Mahavir Education Trust's
SHAH & ANCHOR KUTCHHI ENGINEERING COLLEGE
Chembur, Mumbai - 400 088
UG Program in Information Technology

- **Student Name: Purva Mahamunkar**
- **Roll Number: 12/28**
- **PRN : 22UF17661IT090**
- **Department: IT**
- **Batch/Division: BTECH-1**
- **Experiment Title: Predict Disease Risk from Patient Data**
- **Date: 25/08/2025**

| Experiment No. 6 | | | | |
|---|---|---|---|---|
| **Date of Performance:** | **25/08/2025** | | | |
| **Date of Submission:** | **25/08/2025** | | | |
| Program Execution/ formation / correction/ ethical practices | Timely Submission | Viva | Experiment Total | Sign with Date |
| | | | | |

# Experiment No 6

**6.1 Aim:** Predict Disease Risk from Patient Data

**6.2 Course Outcome:** Train and evaluate Machine Learning models to predict disease risks using structured patient data.

**6.3 Learning Outcome:** Train and evaluate Machine Learning models to predict disease risks using structured patient data.

**6.4 Requirement:** Python (with TensorFlow/PyTorch), Jupyter/Colab, OpenCV, NumPy, Matplotlib, and a labeled Diseases Risk Patient dataset.

## 6.5 Related Theory:

### 1. Introduction

- Disease risk prediction is a key application of **AI/ML in healthcare**.
- Structured patient data includes demographic, clinical, and lifestyle features.
- Goal: Use ML models to identify patients at high risk of developing a disease early.

### 2. Data Collection

- Gather structured data from electronic health records, hospital databases, or health surveys.
- Features may include:
  - Age, Gender, BMI.
  - Medical history (blood pressure, cholesterol, sugar levels).
  - Lifestyle factors (smoking, diet, exercise).
  - Lab test results.

- Target variable: Presence or risk of disease (Yes/No).

**3. Data Preprocessing**

- **Data Cleaning:** Handle missing values, outliers, and inconsistencies.
- **Feature Encoding:** Convert categorical data (e.g., gender: Male/Female → 0/1).
- **Feature Scaling:** Normalize/standardize numerical values for uniformity.
- **Train-Test Split:** Divide dataset (e.g., 80% training, 20% testing).

**4. Model Selection**

- Choose suitable ML algorithms for classification tasks:
    - **Logistic Regression** → interpretable, baseline model.
    - **Decision Trees / Random Forests** → handle complex relationships, non-linear data.
    - **Support Vector Machines (SVM)** → good for high-dimensional data.
    - **Gradient Boosting (XGBoost/LightGBM)** → strong predictive performance.
    - **Neural Networks** (if the dataset is large and complex).

**5. Model Training**

- Feed training data (features + labels) into chosen ML model.
- The model learns patterns between patient attributes and disease risk.
- Hyperparameter tuning (e.g., using Grid Search, Cross-Validation) improves accuracy.

**6. Model Evaluation**

- Test the model on unseen **test dataset**.
- Metrics used:
    - **Accuracy** – overall correct predictions.
    - **Precision** – correctly predicted positives out of all predicted positives.
    - **Recall** – correctly predicted positives out of actual positives
    - **F1-Score** – balance between precision & recall.
    - **ROC-AUC** – overall ability of model to distinguish between classes.

**7. Model Deployment**

- Once validated, the model can be deployed in a clinical decision support system.
- Doctors can input patient data → the model predicts the probability of disease risk.

