Project 5: Early Diagnosis of Chronic Kidney Disease (CKD)

Company: MediScan Diagnostics (A Network of Diagnostic Laboratories)

Project Background & Business Problem

MediScan Diagnostics is looking to implement an AI-driven screening tool to assist doctors in the early detection of Chronic Kidney Disease (CKD). Currently, many patients are only diagnosed in advanced stages because the symptoms are subtle. This leads to:

- **High Treatment Costs:** Late-stage CKD requires dialysis or transplants, which are significantly more expensive than early-stage management.
- **Diagnostic Bottlenecks:** Specialists are often overwhelmed with routine screenings, slowing down the time to diagnosis for critical patients.
- **Inconsistent Screening:** General practitioners sometimes overlook minor variations in lab results that, when combined, strongly indicate early-stage CKD.

Objective: You are to build a **Classification Risk Engine**. The model must analyze biochemical markers to predict whether a patient is at risk of CKD. Because medical errors are costly, the company prioritizes **Recall** (not missing any positive cases) and **Explainability** (doctors must understand why a patient was flagged).

Dataset Specifications

The dataset, mediscan_ckd_diagnostic.csv, contains **50,000 rows** of patient health records.

Feature Name	**Description**
Patient_ID	Unique identifier for each patient.
Age	Age of the patient.
Blood_Pressure	Diastolic blood pressure (mm/Hg).
Specific_Gravity	Urine density (1.005 to 1.025).
Albumin	Level of albumin in urine (0 to 5).
Sugar	Level of sugar in urine (0 to 5).
Red_Blood_Cells	Presence of RBCs in urine (Normal/Abnormal).

Feature Name	**Description**
Blood_Glucose_Random	Random blood glucose level (mgs/dl).
Blood_Urea	Level of urea in the blood (mgs/dl).
Serum_Creatinine	Level of creatinine in the blood (mgs/dl).
Sodium	Level of sodium in the blood (mEq/L).
Hemoglobin	Hemoglobin levels (gms).
White_Blood_Cell_Count	WBC count (cells/cmm).
Hypertension	Binary (Yes/No).
Diabetes_Mellitus	Binary (Yes/No).
CKD_Status	**Target Variable:** (1 = Patient has CKD, 0 = Healthy).

Project Deliverables

1. **Handling Missing Data:** Clinical data often has missing lab values. Apply **KNN Imputation** as per your syllabus to fill these gaps realistically.
2. **Model Selection & Tuning:** Implement a **Support Vector Machine (SVM)** and a **Random Forest**. Use **GridSearchCV** to find the optimal hyperparameters.
3. **Explainable AI (XAI):** Medical professionals require transparency. Use **SHAP** to visualize which features (e.g., Creatinine vs. Hemoglobin) are the strongest predictors for specific patient cases.
4. **Deployment:** Develop a **Streamlit Dashboard** where a lab technician can input a patient's lab results to get a risk probability and a feature importance plot.

Final Evaluation & Presentation Guidelines

Since these students have successfully completed their training on machine learning for business intelligence, the final evaluation should focus on their ability to translate technical metrics into business value. Each group of five will be assessed on their end-to-end implementation—from data quality assessment to model explainability.

Presentation Structure (15 Minutes)

1. **Business Problem & Gap Analysis (2 Mins):** Define the company's pain points and what the "to-be" state looks like after ML implementation.
2. **Data Preprocessing & EDA (3 Mins):** Demonstrate how they handled missing values, outliers, and feature transformation.
3. **Modeling & Tuning (4 Mins):** Explain the choice of algorithms and the results of hyperparameter tuning using GridSearchCV or RandomizedSearchCV.
4. **Explainability & Diagnostics (3 Mins):** Use SHAP or LIME to explain a single prediction and discuss the bias-variance tradeoff.
5. **Live Streamlit Demo (3 Mins):** A functional demonstration of the deployed model pipeline.

Evaluation Rubric

Criteria	**Exceptional (5)**	**Proficient (3-4)**	**Developing (1-2)**
Data Engineering	Advanced feature engineering; pipeline is automated.	Standard cleaning and scaling applied.	Minimal cleaning; data leaks present.
Model Rigor	Multiple models compared with cross-validation.	Single model with basic evaluation metrics.	Overfitted model with no validation.
Explainability	Clear XAI insights linked to business decisions.	SHAP/LIME plots included but not explained.	No interpretability analysis.
Deployment	Intuitive Streamlit UI with real-time error handling.	Basic functional Streamlit app.	Script runs only in a notebook.