**Precision-Driven Breast Cancer Diagnosis: Feature Engineering Meets Machine Learning**

**Project Progress Report**

## Completed Tasks

We have made significant progress in our project, focusing on the classification of breast tumors as benign or malignant using machine learning models. The following tasks have been completed:

- **Data Acquisition and Preprocessing:**
  - Collected the Breast Cancer Diagnosis dataset from Kaggle.
  - Handled missing values using imputation techniques.
  - Removed outliers and scaled numerical features.
  - Encoded categorical variables for compatibility with machine learning models.
  - Applied SMOTE to balance class distribution.
- **Exploratory Data Analysis (EDA):**
  - Analyzed correlations to identify significant predictive factors.
  - Visualized data distributions and relationships between features.
  - Identified tumor size, regional node involvement, and age as key variables.
- **Model Selection and Initial Implementation:**
  - Implemented multiple machine learning models, like Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and Gradient Boosting.
  - Used GridSearchCV and k-fold cross-validation for hyperparameter tuning.
  - Conducted initial performance evaluation using accuracy, precision, recall, and F1-score.

## Challenges

### Encountered Difficulties

- Target Variable Clarification: Based on professor feedback, we refined our objective to focus on classifying tumors as benign or malignant rather than predicting survival time.
- Feature Engineering Complexity: Identifying the most relevant predictive features required extensive correlation analysis and domain knowledge.

- Class Imbalance: The dataset showed an uneven distribution of benign and malignant cases, requiring the use of SMOTE for data balancing.

**Solutions and Future Approach**

- Feature Selection Optimization: We are using feature importance analysis to refine our selected features.
- Handling Class Imbalance: We plan to test cost-sensitive learning as an alternative balancing method.
- Computational Challenges: Optimizing grid search parameters to reduce processing time.

**Collaboration**

- Meeting Frequency: Our group meets weekly to discuss progress, challenges, and next steps.
- We resolved disagreements through thorough discussions and consensus.
- Task Contributions: Each team member is actively contributing to different phases of the project. So far, there have been no concerns about workload distribution.
- **Tools Used:**
    - Data Processing: Pandas, NumPy
    - Visualization: Matplotlib, Seaborn
    - Model Implementation: Scikit-learn, XGBoost, LightGBM
    - Version Control: Git and GitHub for collaborative coding

**Next Steps**

**Remaining Tasks**

1. Feature Selection Refinement: Conduct feature importance analysis to optimize predictive power.
2. Model Training and Evaluation: Train classification models using optimized hyperparameters.
3. Comparative Analysis: Assess model performance based on key evaluation metrics (accuracy, precision, recall, F1-score, confusion matrix).

4. Final Report and Visualization: Summarize findings, create meaningful visualizations, and document key insights.

**Plan to Complete Remaining Tasks**

- Allocated tasks to team members based on expertise.
- Maintain weekly check-ins to monitor progress and adjust timelines.
- Utilize cloud computing resources if necessary for computational efficiency.

**Potential Challenges**

- Refining model performance without overfitting.
- Ensuring balanced contributions from all team members as workloads increase.
- Finalizing a compelling and well-structured report for submission.

This progress update reflects our ongoing commitment to achieving high predictive accuracy in breast cancer classification while adhering to project deadlines and quality standards.