**📝 Summary – Diabetes Prediction System**

📌 **Dataset:**

* 768 patient records with 9 health-related features.

🔧 **Data Preprocessing:**

* Fixed incorrect/shifted column names.
* Replaced invalid medical values (0 for Insulin, BMI, etc.) with median.
* Removed extreme outliers using **Z-Score**.
* No missing or duplicate records found.

📊 **Exploratory Data Analysis (EDA):**

* Key features linked to diabetes: **Pregnancies, Insulin, BMI, and Family History (DPF)**.
* No severe multicollinearity (correlation < 0.7).
* Insulin values had many outliers.

**🤖 Modeling:**

* **Without Hyperparameter Tuning:**👉 Logistic Regression → 75.58%  
  👉 Decision Tree → 67.46%  
  👉 Random Forest → 77.91%  
  👉 Gradient Boosting → 76.46%  
  👉 SVM → 75.29%  
  👉 Voting Classifier (CV) → 75.89%  
  👉 Stacking → 71.74%  
  👉 XGBoost → 71.01%  
  👉 SMOTE + SVM → 56.52%
* **With Hyperparameter Tuning:**
* **Model**: SMOTE (Synthetic Minority Over-sampling Technique) + XGBoost
* **Validation Strategy**: Stratified K-Fold Cross-Validation with **GridSearchCV** for hyperparameter tuning
* **Test Accuracy**: **73.91%**

**📌 Conclusion**

* **Logistic Regression, SVC, and KNN** are the most reliable models with **~80% accuracy**.
* **Ensemble methods (Voting & Stacking)** give stable performance but do not significantly improve accuracy.
* **Complex models (Random Forest, Decision Tree, XGBoost)** tend to overfit unless tuned properly.
* To improve beyond 80%, further **feature engineering, advanced scaling, encoding, and better handling of outliers** is required.

⚖️ **Final Insight:**

* This project highlights that **simpler, well-preprocessed models can perform just as effectively as complex models** for diabetes prediction, making them practical and reliable for real-world healthcare applications.