PART 2: PREDICTING TROOP BETRAYAL IN THE WAR AGAINST THE PHRYGIANS

Team members:

Akshat Namdeo
Debangan Sarkar
Abhinav Singh Naruka
Ashi Gupta

TEAM GREY

IIT ROORKEE

Overall Workflow

1. Step 1: Data Preparation
2. Step 2: Statistical Analysis
3. Step 3: Feature Selection
4. Step 4: Data Preprocessing
5. Step 5: Handling Class Imbalance
6. Step 6: Model Training and Hyperparameter Tuning
7. Step 7: Model Evaluation
8. Step 8: Scalability

Our idea

(Important key factors)

Data Collection:
All Key factors

(greed, loyalty, etc)

Providing significance to each factor

(Using Z-test and hypothesis testing)

Feature Extraction: Taking into consideration only the factors having significance >0.05

Risk Assessment: Based on the model's predictions, rank each soldier by their betrayal risk.

Dynamic Updates: As new data is collected (e.g., new offers made by the Phrygians, mission success/failure updates), re-train the model periodically to ensure accuracy.

Hypothesis Formation (Key Factors):
Following are the factors that could influence a soldier's decision to betray their clan. These factors could be divided into different categories:

1. **Greed Index**: Measure of a soldier's materialistic tendencies
2. **Loyalty Score**: Based on past actions and time served
3. **Family** :Number of close family members in the clan
4. **Performance Rating**: Recent combat and mission performance
5. **Disciplinary Record**: Number and severity of infractions
6. **Social Connections**: Strength of bonds with fellow soldiers
7. **Ideology Alignment**: How closely their beliefs align with the clan's
8. **Financial Stress**: Level of personal debt or financial difficulties
9. **Promotion Prospects**: Likelihood of advancing in rank
10. **Job Satisfaction**: Overall contentment with their role

11. **Psychological Resilience**: Ability to cope with stress and adversity
12. **Cultural Integration**: How well the soldier has adapted to clan culture
13. **External Influences**: Exposure to Phrygian propaganda or contacts
14. **Health Status**: Physical and mental health condition
15. **Personal Grievances**: Number of unresolved complaints or disputes
16. **Mission Success Rate**: Percentage of successful missions participated in
17. **Resource Access:** Level of access to sensitive information or resources
18. **Leadership Potential**: Assessed capability for future leadership roles
19. **Peer Evaluation**: Average rating given by fellow soldiers
20. **Training Performance**: Scores from regular training exercises
21. **Off-Duty Behavior**: Conduct during leisure time (e.g., substance use, gambling)
22. **Communication Patterns**: Frequency and nature of communications with outsiders
23. **Adaptability Score**: Ability to adjust to new situations or technologies
24. **Clan Heritage**: Number of generations the soldier's family has served the clan
25. **Reward History**: Frequency and significance of received accolades

PROVIDING DATASET

Based on the collected data, engineer the following features:

- Greed Score: Calculated based on economic status and interest in rewards.
- Loyalty Score: Determined by history of dedication, mission success rate, and respect from superiors.
- Temptation Score: Based on external pressures, economic background, and proximity to the Phrygians.
- These scores would be represented as numerical values (e.g., between 0 and 1), making them suitable for feeding into a machine learning model.

STATISTICAL ANALYSIS & FEATURE SELECTION

(PROVIDING WEIGHTAGE/SIGNIFICANCE TO EACH FACTOR)

- To determine whether a given factor (greed, loyalty, etc.) has a significant impact on the decision making we use a statistical model that determines the same.

- **Z-Test**: For each feature (excluding the target variable), a Z-test is performed to compare the means of the loyal and betrayal risk groups.
- **Hypothesis Testing**

  Null Hypothesis (H_0): There is no significant difference in the means of the two groups

  Alternative Hypothesis (H_1): There is a significant difference in the means of the two groups.

- **Results**: The Z-statistic and p-value (alpha) are calculated. If alpha < 0.05, the feature is considered statistically significant.

- **Results Summary**: The Z-statistics and significance levels are printed for each feature, providing insight into which features may influence betrayal risk.

