# Experiment Number 6

Design any case study on the following topic and present in front of external

Case study Topic: Diagnose disease risk from Patient data.

Sample Case study:

1. Predicting Cardiovascular Disease Risk: A Case Study Using Patient Health Data"
2. "Leveraging Electronic Health Records to Forecast Disease Risk: A Predictive Model Case Study"
3. "Early Detection of Disease Risk Through Patient Data Analysis: A Case Study on Chronic Disease Prediction"
4. "Identifying High-Risk Patients: A Case Study on Predictive Modeling for Cardiovascular Disease"
5. "From Data to Diagnosis: Using Machine Learning to Predict Disease Risk in Patient Populations"
6. "Improving Preventative Care with Patient Data Analytics: A Case Study in Disease Risk Prediction"
7. "Harnessing Patient Data for Early Disease Detection: A Case Study on Risk Stratification"

You can prepare any one out of the above topic

Points to be included

* **Data Collection and Preparation**
* **Data Source**
* **Patient Variables**
* **Data Cleaning**
* **Exploratory Data Analysis (EDA)**
* **Feature Engineering and Selection**
* **Feature Selection**
* **Train-Test Split**
* **Hyperparameter Tuning**
* **Model Evaluation**
* **Evaluation Metrics**
* **Model Interpretation and Validation**
* **Interpretability Techniques**
* **Validation on External Data**
* **Risk Stratification and Clinical Interpretation**
* **Risk Categories**

1. **Ethical and Privacy Considerations**
2. **Expected Outcomes**

**Title:** "Predicting Cardiovascular Disease Risk: A Case Study Using Patient Health Data"

**1. Introduction**

Cardiovascular disease (CVD) is a major health concern worldwide, contributing to millions of deaths every year. Early prediction and prevention can save lives and improve healthcare quality. This case study explores the use of patient health data to predict cardiovascular disease risk using machine learning models. The study leverages patient electronic health records (EHR) and other healthcare data to classify individuals into various risk categories, enabling healthcare providers to intervene early.

**2. Data Collection and Preparation**

**Data Source**

For this study, data was sourced from a large dataset of patient EHRs provided by a healthcare organization, incorporating variables relevant to cardiovascular health risk factors. The data includes records of over 50,000 patients across different age groups, genders, and ethnicities.

**Patient Variables**

The patient health data consisted of multiple variables, categorized as follows:

* **Demographic Information:** Age, gender, ethnicity
* **Clinical Measurements:** Blood pressure, cholesterol levels, BMI, heart rate
* **Lifestyle Factors:** Smoking status, physical activity, diet
* **Medical History:** Previous cardiovascular events, family history of heart disease, comorbidities such as diabetes and hypertension

**3. Data Cleaning**

* **Handling Missing Values:** Imputed missing data using the median for continuous variables and mode for categorical variables.
* **Outlier Detection:** Analyzed outliers in critical health indicators (e.g., blood pressure, BMI) and decided to retain or adjust them based on clinical plausibility.
* **Standardization:** Standardized continuous variables like cholesterol and blood pressure for consistency across different patient populations.

**4. Exploratory Data Analysis (EDA)**

During EDA, several insights were uncovered:

* A higher risk of CVD was observed among patients with high cholesterol, obesity, and a history of smoking.
* Strong correlations were identified between age, high blood pressure, diabetes, and CVD risk.
* Age and gender had noticeable impacts on CVD risk distribution.

**5. Feature Engineering and Selection**

**Feature Engineering:**

* Created composite indicators such as a “Lifestyle Risk Score” based on diet, exercise, and smoking habits.
* Developed an “Age-Adjusted Risk” factor for individuals with a family history of cardiovascular disease.

**Feature Selection:**

* Applied statistical methods (e.g., correlation analysis and recursive feature elimination) to narrow down the most predictive features.
* Chose top features like age, blood pressure, cholesterol levels, smoking status, and diabetes history for modeling.

**6. Model Development**

**Train-Test Split:**

* Split the dataset into training (80%) and testing (20%) subsets.

**Hyperparameter Tuning:**

* Conducted grid search and random search for model hyperparameters to enhance performance.

**Model Evaluation:**

* Utilized various machine learning models, including Logistic Regression, Random Forest, and Gradient Boosting.
* Gradient Boosting was selected as the final model due to its superior accuracy and interpretability.

**Evaluation Metrics:**

* Evaluated models using metrics such as Accuracy, Precision, Recall, F1 Score, and AUC-ROC to gauge model performance.

**7. Model Interpretation and Validation**

**Interpretability Techniques:**

* Utilized SHAP (SHapley Additive exPlanations) values to determine feature importance and understand model predictions.
* Observed that high blood pressure, high cholesterol, and a history of smoking were significant predictors of cardiovascular risk.

**Validation on External Data:**

* Validated the model on an external dataset from a different healthcare provider to test model robustness and generalizability. Performance remained consistent with the initial test data.

**8. Risk Stratification and Clinical Interpretation**

**Risk Categories:** The model stratified patients into three risk categories:

* **Low Risk:** Patients with minimal risk factors and a low probability of developing CVD.
* **Moderate Risk:** Patients with a few risk factors, including slightly elevated blood pressure or cholesterol.
* **High Risk:** Patients with multiple risk factors, such as obesity, smoking, high cholesterol, and hypertension.

**9. Ethical and Privacy Considerations**

* **Data Privacy:** Ensured all patient data was de-identified to protect privacy.
* **Bias Mitigation:** Performed bias testing to verify fairness across demographics, including race, gender, and age, to avoid discriminatory outcomes.

**10. Expected Outcomes**

This predictive model is expected to enable healthcare providers to:

* Identify high-risk patients early, offering targeted preventive care.
* Allocate resources efficiently by focusing on patients with elevated risk profiles.
* Track and manage patient health better, ultimately aiming to reduce CVD incidence and improve population health.

**Conclusion**

The successful development of this CVD risk prediction model demonstrates the potential for machine learning in healthcare. By leveraging patient health data, healthcare providers can make informed, data-driven decisions to enhance preventive care, optimize treatment plans, and ultimately reduce cardiovascular disease prevalence.