Project Report On

**Machine Learning Algorithms for Diabetes**

****

Submitted in the partial fulfillment for the award of Post Graduate Diploma in Big Data Analytics (PG-DBDA)

from Know-IT ATC, CDAC ACTS, Pune

**Guided by:**

**Mr. Milind Kapse**

Submitted By:

Aditya Sanjay Awati(24084302503)

Prachi Sohani(240843025027)

Rutuja Deepak Jadhav (240843025055)

Disha Vijay Dalvi(240843025013)

# CERTIFICATE

**TO WHOMSOEVER IT MAY CONCERN**

### This is to certify that

Aditya Sanjay Awati(24084302503)

Prachi Sohani(240843025027)

Rutuja Deepak Jadhav (240843025055)

Disha Vijay Dalvi(240843025013)

### have successfully completed their project on

**Machine Learning Algorithms for Diabetes Prediction**

**Under the guidance of Mr. Milind Kapse**

# ACKNOWLEDGEMENT

This project **“Machine Learning Algorithms for Diabetes Prediction”** was agreat learning experience for us and we are submitting this work to Know-IT ATC, CDAC ACTS,Pune.

We are all very glad to mention the name of **Mr. Milind Kapse** for his valuable guidance on this project. Their continuous guidance and support helped us overcome various obstacles and intricacies during the course of the project work.

We are highly grateful to **Mr. Vaibhav Inamdar,** the Center Coordinator at Know-it in Pune, for his guidance and support whenever necessary while we were pursuing the Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS in Pune.

Our most heartfelt thanks go to **Mr. Shrinivas Jadhav** (Vice-President, Know-it, Pune) who provided all the required support and his kind co-ordination to provide all the necessities like required hardware, internet facility and extra lab hours to complete the project, throughout the course and till date, here in Know-IT ATC, CDAC ACTS, Pune.

#### From:

**Aditya Sanjay Awati (24084302503)**

**Prachi Sohani (240843025027)**

**Rutuja Deepak Jadhav(240843025055)**

**Disha Vijay Dalvi (240843025013)**

## TABLE OF CONTENTS

**ABSTRACT**

1. **INTRODUCTION**
2. **SYSTEM REQUIREMENTS**
   1. Software Requirements
   2. Hardware Requirements
3. **FUNCTIONAL REQUIREMENTS**
4. **SYSTEM ARCHITECTURE**
5. **METHODOLOGY**
6. **MACHINE LEARNING ALGORITHMS**
7. **DATA VISUALIZATION AND REPRESENTATION**
8. **CONCLUSION AND FUTURE SCOPE**
9. **REFERENCES**

## ABSTRACT

Diabetes is a chronic disease that affects millions worldwide. Early detection and prediction of diabetes can help individuals take preventive measures to manage their health. This project focuses on building a machine learning model to predict diabetes based on survey data collected by the CDC BRFSS2015 dataset. The study involves data preprocessing, exploratory data analysis (EDA), feature selection, and model training using various ML algorithms such as Random Forest, XGBoost, CatBoost,KNN. The goal is to identify patterns that can aid in the early diagnosis of diabetes.

## INTRODUCTION

Diabetes is a chronic metabolic disorder that affects millions worldwide, characterized by high blood glucose levels due to the body's inability to produce or use insulin effectively. Early detection and management of diabetes are crucial for preventing severe complications such as cardiovascular diseases, kidney failure, and neuropathy. Given the increasing availability of health data, machine learning techniques have emerged as powerful tools for predicting diabetes risk and providing personalized recommendations for disease management.

* The system takes user input on health parameters, processes the data, and provides a prediction in real time. The dataset used for training is derived from **CDC's BRFSS2015 dataset**, but only two classes will be considered:
* **0** → No Diabetes
* **1** → Diabetes

The main objectives of this project are:

* Develop an machine learning model for binary diabetes classification.
* Build an interactive web application using Streamlit to enable real-time predictions.
* Optimize model accuracy through feature selection and hyperparameter tuning.
* Ensure an easy-to-use interface for healthcare professionals and individuals.
* We will use this dataset to develop and evaluate machine learning models that can predict crop yields based on the available features.
* The main objective of this study is to explore the performance of different machine learning algorithms for crop production prediction and identify the most accurate model. We will also investigate the significance of different features in predicting crop yields and explore ways to improve the model’s accuracy.