PROVIDING WEIGHTAGE/SIGNIFICANCE TO EACH FACTOR

❖ **This is the part that shows uniqueness of our solution**

```python
# Step 1: Statistical Analysis using Z-Test
results = {}
for feature in df.columns[:-1]:  # Exclude the target variable
    loyal_scores = df[df['betrayal_risk'] == 0][feature]
    betrayal_scores = df[df['betrayal_risk'] == 1][feature]

    # Perform Z-test
    z_stat, alpha = stats.ttest_ind(loyal_scores, betrayal_scores, equal_var=False)
    results[feature] = {
        'Z-statistic': z_stat,
        'alpha': alpha,
        'Significant': alpha < 0.05
    }
```

- **This code uses probability and statistics to determine the significant factors for a particular soldier, the data taken here is of various soldiers and mean and standard deviation is calculated accordingly for Z-test.**

  This data is then used to train the model and rest of the operations are performed.

Data Preprocessing

• Standardization: The selected features are standardized using StandardScaler. This transforms the data to have a mean of 0 and a standard deviation of 1, ensuring that all features contribute equally to the model training.

Handling Class Imbalance

• SMOTE: To address potential class imbalance (where one class may have significantly fewer samples), the Synthetic Minority Over-sampling Technique (SMOTE) is applied. This generates synthetic samples for the minority class (betrayal risk) to balance the dataset, improving model performance.

MODEL TRAINING AND HYPERPARAMETER TUNING

• *Train-Test Split:* The dataset is divided into training and testing sets using an 80-20 split, ensuring that the model can be evaluated on
unseen data.

• *Model Selection*: A RandomForestClassifier is chosen as the base model due to its robustness and effectiveness in classification tasks.

• *Hyperparameter Tuning*: GridSearchCV is utilized to find the best hyperparameters for the model. The parameters being tuned include:

- n estimators: Number of trees in the forest.
- max_depth: Maximum depth of each tree.
- min_samples_split: Minimum number of samples required to split an intemal node.

• *Training*: The model is trained using the training data, and the best hyperparameters are determined through cross-validation.

MODEL EVALUATION

- *Prediction*: The best model is used to predict outcomes on the test set.
- *Performance Metrics*: The model's performance is assessed using two key metrics:
  - Accuracy. The proportion of correctly predicted instances out of the total instances.
  - FI Score: The harmonic mean of precision and recall, providing a balance between the two, especially useful for imbalanced datasets.
- *Results Output*: The optimized model's accuracy and FI score are printed, giving an indication of how well the model can predict betrayal risk based on the input features.

Languages:
Python: The entire pipeline is written in Python, a widely used language for data analysis and machine learning.

Python Libraries Used:

- **Pandas** (pd): For data manipulation and reading the CSV dataset.
- **NumPy** (np): For efficient numerical operations.
- **SciPy** (scipy. stats): For performing statistical analysis (t-tests).
- **Scikit-learn** (sklearn): For machine learning models, data preprocessing, and model evaluation.
- **Imbalanced-learn** (imblearn): To handle class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).

FULL STACKS USED IN THE IMPLEMENTATION

1. <u>Data Collection</u>
   - *Pandas* (import pandas as pd) is used for handling and manipulating the dataset, which contains features like greed, loyalty, temptation, and performance.
   - Data is collected manually and loaded into a Pandas DataFrame for analysis.

2. <u>Statistical Analysis</u>
   - *Scipy* (from scipy import stats) is employed for statistical analysis. A Z-test (t-test here) is used to identify significant differences between loyal and betrayal risk groups for each feature.
   - Significant features are selected for further processing.

3. Modelling

- *Scikit-learn* (from sklearn.ensemble import RandomForestClassifier) is used for machine learning. The main model is a Random Forest Classifier, which is chosen for its robustness and ability to handle both continuous and categorical data.
- GridSearchCV and RandomizedSearchCV are used to perform hyperparameter tuning, helping to find the best model configuration.

4. Data Processing

- *StandardScaler* is used to standardize the data, ensuring features have a mean of 0 and a standard deviation of 1, which helps many machine learning algorithms perform better.
- SMOTE (from imblearn.over_sampling import SMOTE) addresses class imbalance by oversampling the minority class (betrayal risk).

5. Model Training and Evaluation

- Data is split into training and testing sets using train_test_split.
- After hyperparameter tuning, the model's performance is evaluated using accuracy and F1 score metrics from *Scikit-learn*.
- Final results are printed, providing insight into how well the model can predict betrayal risk.

Thank you