PHASE-02 : INNOVATION

AI-BASED DIABETES PREDICTION SYSTEM

INTRODUCTION:

We are assigned with the task of designing an AI based Diabetes Prediction System. In this project, our objective is to predict whether the patient has diabetes or not based on various features like Glucose level, Insulin, Age, BMI.

PROCESS INVOLVED IN OUR INNOVATION:

In this document, we discuss about the process involved in the innovation for our problem statement in detail.

Creating an AI Based Diabetes Prediction System involves several steps:

1. **DATA COLLECTION:**

We need a dataset containing medical features such as glucose levels, blood pressure, BMI, etc., along with information about whether the individual has diabetes or not.

1. **DATA PREPROCESSING:**

The medical data needs to be cleaned, normalized, and prepared for training machine learning models.

1. **FEATURE SELECTION:**

We will select relevant features that can impact diabetes risk prediction.

1. **MODEL SELECTION:**

We can experiment with various machine learning algorithms like Logistic Regression, Random Forest, and Gradient Boosting.

LOGESTIC REGRESSION:

Logistic regression is a statistical method used in machine learning for binary classification tasks, where the goal is to predict one of two possible outcomes or classes, typically labeled as 0 and 1. It's a type of generalized linear model (GLM) that models the relationship between the independent variables (features) and the probability of a particular outcome.

Here's an overview of how logistic regression works:

Sigmoid Function: Logistic regression uses the logistic (or sigmoid) function to model the probability that a given input belongs to the positive class (class 1). The sigmoid function is defined as:

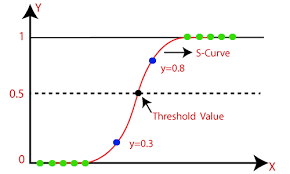
σ(z)=1+e−z1​

Here, $z$ is a linear combination of the input features and their associated coefficients, and $\sigma(z)$ outputs a value between 0 and 1, which can be interpreted as a probability.

Model Training: The goal of logistic regression is to find the optimal coefficients that maximize the likelihood of the observed data. This is typically done through a process called "maximum likelihood estimation" or "gradient descent." The coefficients are adjusted iteratively to minimize a cost function, which measures the difference between the predicted probabilities and the actual class labels.

Decision Boundary: Once the model is trained, you can use it to make predictions. A common threshold for classification is 0.5. If $\sigma(z) > 0.5$, the input is classified as class 1; otherwise, it's classified as class 0. The decision boundary is the line or hyperplane that separates the two classes.

Evaluation: Logistic regression models can be evaluated using various metrics, such as accuracy, precision, recall, F1 score, and the ROC curve. These metrics help you assess how well the model is performing on your dataset.



**5. MODEL EVALUATION:**

We will evaluate the model’s performance using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

**6. VALIDATION:**

Validating the model using cross-validation techniques or external datasets to ensure its robustness and reliability.

**7. MODEL DEPLOYMENT:**

Deploying the trained model as a software application that can be used by healthcare professionals or patients to predict the risk of diabetes and provide personalized recommendations for prevention or management.

**8. ITERATIVE IMPROVEMENT:**

We will fine-tune the model parameters and explore techniques like feature engineering to enhance prediction accuracy.

**9. MODEL TESTING:**

Testing the deployed model with real-world users and collecting their feedback on its usability, effectiveness, and satisfaction.

CONCLUSION:

AI aims to make accurate and advanced predictions for a large amount of knowledge data. With patient records, we are able to accurately predict whether or not the patients in the dataset have diabetes.