This project aims to create a lightweight, accurate, and accessible diabetes prediction **system** that can:

* Help individuals assess their diabetes risk.
* Provide quick and accurate predictions.
* Serve as a foundation for future healthcare applications.

### Datasets and features:

 Source: CDC's BRFSS2015 Survey

 TotalRows: 253,680

 Total Features: 21 (excluding the target variable)

 Target Variable: Diabetes\_Binary

* 0 : No Diabetes
* 1: Diabetes

| **Feature** | **Description** |
| --- | --- |
| HighBP | 1 = High Blood Pressure, 0 = Normal BP |
| HighChol | 1 = High Cholesterol, 0 = Normal |
| CholCheck | 1 = Cholesterol check in past 5 years, 0 = No check |
| BMI | Body Mass Index (continuous) |
| Smoker | 1 = Smoker, 0 = Non-smoker |
| Stroke | 1 = History of stroke, 0 = No stroke |
| HeartDiseaseorAttack | 1 = Heart disease/attack history, 0 = No history |
| PhysActivity | 1 = Active, 0 = Inactive |
| Fruits | 1 = Eats fruits daily, 0 = Does not eat daily |
| Veggies | 1 = Eats vegetables daily, 0 = Does not eat daily |
| HvyAlcoholConsump | 1 = Heavy Alcohol Consumption, 0 = No heavy drinking |
| AnyHealthcare | 1 = Has healthcare access, 0 = No access |
| NoDocbcCost | 1 = Could not see a doctor due to cost, 0 = No issue |
| GenHlth | General health rating (1 = Excellent to 5 = Poor) |
| MentHlth | Mental health (number of bad mental health days in the last month) |
| PhysHlth | Physical health (number of bad physical health days in the last month) |
| DiffWalk | 1 = Difficulty walking, 0 = No difficulty |
| Sex | 1 = Male, 0 = Female |
| Age | Age category (13 levels: 1 = 18-24, 13 = 80+) |
| Education | Education level (1 = No schooling, 6 = College graduate) |
| Income | In+A61:B82category (1 = Low income, 8 = High income) |

## SYSTEM REQUIREMENTS

#### Hardware Requirements:

* Processor: Intel i5 or above
* RAM: 8GB or higher
* Storage: 50GB free space
* GPU (Optional for deep learning models)

#### Software Requirements:

* Operating System: Windows/Linux/MacOS
* Programming Language: Python 3.8+
* Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, XGBoost, PySpark
* IDE: Jupyter Notebook, VS Code

## FUNCTIONAL REQUIREMENTS

## User Functionalities

|  |  |  |
| --- | --- | --- |
| ID | Requirement | Description |
| 1 | User Input Form | Users should be able to enter their health-related details via a web form. |
| 2 | Diabetes Prediction | The system should process the input and predict whether the user has diabetes or not (0 = No Diabetes, 1 = Diabetes). |
| 3 | Real-time Processing | The prediction should be displayed instantly after submission. |
| 4 | Result Explanation | The model should provide basic reasoning behind the prediction (e.g., feature importance). |

## Data Processing Functionalities

|  |  |  |
| --- | --- | --- |
| ID | Requirement | Description |
| 1 | Data Preprocessing | The dataset should undergo data cleaning, normalization, and encoding before being fed into the model. |
| 2 | Feature Selection | The system should use feature importance techniques to select the most relevant variables. |
| 3 | Handling Missing Values | Missing values should be handled using mean/median imputation or dropping irrelevant records. |
| 4 | Data Standardization | Certain features like BMI should be normalized for better model performance. |

