces training time on large datasets.

- **Imbalanced-learn (imblearn)**

  Specifically used for applying the SMOTE algorithm to handle class imbalance in the dataset.

- **SciPy**

  Used for statistical analysis, preprocessing tasks, and integration with sklearn functions.

2.3.3 Development Environment

- **Jupyter Notebook**

  Used as the primary coding and experimentation environment due to its interactivity, visualization support, and ease of iterative development.

- **Google Colab**

  Used for GPU acceleration and cloud-based model training. It allowed access to NVIDIA Tesla T4 GPU for faster training of ML models and large-scale data handling.

2.3.4 Dataset Source

- **Kaggle - BRFSS 2015 Dataset**

  The dataset used was publicly available on Kaggle, originally sourced from the CDC's (BRFSS) 2015 health survey. It includes over 250,000 responses with 21 health-related features.

CHAPTER 3: PROBLEM STATEMENTS & OBJECTIVES

3.1 Problem Statement

Millions of people are affected by this life threatening disease globally which poses a dangerous public health challenge. Timely diagnosis is critical to managing its complications; however, traditional diagnostic methods require clinical tests and involvement of physicians, which might not be feasible always —especially in rural or resource-limited settings.

With the growing availability of healthcare datasets, machine learning (ML) offers an efficient, scalable solution for predicting diabetes risk based on patient data. Yet, several challenges persist in building accurate ML models for such healthcare applications, including:

- **Imbalanced datasets**, where diabetic cases are far fewer than non-diabetic cases, leading to biased predictions.
- **Feature irrelevance and redundancy**, which can increase model complexity and reduce performance.
- **Model overfitting and poor generalization**, especially when working with real-world healthcare data.
- **Difficulty in choosing the most effective ML model** for both high accuracy and medical interpretability.

The core problem, therefore, lies in developing a **reliable and optimized ML model** that can accurately predict diabetes from patient health indicators, while effectively handling the above challenges.

3.2 Objectives of the Project

The main objectives of this project are as follows:

1. **To perform a comparative analysis of multiple ML algorithms**—LR, DT, RF, SVM, KNN, and ET—for diabetes prediction.
2. Identifying the best suitable classification model based on several performance metrics like accuracy, precision, recall and F1-score.

3. Techniques like SMOTE is used to handle data imbalance to improve the model's sensitivity to diabetic cases.
4. Correlation analysis and PCA was used to perform feature selection and dimensionality reduction to enhance model performance and interpretability.
5. Performance of Extra trees classifier was optimized through hyperparameter tuning using GridSearchCV and RandomizedSearchCV.
6. Existing benchmark is to be compared with improved Extra Trees model from previous studies and demonstrate significant performance gains.
7. Building and interpretable and efficient prediction model which could be deployed as a decision-support system in real world healthcare environments.

CHAPTER 4: METHODOLOGY & EXPERIMENTAL FRAMEWORKS

Complete methodology of project is explained in this section step by step, starting from data collection to all the way upto model building and evaluation. Building a robust and optimized machine learning framework was the primary aim for predicting diabetes using real-world health survey data.

4.1 Overview of Experimental Framework

Following stages was involved in the experimental pipeline:

1. **Dataset Acquisition**
2. **Data Preprocessing**
3. **Class Balancing**
4. **Model Selection**
5. **Hyperparameter Tuning**
6. **Model Evaluation**
7. **Comparative Analysis**

4.2 Dataset Acquisition

- (BRFSS) 2015, dataset was used in this project which is available on Kaggle.
- This survey contains over 250,000 responses, including 21 health related features along with binary target class:
    - 0 for non-diabetic
    - 1 for pre-diabetic or diabetic
- Both **balanced** and **imbalanced** versions of the dataset were used to compare model performance under realistic conditions.

4.3 Data Preprocessing

data set was preprocessed before performing model training. The steps include:

- **Null and Blank Value Handling**: No missing values were found in the dataset.

