演示内容应包括背景、现有解决方案的文献综述、提出的方法、评估结果以及结论和讨论。演示时间应少于10分钟，并且所有组员都应参与（准备幻灯片、演示或两者兼有）。将视频保存为.mp4格式，并将其提交到Moodle（文件大小应小于250MB）。

### ****Presentation: Predicting Breast Cancer Treatment Outcomes Using Machine Learning****

#### ****Speaker 1: Background (2 minutes)****

Opening lines:  
“Good morning/afternoon, everyone! Today, we will present our work on predicting breast cancer treatment outcomes using machine learning, focusing on pathological complete response (PCR) and relapse-free survival (RFS).”

**Key Points**:

**Prevalence of Breast Cancer**:

* 1. Breast cancer is the most common cancer among women in the UK. Chemotherapy is a standard treatment to shrink tumors before surgery.
  2. However, chemotherapy is toxic and only 25% of patients achieve PCR, which correlates with better outcomes and longer RFS.

**Research Motivation**:

* 1. Predicting PCR and RFS before treatment can help personalize patient care, minimizing unnecessary chemotherapy and improving survival rates.

**Objective**:

* 1. This study uses machine learning to predict PCR (classification) and RFS (regression), addressing challenges like data imbalance, missing values, and high dimensionality.

#### ****Speaker 2: Literature Review (2 minutes)****

Opening lines:  
“I will now review existing solutions and their limitations, highlighting the need for improved methods.”

**Key Points**:

**Machine Learning in Breast Cancer**:

* 1. **Support Vector Machines (SVMs)**: Consistently effective in cancer classification due to their ability to handle complex patterns in data.
  2. **Artificial Neural Networks (ANNs)**: Strong performers on complex datasets, especially in deep learning applications.
  3. **Hybrid Models**: Combining SVM with decision trees improves accuracy and adaptability.

**Challenges**:

* 1. **Data Heterogeneity**: Breast cancer datasets often have missing values, inconsistent features, and redundant data.
  2. **Imbalanced Data**: Underrepresentation of PCR cases skews model performance.
  3. **High Dimensionality**: MRI data contributes numerous features, complicating model training.

**Takeaway**:

* 1. While existing methods show promise, there’s a need for more robust, scalable solutions that integrate multi-modal data and address data-specific challenges.

#### ****Speaker 3: Proposed Method (2 minutes)****

Opening lines:  
“Next, I will explain the methods we proposed to overcome these challenges and achieve better predictions.”

**Key Points**:

**Data Preprocessing**:

* 1. Missing values were imputed using the median or encoded as a separate category for categorical data like “Gene.”
  2. Data normalization and scaling ensured standardized feature ranges, improving model performance.

**Feature Selection and Dimensionality Reduction**:

* 1. **Principal Component Analysis (PCA)**: Reduced data complexity while retaining critical variance.

**Handling Class Imbalance**:

* 1. **SMOTENC**: Synthetic Minority Over-sampling Technique for Nominal and Continuous features, balanced the dataset, enhancing model recall for the minority PCR class.

**Model Choices**:

* 1. **PCR Prediction**: Random Forest, AdaBoost, Logistic Regression.
  2. **RFS Prediction**: Linear Regression, Random Forest, Lasso Regression.

**Optimization**:

* 1. Hyperparameter tuning and cross-validation were applied to enhance model robustness.

#### ****Speaker 4: Evaluation Results (2 minutes)****

Opening lines:  
“Now, I will share the evaluation results of our models and their implications.”

**Key Points**:

**PCR Classification Results**:

* 1. **AdaBoost**: Achieved the best performance with 75.33% balanced accuracy and ROC-AUC of 0.7533. Its adaptive boosting mechanism effectively handled imbalanced data.
  2. **Random Forest**: Delivered robust performance but slightly lagged behind AdaBoost.
  3. **Logistic Regression**: Moderate performance, limited by linear assumptions but efficient and interpretable.

**RFS Regression Results**:

* 1. **Linear Regression**: Performed best with the lowest Mean Absolute Error (20.435) and highest R-squared (0.0785), indicating a predominantly linear feature space.
  2. **Random Forest**: Strong nonlinear modeling capabilities but less effective due to PCA-based information loss.
  3. **Lasso Regression**: Regularization did not improve performance significantly, likely due to optimized PCA-based features.

**Takeaway**:

* 1. Machine learning effectively addressed the challenges of imbalance and high dimensionality, providing reliable predictions for both tasks.

#### ****Speaker 5: Conclusion and Discussion (2 minutes)****

Opening lines:  
“Finally, I will summarize our findings and discuss potential future directions.”

**Key Points**:

**Conclusions**:

* 1. AdaBoost and Linear Regression emerged as the best-performing models for PCR and RFS predictions, respectively.
  2. Methods like SMOTENC and PCA significantly improved model performance.

**Study Limitations**:

* 1. Limited dataset size may affect model generalization.
  2. Single-modality data (e.g., clinical and MRI features) restricts model capacity to capture complex patterns.

**Future Directions**:

* 1. Integrating multi-modal datasets (e.g., genomic, imaging, and clinical data) for more robust predictions.
  2. Exploring advanced models like XGBoost and feature engineering techniques to capture nonlinear relationships.

**Closing Lines**:  
“This study demonstrates how machine learning can transform cancer treatment strategies by predicting outcomes and personalizing care. Thank you for your attention, and we welcome your questions.”