## Machine Learning Model Functionalities

|  |  |  |
| --- | --- | --- |
| ID | Requirement | Description |
| 1 | Model Selection | The system should use XGBoost as the primary model for diabetes classification. |
| 2 | Model Training | The model should be trained on the preprocessed dataset to maximize accuracy. |
| 3 | Hyperparameter Tuning | The model should optimize parameters such as learning\_rate, max\_depth, and n\_estimators. |
| 4 | Evaluation Metrics | The model should be evaluated using accuracy, precision, recall, |
|  |  |  |

## Web Application Functionalities

|  |  |  |
| --- | --- | --- |
| ID | Requirement | Description |
| 1 | Frontend with Streamlit | The system should have a simple Streamlit-based UI for easy interaction. |
| 2 | Graphical Data Representation | The system should display relevant charts and graphs for better visualization. |
| 3 | Deployment & Accessibility | The web app should be deployable via Streamlit Cloud or AWS/GCP. |

## System Functionalities

|  |  |  |
| --- | --- | --- |
| ID | Requirement | Description |
| 1 | Backend Processing | The system should handle user input, process the data, and make predictions. |
| 2 | Security & Data Privacy | User data should not be stored or shared, ensuring privacy. |
| 3 | Performance Optimization | The system should be optimized for fast predictions (<1 second). |

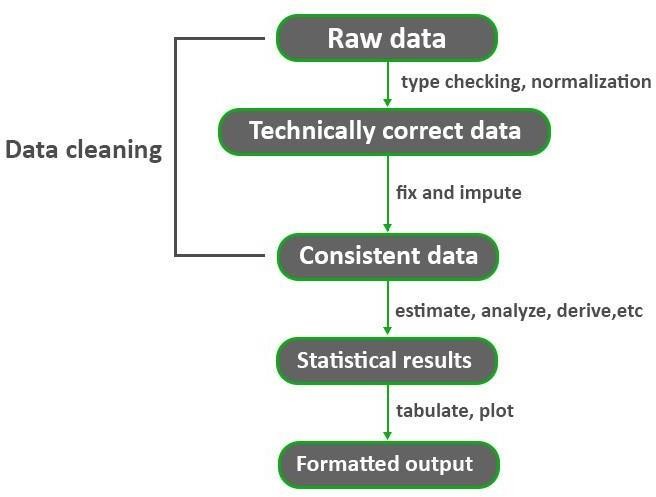
#### Python 3:

* + Python is a high-level programming language that is easy to learn and use.
  + Python is an interpreted language, which means that code can be executed on the fly, without the need for compilation.
  + Python is open source and free to use, with a large and active community of developers contributing to its development and maintenance.
  + Python has a vast collection of third-party libraries and packages, such as NumPy, Pandas, Matplotlib, and Scikit-learn, among others, that make it easy to perform data analysis.

#### Tableau:

* + Tableau is a data visualization and business intelligence software that allows users to connect, analyse, and share data in a visual and interactive way.
  + It offers a user-friendly drag-and-drop interface that enables users to create interactive dashboards, reports, and charts without the need for complex coding or programming.
  + Tableau supports various data sources, including spreadsheets, databases, cloud services, and bigdata platforms, such as Hadoop and Spark.

#### Data Cleaning:

****

**Fig: Data Cleaning Process**

* + Data cleaning is a crucial process in Data Mining. It carries an important part in the building of a model. Data Cleaning can be regarded as the process needed, but everyone often neglects it. Data quality is the main issue in quality information management. Data quality problems occur anywherein information systems. These problems are solved by data cleaning.
  + Without proper data cleaning, data analysis and modelling can lead to erroneous or biased results,which can have serious consequences for businesses and organizations.
  + Hence, it is a critical step in the data preparation process, as it can significantly impact the accuracyand reliability of the insights and decisions that are derived from the data. By improving the quality of data, organizations can gain a better understanding of their operations, customers, and market trends, and make more informed and effective decisions.