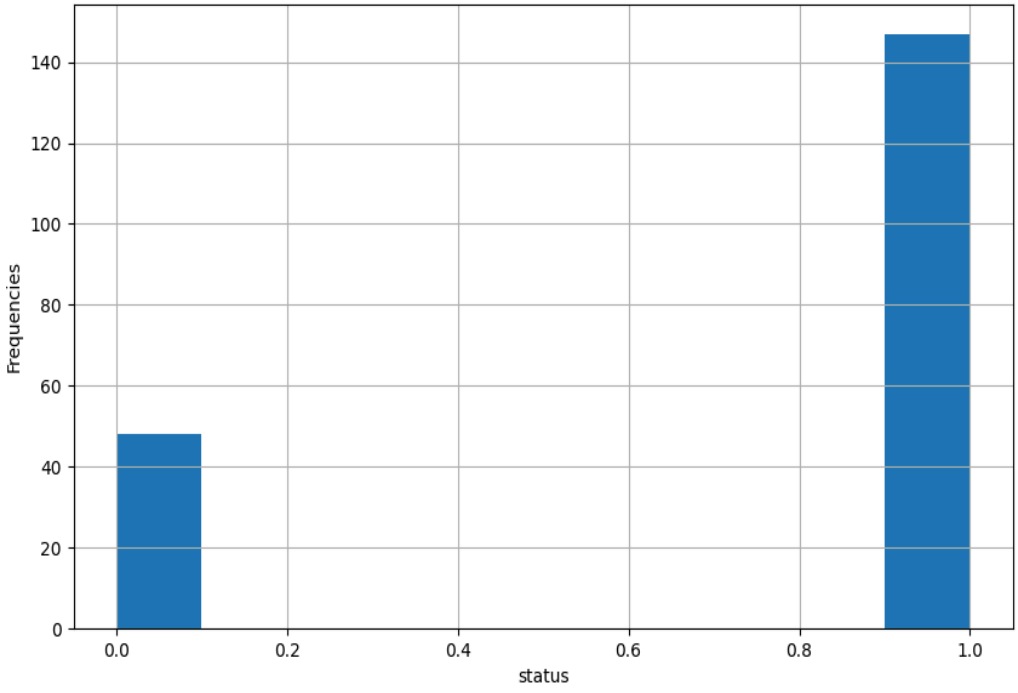
**8. Ethical Considerations**

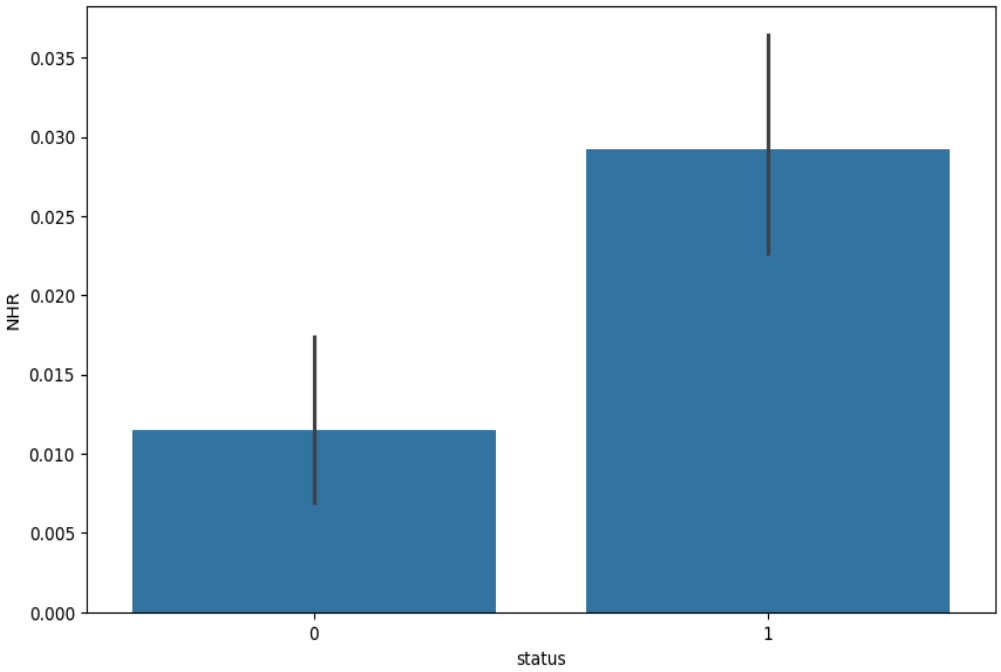
- Ensure patient data privacy and compliance with medical standards (HIPAA, GDPR).
- Avoid algorithmic bias (train with diverse datasets).
- Use explainable ML models for doctor trust and clinical acceptance.
- ML models trained on structured patient data can significantly improve early detection and prevention of diseases.