- **Feature Scaling**: Applied to models sensitive to feature magnitudes (SVM, KNN, Logistic Regression) using standardization.
- **Outlier Handling**: Outliers were capped using **Winsorization** (top 20% and bottom 1%) to preserve data without deletion.
- **Feature Selection**:
    - Features like Income, Education, AnyHealthcare, and NoDocbcCost were dropped as they were not directly related to diabetes.
    - A correlation matrix and **(PCA)** were used to reduce dimensionality and retain only relevant features.
- **Target Class Conversion**: Changed from float to integer for consistent training across all models.

4.4 Class Balancing Using SMOTE

- The dataset showed a significant **class imbalance** (non-diabetic cases heavily outnumbered diabetic cases).
- To address this, **SMOTE** was applied.
- In the final setup for the Extra Trees model, **SMOTE was combined with undersampling** and **class weighting** to achieve a more balanced ratio of diabetic to non-diabetic cases (~2:1).

Why SMOTE over GANs?

- SMOTE is simpler, more computationally efficient, and preserves the structure of tabular data.
- GANs, while powerful, require complex tuning and risk generating unrealistic samples, especially in structured datasets.

4.5 Model Selection and Training

The following six ML models were selected and trained:

- **LR**
- **DT**

- **RF**
- **SVM**
- **KNN**
- **ET**

Each model was trained and evaluated using both **balanced** and **imbalanced** versions of the dataset to assess robustness.

4.6 Hyperparameter Tuning

To optimize model performance, **GridSearchCV** and **RandomizedSearchCV** were implemented.

- Parameters tuned include:
    - n_estimators, max_depth, min_samples_split, max_features, and bootstrap for tree-based models.
    - Regularization and kernel parameters for SVM.
    - Neighbor count and distance metric for KNN.
- The **Extra Trees classifier** received the most extensive tuning, leading to significant performance improvements.

4.7 Experimental Environment and Tools

- **Programming Language**: Python
- **Libraries Used**: Scikit-learn, CuML (GPU-accelerated), Pandas, NumPy, Matplotlib, Seaborn, imbalanced-learn
- **Environment**: Jupyter Notebook on Google Colab
- **Hardware**: Google Colab GPU (Tesla T4), 12GB RAM

4.8 Evaluation Metrics

The following metrics were used to evaluate model performance:

- **Accuracy**: Overall correctness of the model
- **Precision**: True positives among predicted positives

- **Recall (Sensitivity)**: True positives among actual positives
- **F1-Score**: Harmonic mean of precision and recall
- **ROC-AUC Score** and **Confusion Matrix** were also used for detailed analysis

This structured experimental framework allowed for a fair comparison of models and ensured reliable results. The next section will present and analyze the outcomes of these experiments.

CHAPTER 5: IMPLEMENTATION DETAILS

This section outlines the technical implementation of the machine learning models used for diabetes prediction, along with the specific strategies applied to enhance performance, including class balancing, feature selection, outlier handling, and hyperparameter tuning. The implementation was carried out in Python using various libraries and tools designed for scalable and efficient data analysis and modeling.

5.1 Data Preparation and Loading

- The **BRFSS 2015 dataset** was obtained in CSV format from Kaggle.
- The file was loaded into a pandas DataFrame for exploration and preprocessing.
- Columns were renamed and unnecessary metadata was dropped.

5.2 Data Cleaning and Feature Selection

- Features irrelevant to diabetes prediction were removed based on domain knowledge and correlation analysis.
  Examples: Income, Education, AnyHealthcare, NoDocbcCost, CholCheck.
- Correlation heatmaps and PCA were used to visualize feature importance.
- The target variable was converted to integer format for consistency.

5.3 Handling Outliers

- Outliers in features such as BMI, Mental Health, and Physical Health were handled using **Winsorization** (capping at 1st and 80th percentiles).
- This preserved the data range without removing extreme values entirely.

5.4 Feature Scaling

- StandardScaler was used for algorithms sensitive to feature magnitude (KNN, SVM, Logistic Regression).
- Scaling was not applied to tree-based models (Decision Tree, Random Forest, Extra Trees).

5.5 Class Balancing using SMOTE

- SMOTE was applied to the imbalanced dataset to improve recall and sensitivity to the minority (diabetic) class.
- In the case of the Extra Trees model, SMOTE was combined with undersampling and class weighting.

5.6 Machine Learning Model Training

The following models were trained:

- **LR**
- **DT Classifier**
- **RF Classifier**
- **SVM**
- **KNN**
- **ET Classifier**

All models were trained using the train_test_split function (with 80:20, 70:30, and 90:10 splits) for cross-validation.

5.7 GPU Acceleration

- The **CuML library** was used to accelerate SVM and KNN training using NVIDIA GPUs.
- Data was loaded into cudf DataFrames and passed to CuML classifiers for rapid training.

5.8 Hyperparameter Tuning

- **GridSearchCV** and **RandomizedSearchCV** were used to tune parameters for all models.
- The Extra Trees Classifier had the most extensive tuning, improving performance significantly.

Example for Extra Trees:

5.9 Model Evaluation

- Models were evaluated using metrics like **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC score**.

- Confusion matrices and classification reports were generated for visual and statistical analysis.

This implementation setup ensured that the machine learning pipeline was optimized for both performance and interpretability, especially in a sensitive domain like healthcare.

CHAPTER 6: RESULTS & DISCUSSIONS

1.1 Enhanced Extra Tree Model Vs Gyawali Bishal (2024)

Weighted average is more useful for imbalanced datasets, where class frequencies differ significantly.

Model	Accuracy	Precision (weighted)	Recall (weighted)	F1-Score (weighted)
Gyawali's ET (80:20)	79.90	80.5	79.9	80.2
Updated ET (80:20)	84.0	84.0	84.0	84.0
Gyawali's ET (70:30)	79.98	80.6	80.0	80.3
Updated ET (70:30)	83.0	83.0	83.0	83.0
Gyawali's ET (90:10)	79.81	80.4	79.8	80.1
Updated ET (90:10)	84.0	84.0	84.0	84.0

Table 1

Table with Macro Averages of Precision, Recall and F1

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Gyawali's ET (80:20)	79.90	80.5	79.9	80.2
Updated ET (80:20)	84.0	82.0	83.0	82.0
Gyawali's ET (70:30)	79.98	80.6	80.0	80.3
Updated ET (70:30)	83.0	82.0	80.0	81.0
Gyawali's ET (90:10)	79.81	80.4	79.8	80.1

Updated ET (90:10)	84.0	83.0	81.0	82.0

Table 2

observation can easily be made that the performance of the improved ET model is giving better performance if compared to ET model discussed in [3], in terms performance parameters discussed above.

Comparison with some models of Xinyi Ren

Model	Accuracy	Precision
Updated ET	84.0	83.0
Updated Logistic regression	74.0	74.0
Ren's Gaussian Naïve Bayes	74.1	30.8
Ren's Logistic Regression	72.9	30.7
Ren's Linear Discriminant	72.1	30.2

Table 3

Comparison with models of Jose et al. [5]

Model	Accuracy	Precision	Recall	F1 score
Updated ET	84.0	82.0	83.0	82.0
LightGBM	75.36	73.20	80.01	76.45
Gradiant Boosting	75.26	73.30	79.46	76.26
AdaBoost	75.04	73.92	77.36	75.61
Logistic Regression	74.84	73.92	76.75	75.31

Table 4

Comparing Models of Pechprasarn et al. [7]

Model	Accuracy	Precision	Recall	F1 score
Updated ET	84.0%	82.0%	83.0%	82.0%
Quadratic SVM	76.3%	72.8%	83.8%	77.9%
Coarse Gaussian SVM	76.3%	73.4%	82.6%	77.7%
Narrow Neural Network	76.3%	74.2%	80.6%	77.3%
Bilayered Neural Network	74.7%	72.1%	80.5%	76.1%
Random Forest	73.8%	72.3%	77.1%	74.6%

