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 Script: Predicting Breast Cancer Treatment Outcomes Using Machine Learning****

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

Opening lines:  
“Hello, everyone! Today, our team will present our research on predicting breast cancer treatment outcomes, focusing on pathological complete response (PCR) and relapse-free survival (RFS) using machine learning.”

**Key Points**:

**Breast Cancer Overview**

* 1. Breast cancer is the most common cancer among women in the UK, and chemotherapy is a common pre-surgical treatment. However, it is not effective for all patients, with only about 25% achieving PCR.

**Research Significance**

* 1. Achieving PCR is strongly associated with better prognosis and longer relapse-free survival (RFS). Predicting treatment outcomes can optimize decision-making and reduce chemotherapy toxicity.

**Objective**

* 1. Our study aims to leverage machine learning to predict PCR and RFS using pre-treatment data, facilitating personalized treatment strategies.

#### ****Speaker 2: Literature Review of Existing Solutions (2 minutes)****

Opening lines:  
“Next, I will introduce the current research progress, existing solutions, and challenges in this field.”

**Key Points**:

**Machine Learning in Breast Cancer Prediction**

* 1. **Support Vector Machines (SVMs)**: Studies have demonstrated SVM’s high accuracy and reliability in breast cancer classification.
  2. **Artificial Neural Networks (ANNs)**: Deep learning techniques have shown great potential in handling complex datasets.
  3. **Hybrid Models**: Ensemble techniques (e.g., combining SVM with decision trees) improve diagnostic accuracy and adaptability.

**Data Challenges**

* 1. **Imbalanced Data**: PCR cases are often underrepresented, affecting model performance.
  2. **High-Dimensional Data**: The dataset includes 107 MRI-based features, which increases computational complexity.
  3. **Missing Data**: For example, the “Gene” feature has a 28.2% missing rate, requiring careful handling.

**Summary**:

* Effective data preprocessing and feature selection are crucial to improving model performance.

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

Opening lines:  
“Next, I will detail the methods we proposed, including data processing, feature selection, and model design.”

**Key Points**:

**Data Preprocessing**

* + **Handling Missing Values**:
    - Continuous data was filled with the median. For categorical features like “Gene,” missing values were treated as a separate category to preserve information.
  + **Normalization and Scaling**: Ensured balanced feature values for better model performance.

**Feature Selection and Dimensionality Reduction**

* + **Principal Component Analysis (PCA)**: Reduced data complexity while preserving key information.

**Handling Imbalanced Data**

* + **SMOTENC**: Balanced class distributions and improved recall for minority PCR cases.

**Model Selection**

* + **PCR (Classification)**: Random Forest, AdaBoost Classifier, and Logistic Regression were used.
  + **RFS (Regression)**: Random Forest and Linear Regression were implemented.

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

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

**Key Points**:

**Evaluation Metrics**

* + **Classification Task (PCR)**: Balanced accuracy, precision, recall, F1-score, and ROC-AUC were used.
  + **Regression Task (RFS)**: Mean Absolute Error (MAE) was used.

**Classification Results**

* + **AdaBoost Classifier**: Achieved the highest balanced accuracy (75.33%) and ROC-AUC (75.33%) for PCR prediction.
  + **Impact of SMOTE**: Significantly improved the model’s ability to recognize minority classes.

**Regression Results**

* + **Random Forest**: Outperformed linear regression, demonstrating better handling of high-dimensional data and more stable predictions.

**Summary**:

* Our models performed well in imbalanced data scenarios, providing reliable prediction results.

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

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

**Key Points**:

**Research Conclusion**

* + Our models effectively predicted PCR and RFS, demonstrating the potential of machine learning in personalized medicine.
  + Data preprocessing techniques (e.g., SMOTE, PCA) and ensemble models (e.g., AdaBoost) significantly enhanced performance.

**Research Limitations**

* + The dataset size was limited, requiring validation on larger, independent datasets for generalization.
  + Single-modality data (e.g., clinical and MRI features) restricted prediction accuracy.

**Future Directions**

* + Integrating multi-modal data (e.g., genomic and imaging features) to improve prediction robustness.
  + Developing lightweight models suitable for resource-constrained settings.

**Closing Remarks**:  
“This research highlights the immense potential of combining technology with medicine. Thank you for your attention, and we welcome your questions.”