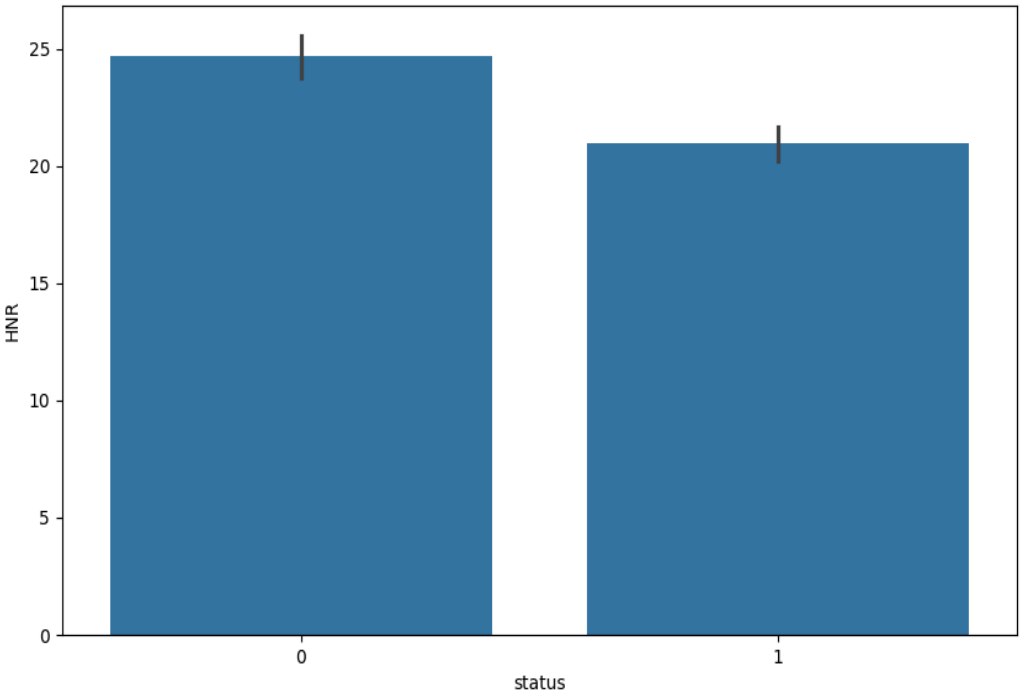
## 6.6 Program and Output:

