**📅 Project Summary – Cervical Cancer Risk Prediction (Day 4)**

**🎯 Goal:** Predict cervical cancer risk using behavioral and medical data with a focus on explainability and fair evaluation.

**📊 Dataset:**

* Kaggle: Cervical Cancer Behavior Risk
* 858 rows, 36 features
* Target: Biopsy (0 = No risk, 1 = At risk)

**⚖️ Machine Learning Type:**

* Supervised Learning – Binary Classification

**🔄 Techniques Used:**

* Exploratory Data Analysis (distributions, imbalance, correlation)
* Feature Selection (removed leakage & redundancy)
* Data Preprocessing (standardization)
* Stratified K-Fold Cross-Validation (5 folds)
* Logistic Regression Model
* SHAP for Explainability
* Saved model, scaler, and schema (.pkl/.json)

**📊 Results (StratifiedKFold Evaluation):**

| **Metric** | **Value** | **Description** |
| --- | --- | --- |
| Accuracy | 0.958 | High overall correctness |
| Precision | 0.643 | Proportion of correct positive predictions |
| Recall | 0.818 | Captures high-risk patients (critical) |
| F1 Score | 0.720 | Balanced metric for imbalanced data |

**🔍 SHAP Explainability:**

* Visualized global feature impact on predictions
* Key predictors: STDs, Dx:CIN, HPV, Age, Sexual Partners
* SHAP plot helped explain decisions to non-technical stakeholders

**📆 Deliverables:**

* cervical\_cancer\_model.pkl
* scaler.pkl
* feature\_columns.json

**📖 Key Takeaways:**

* Stratified K-Fold is essential for imbalanced classification
* SHAP ensures transparency in medical ML
* High Recall is more important than Accuracy in healthcare
* Project is deployment-ready for real-world use

**👩‍💻 Author:** Dilrabo Khidirova  
**🌐 Role:** Data Scientist | ML Engineer | FAANG Prep Journey