

PROJECT TITLE:

**Diabetes Prediction Using Machine Learning with an
End-to-End MLOps Pipeline**

Team 09

PROBLEM STATEMENT:

Diabetes is a rapidly increasing global health issue where early detection is crucial to prevent severe complications. This project aims to develop a machine learning model that predicts diabetes using key health parameters such as blood glucose levels, BMI, age, and insulin. An end-to-end MLOps pipeline will be implemented to automate model training, evaluation, and deployment, enabling scalable and reliable real-time predictions.

Key Challenges

- Data Quality & Missing Values: Medical datasets often contain missing, inconsistent, or zero values (e.g., insulin, BMI), which can mislead the model and reduce prediction accuracy.
- Class Imbalance: Diabetes datasets typically have fewer positive (diabetic) cases than negative ones, causing biased predictions and poor detection of high-risk patients.
- Model Drift & Deployment Challenges: Changes in patient health patterns over time can degrade model performance, requiring continuous monitoring, retraining, and robust MLOps automation.

Abstract:

Diabetes mellitus represents one of the most pressing chronic health challenges worldwide, affecting over 500 million individuals and projected to reach 700 million by 2045, with severe complications including cardiovascular disease, kidney failure, neuropathy, and vision loss that significantly diminish quality of life and escalate healthcare costs. Early and accurate diagnosis remains critically important, yet traditional screening methods often prove inadequate for population-scale deployment due to resource limitations and inconsistent access to clinical diagnostics.

This project introduces **Mellitus - MLOps**, an advanced machine learning-based predictive system designed to accurately forecast diabetes risk using a comprehensive set of behavioral, clinical, and socioeconomic health parameters derived from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) dataset—comprising over 70,000 patient records with 21 key features including BMI categories, physical activity levels, dietary habits, hypertension status, cholesterol levels, smoking history, healthcare access, and demographic factors. Unlike conventional approaches limited to basic clinical metrics, this holistic feature set enables more nuanced risk stratification.

The champion model deploys as a **FastAPI** web service in **Docker** containers, enabling real-time predictions for healthcare apps.

This demonstrates production-grade ML practices—MLflow experiment tracking — ensuring scalability, reproducibility, and continuous improvement for real-world diabetes screening.