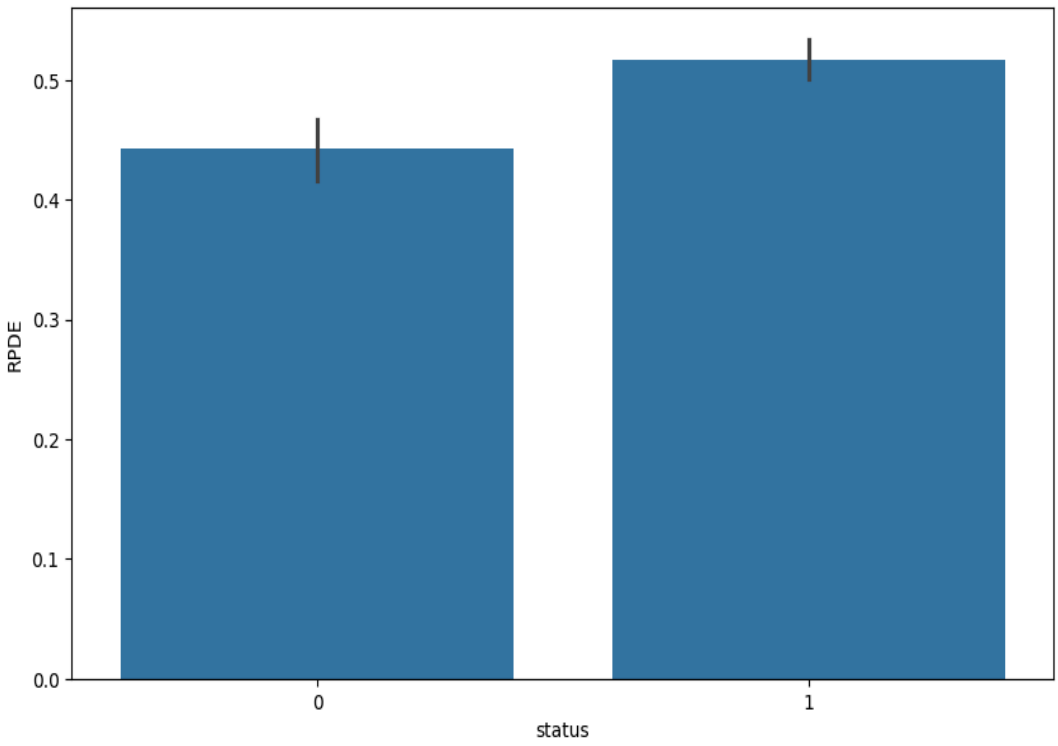
Completed Google Colab AIML6 Link

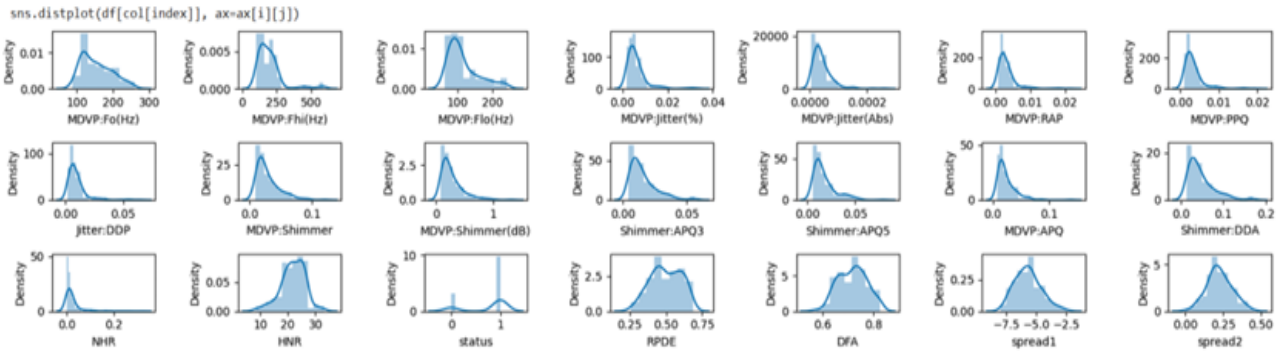
## 4.7 Result SnapShots:




25
20
15
10
5
0
HNR
0
1
status


0.5
0.4
0.3
0.2
0.1
0.0
RPDE
0
1
status




## 6.8 Conclusion:

Machine Learning models trained on structured patient data provide an efficient and reliable way to predict disease risks. By carefully preprocessing patient information, selecting suitable algorithms, and evaluating models with appropriate metrics, healthcare providers can identify at-risk individuals early. This not only supports timely treatment and prevention but also improves patient outcomes and reduces healthcare costs. However, successful deployment must ensure data privacy, fairness, and transparency to build trust in real-world medical use.

## 6.9 Questions:

1. **Why do we split the dataset into training and testing sets?**
   To ensure unbiased evaluation of the model. The training set is used to learn, and testing set is used to check generalization ability.

2. **What is the difference between Logistic Regression and Random Forest in disease prediction?**
   Logistic Regression is a linear classifier, simple and interpretable. Random Forest is an ensemble of decision trees, more powerful, handles non-linear data well, and usually provides higher accuracy.

3. **Why is feature scaling not necessary for Random Forest but needed for Logistic Regression?**
   Random Forest is tree-based and splits based on thresholds, unaffected by scaling. Logistic Regression depends on coefficients, so feature scaling improves its performance.

4. **How do you interpret a confusion matrix in this experiment?**
   It shows the number of correctly predicted healthy/disease patients (True Positive, True Negative) and misclassified cases (False Positive, False Negative).