## SYSTEM ARCHITECTURE

Performing Ananlysis

Finding the best accuracy

Model Training

Machine Learning Using Python

Data Pre- processing

Raw data

Visualisation

**Fig: System Architecture of Diabetes Prediction Using Machine Learning**

## METHODOLOGY

Content:

1.Dataset Description:

* CDC BRFSS2015 dataset (253,680 records, 21 features).

2.Data preprocessing steps :

* Converted to Binary Classification (0 = No Diabetes, 1 = Diabetes)
* Feature Selection & Cleaning

3.Model Training:

* Algorithms Used: Random Forest, XGBoost, CatBoost, KNN

4.Performance Evaluation

5.Deployment:

* Developed Streamlit Web App.

## MACHINE LEARNING ALGORITHMS

Several machine learning algorithms were applied to predict diabetes based on selected features. Each model was trained and evaluated for accuracy, precision, recall, and F1-score. Below is a brief description of the algorithms used:

**K-Nearest Neighbors (KNN)**

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm that classifies data points based on their proximity to neighboring points in the feature space. The number of neighbors (k) is a key hyperparameter that influences model performance. In this study, the KNN model was tuned to select the optimal k-value for improved classification accuracy.

**Extreme Gradient Boosting (XGBoost)**

XGBoost is an optimized gradient boosting algorithm designed for speed and performance. It builds sequential decision trees, where each new tree corrects the errors of the previous ones. XGBoost employs regularization techniques to prevent overfitting and supports parallel computing, making it highly efficient for large datasets.

**Categorical Boosting (CatBoost)**

CatBoost is a gradient boosting algorithm that excels in handling categorical variables without requiring extensive preprocessing. It efficiently encodes categorical features and prevents overfitting using ordered boosting. This model is particularly useful for structured datasets and provides a balance between accuracy and interpretability.