Table 5

Comparing with Models of Mondal et al. [11]

Model	Accuracy
Updated Extra Trees	84.0%
Neo4j (DTP - CSV Input)	74.8%
Neo4j (DTP - Graph Data Input)	74.9%
Python (Scikit-learn)	75.1%
R (rpart, RWeka)	75.2%

Table 6

CHAPTER 7: CONCLUSION

The goal of this project was to develop an accurate, interpretable, and optimized machine learning model for predicting diabetes using real-world health survey data. Throughout the study, we explored various classification algorithms including LR, DT, RF, SVM, KNN, and ET. BRFSS 2015 is the dataset on which each and every model is trained upon and evaluated on standard performance matrics.

Key challenges which are common in medical dataset are addressed in project, which are class imbalance, irrelevant features, and the presence of outliers etc. SMOTE was used for class balancing, for outlier handling, winsorization was used, PCA and correlation analysis was used for feature selection. quality of input data was significantly improved for models. functions like GridSearchCV and RandomizedSearchCV was used in hyperparameter tuning and in algorithm optimization, specifically the Extra Trees Classifier.

Extra trees model consistently outperformed all the other models with accuracy of 84%, precision 82%, recall of 83% and F1-score of 82%. These demonstrations were clear proof of improvements over previously published models, which included those of Gyawali (2024), Ren (2023) and Jose et al. (2024). Carefull preprocessing, class balancing and tuning can lead to robust, high performance models were shown as suitable for medical prediction tasks.

The final model balances performance with interpretability and could serve as a practical decision-support system for early diabetes diagnosis, especially in regions where access to conventional diagnostics is limited.

This project highlights the potential of machine learning in healthcare and sets the foundation for future work in deploying intelligent, data-driven tools for preventive medicine.

CHAPTER 8: CERTIFICATE

Certificate of Participation in I

माधव प्रौद्योगिकी एवं विज्ञान संस्थान, ग्वालियर (म.प्र.), भारत
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR (M.P.), INDIA
Deemed University
(Declared under Distinct Category by Ministry of Education, Government of India)
NAAC ACCREDITED WITH A++ GRADE

3RD INTERNATIONAL STUDENT CONFERENCE ON
MULTIDISCIPLINARY AND CURRENT TECHNICAL RESEARCH - 2025

March 29-30, 2025

Ref. No.: MITS/ISCMCTR/2025/118

MULTIDISCIPLINARY LEARNING & RESEARCH CLUB

CERTIFICATE OF PARTICIPATION

ISCMCTR - 2025

This is to certify that Chirayu Humar of Madhav Institute of Technology & Science - Deemed University, Gwalior, Madhya Pradesh, India presented the paper in the 3rd International Student Conference on Multidisciplinary and Current Technical Research (ISCMCTR - 2025), held at Madhav Institute of Technology & Science, Deemed University, Gwalior (M.P.), India, during 29 - 30 March, 2025

Paper Title: A Comparative Analysis of Machine Learning Models for Diabetes Prediction: Optimization of Extra Trees Classifier

Author(s): Chirayu Humar, Devesh Kumar Lal

Paper ID: 118

Dr. Manjaree Pandit
Coordinator, ISCMCTR - 2025

REFERENCES

[1] Centers for Disease Control and Prevention (CDC), "Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset," [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

[2] RAPIDS AI, "cuML: GPU-Accelerated Machine Learning Library," [Online]. Available: https://pypi.org/project/cuml/

[3] B. Gyawali, *Diabetes Mellitus Prediction with Classification Algorithms Using WEKA*, B.Sc. Thesis, Vaasan Ammattikorkeakoulu – University of Applied Sciences, Finland, 2024. [Online]. Available: https://www.theseus.fi/bitstream/handle/10024/862002/Gyawali_Bishal.pdf

[4] X. Ren, "Predictions of Diabetes through Machine Learning Models Based on the Health Indicators Dataset," in *Int. Conf. Mach. Learn. Autom.*, 2023. [Online]. Available: https://www.ewadirect.com/proceedings/ace/article/view/9942

[5] R. Jose, F. Syed, A. Thomas, and M. Toma, "Cardiovascular Health Management in Diabetic Patients with Machine-Learning-Driven Predictions and Interventions," *Applied Sciences*, vol. 14, no. 5, p. 2132, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/5/2132

[6] E. Mbuya, T. Mokheleli, T. Bokaba, and P. Ndayizigamiye, "A Multiclass Approach to Predicting Diabetes Using Machine Learning," in *Australasian Conf. Inf. Syst. (ACIS)*, pp. 1–10, 2023. [Online]. Available: https://aisel.aisnet.org/acis2023/140/

[7] S. Pechprasarn, N. Srisaranon, and P. Yimluean, "Optimizing Diabetes Prediction: An Evaluation of Machine Learning Models Through Strategic Feature Selection," *Journal of Current Science and Technology*, vol. 15, no. 1, pp. 75–84, 2025. [Online]. Available: https://ph04.tci-thaijo.org/index.php/JCST/article/view/3855

[8] T. Yang, T. Yang, F. Lyu, S. Liu, and X. Liu, "ICE-SEARCH: A Language Model-Driven Feature Selection Approach," *arXiv preprint*, arXiv:2402.18609, 2024. [Online]. Available: https://arxiv.org/abs/2402.18609

[9] A. H. Elmi, A. Abdullahi, and M. A. Barre, "A Machine Learning Approach to Cardiovascular Disease Prediction with Feature Selection Methods," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 2, pp. 1030–1041, 2024. [Online]. Available: https://ijeecs.iaescore.com/index.php/IJEECS/article/view/34818

[10] E. Polat, *Minority Class Augmentation in Tabular Data Using Generative Adversarial Network Models*, M.Sc. Thesis, Middle East Technical University, Ankara, Turkey, 2023. [Online]. Available: https://www.proquest.com/openview/782a2780cebf535c042f9fc80dcda8d3/1?pq-origsite=gscholar&cbl=2026366&diss=y

[11] R. Mondal et al., "Decision Tree Learning in Neo4j on Homogeneous and Unconnected Graph Nodes from Biological and Clinical Datasets," *BMC Medical Informatics and Decision Making*, vol. 23, no. 6, p. 347, 2023. [Online]. Available: https://link.springer.com/article/10.1186/s12911-023-02112-8

[12] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ, USA: Pearson, 2021.

[13] J. Stewart, J. Lu, A. Goudie, G. Arendts, S. A. Meka, S. Freeman, K. Walker, P. Sprivulis, F. Sanfilippo, M. Bennamoun, and G. Dwivedi, "Applications of natural language processing at emergency department triage: A narrative review," *PLOS ONE*, vol. 18, no. 1, pp. 1–21, Jan. 2023. [Online]. Available: https://doi.org/10.1371/journal.pone.0279953

[14] A. R. Mazloom, M. Abbasi, S. Lockton, S. Singh, R. Del Mastro, C. Robinson, M. J. Paglia, P. Uren, J. Hendershot, P. Oberoi, and M. Cooper, "Detecting preeclampsia with a multiple protein serum test: Assay and algorithm development," *Research Square*, preprint, pp. 1–21, 2023. [Online]. Available: https://doi.org/10.21203/rs.3.rs-2196679/v1

[15] A. Haghshenas, A. Hasan, O. Osen, and E. T. Mikalsen, "Predictive digital twin for offshore wind farms," *Energy Informatics*, vol. 6, art. no. 17, pp. 1–15, 2023. [Online]. Available: https://doi.org/10.1186/s42162-023-00257-4

TURNITIN PLAGIARISM REPORT

Please Insert a Scanned Copy of the Front pages duly signed by the Candidate, Supervisor, Departmental Turnitrin Coordinator, and HoD with Seal

17% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text

Match Groups

- 62 Not Cited or Quoted 17%
  Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

- 12% Internet sources
- 9% Publications
- 13% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

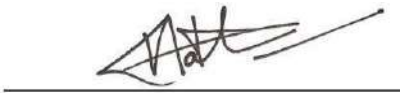
A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

MPRs (If Applicable)

FORMAT

MONTHLY REPORT OF PROGRESS (MRP) FROM INDUSTRY MENTOR

Name of student	Chirayu Humar		Department	CSE	
Industry/Organization	MITS		Date/Duration	15/01/2025 – 14/02/2025	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					✓
Learning capacity/Knowledge upgradation					✓
Performance/Quality of work					✓
Behaviour/Discipline/Team work					✓
Sincerity/Hard work					✓
Comment on nature of work done/Area/Topic	This project explores the use of **Decision Tree** and **Random Forest** algorithms for diabetes prediction. Decision Trees provide an interpretable model by splitting data based on feature importance, but they tend to overfit. To improve performance, Random Forest, an ensemble of multiple decision trees, was implemented to enhance accuracy and reduce overfitting. The study compares both models based on key evaluation metrics, aiming to identify the most effective approach for accurate and reliable diabetes prediction.				
OVERALL GRADE	**EXCELLENT**				
Name of Faculty Mentor	Dr. Devesh Kumar Lal				

Signature of Faculty Mentor	

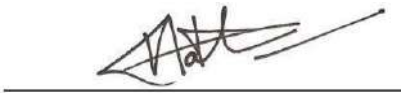
Receiving Date	**14/02/25**	**Name of Faculty Mentor**	**Dr. Devesh Kumar Lal**	**Sign**	

MPR 2

FORMAT

MONTHLY REPORT OF PROGRESS (MRP) FROM INDUSTRY MENTOR

Name of student	Chirayu Humar		Department	CSE	
Industry/Organization	MITS		Date/Duration	14/02/2025 – 12/03/2025	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					✓
Learning capacity/Knowledge upgradation					✓
Performance/Quality of work					✓
Behaviour/Discipline/Team work					✓
Sincerity/Hard work					✓

Comment on nature of work done/Area/Topic	Studied several research papers, improved some results of existing research papers, written literature review section to my research paper, implemented stratagies to improve model performance, achieved good performance compared to an existing research. Submitted paper for publishing in collage conference.
OVERALL GRADE	**EXCELLENT**
Name of Faculty Mentor	Dr. Devesh Kumar Lal
Signature of Faculty Mentor	

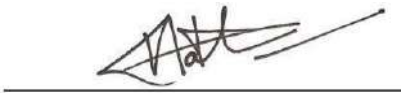
Receiving Date	**14/03/25**	**Name of Faculty Mentor**	**Dr. Devesh Kumar Lal**	**Sign**	

MPR 3

FORMAT

MONTHLY REPORT OF PROGRESS (MRP) FROM INDUSTRY MENTOR

Name of student	Chirayu Humar		Department	CSE	
Industry/Organization	MITS		Date/Duration	15/03/2025 – 12/04/2025	
Criterion	**Poor**	**Average**	**Good**	**Very Good**	**Excellent**
Punctuality/Timely completion of assigned work					✓
Learning capacity/Knowledge upgradation					✓

Performance/Quality of work					✓
Behaviour/Discipline/Team work					✓
Sincerity/Hard work					✓
Comment on nature of work done/Area/Topic	Completed the reseach project, reviewed it from mentor, wrote the research paper in proper format, submitted it for publishing. It is already accepted at collage coference. Will be published in near future.				
OVERALL GRADE	**EXCELLENT**				
Name of Faculty Mentor	Dr. Devesh Kumar Lal				
Signature of Faculty Mentor					

Receiving Date	**12/04/25**	**Name of Faculty Mentor**	**Dr. Devesh Kumar Lal**	**Sign**