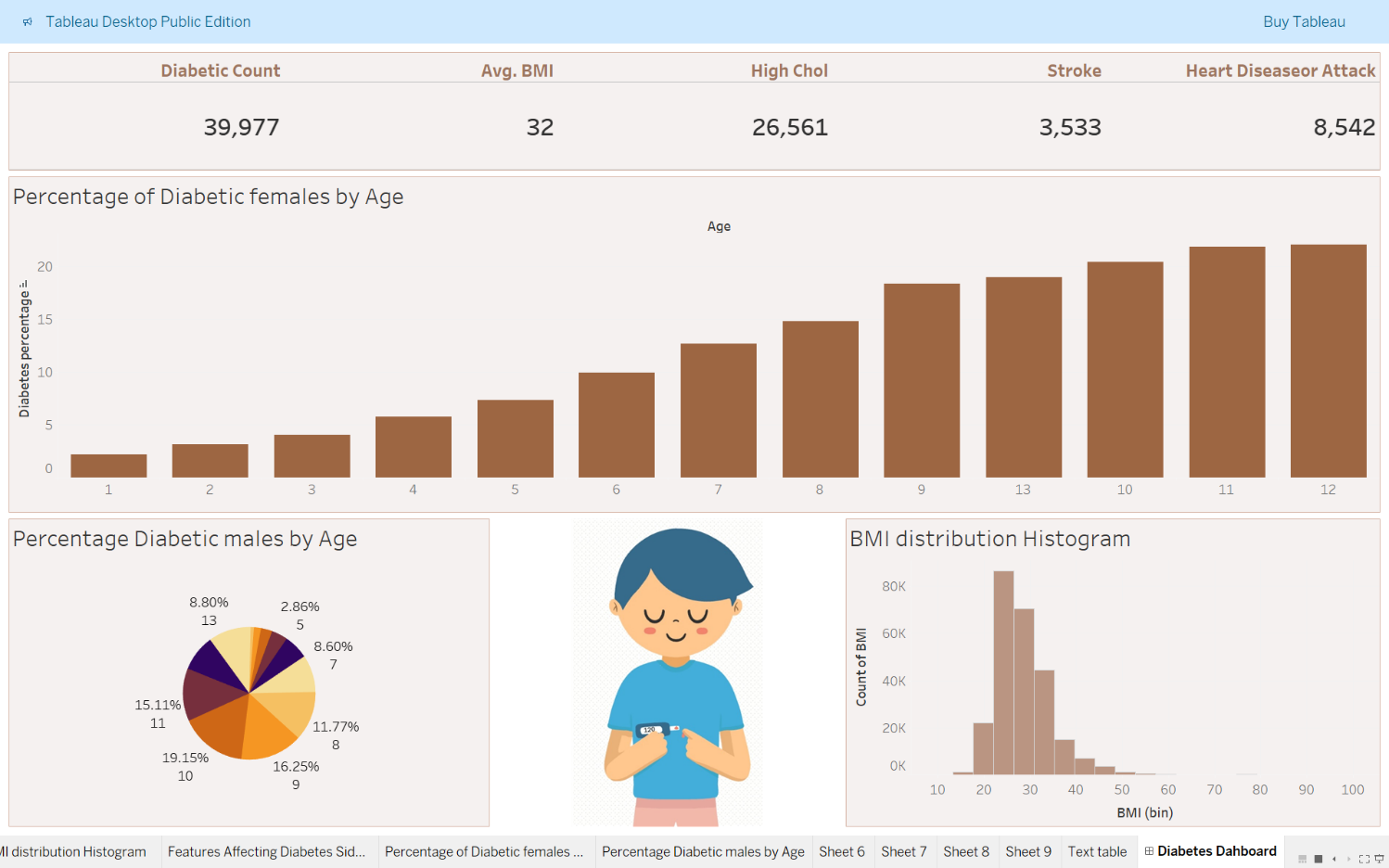
**Random Forest**

Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the majority vote for classification. It improves model stability and reduces variance by aggregating multiple predictions. Random Forest is highly effective for handling complex datasets with high-dimensional features and interactions.

**Model Performance**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision (Class 0)** | **Recall (Class 0)** | **Precision (Class 1)** | **Recall (Class 1)** | **F1-Score (Avg)** |
| **KNN (Tuned)** | **87.78%** | **0.91** | **0.83** | **0.85** | **0.92** | **0.88** |
| **XGBoost** | **88.71%** | **0.84** | **0.95** | **0.94** | **0.82** | **0.89** |
| **CatBoost** | **89.05%** | **0.85** | **0.95** | **0.94** | **0.83** | **0.89** |

## DATA VISUALIZATION AND REPRESENTATION



## CONCLUSION AND FUTURE SCOPE

**4.1 Conclusion**

In this study, multiple machine learning models were applied to predict diabetes based on selected features.

The performance of K-Nearest Neighbors (KNN), XGBoost, CatBoost, and Random Forest was

evaluated using metrics such as accuracy, precision, recall, and F1-score. Among all models**,** Random

Forest achieved the highest accuracy (90.04%), demonstrating its effectiveness in handling complex

patterns within the dataset.

* Random Forest emerged as the best-performing model due to its high accuracy, balanced
* precision-recall tradeoff, and ability to generalize well across both diabetic and non-diabetic
* classes.
* **X**GBoost and CatBoost performed competitively, with CatBoost slightly outperforming
* XGBoost in terms of recall.
* Tuned KNN achieved a strong recall for diabetic cases, making it a suitable choice when
* minimizing false negatives is a priority.

These results indicate that ensemble learning methods (Random Forest, XGBoost, CatBoost)

outperform distance-based methods (KNN) in diabetes prediction. However, each model presents trade-

offs in precision, recall, and computational efficiency, which must be considered based on specific use cases.

**4.2 Future Scope**

* Enhancing the model with deep learning techniques for improved accuracy.Integrating real-time patie-

nt data from wearable devices for continuous monitoring.Expanding the system to provide personalized health recommendations based on dietary and lifestyle factors.Deploying the system as a mobile-friendly application to increase accessibility.

## REFERENCES

Dataset : https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators