**1. Project Motivation & Need**

**Q:** Why did you choose this topic?  
**A:** Himalayan expeditions are high-risk and costly. Accurate prediction can save lives, reduce financial loss, and support informed decision-making for climbers, companies, and insurers.

**Q:** What is the real-world application of your project?  
**A:** It can help expedition companies assess risks before investment, insurers set fair premiums, and climbers make safer decisions.

**2. Problem Statement & Objective**

**Q:** What exact problem are you solving?  
**A:** We are predicting the likelihood of expedition success—defined as reaching the summit and returning safely—using machine learning.

**Q:** Why not use existing methods?  
**A:** Existing methods are mostly descriptive (past trends), lack predictive accuracy, and are not designed for decision-making.

**3. Dataset & Features**

**Q:** What dataset did you use and why?  
**A:** Kaggle’s Himalayan Climbing Expeditions dataset, based on Elizabeth Hawley’s archives. It’s comprehensive, covering 1905–2019 with team, climber, and peak data.

**Q:** What are the key features?  
**A:** Climber age, experience, team size, season, hired staff, oxygen use, peak height/difficulty, and outcome (success/failure).

**Q:** What is your target variable?  
**A:** Expedition outcome – Success (1) or Failure (0).

**4. Research Gap**

**Q:** What gap does your work address?  
**A:** Lack of predictive modeling, no end-to-end ML pipeline, and absence of practical, deployable solutions in previous work.

**5. Methodology**

**Q:** How did you process the data?  
**A:** Data integration, missing value handling, feature engineering (scaling, encoding), and feature selection using correlation analysis.

**Q:** How do you evaluate your models?  
**A:** Using stratified cross-validation and metrics like Accuracy, Precision, Recall, F1-score, and AUC-ROC.

**6. Machine Learning Basics**

**Q:** Why did you choose supervised learning for this problem?  
**A:** Our target variable (success/failure) is labeled, making this a classification problem, which suits supervised learning.

**Q:** What type of classification problem is this?  
**A:** It’s a binary classification problem (Success = 1, Failure = 0).

**Q:** What is overfitting and how do you avoid it?  
**A:** Overfitting is when a model performs well on training data but poorly on unseen data. We prevent it using cross-validation, regularization (L1/L2), and ensemble methods.

**7. Models Used & Specifics**

**Q:** Which models did you use and why?  
**A:** Logistic Regression (baseline), KNN (pattern recognition), Decision Tree (rule-based), Neural Network (non-linear relationships), and Balanced Random Forest (handles imbalance well).

**Q:** Which model performed best?  
**A:** We expect ensemble models like Balanced Random Forest to perform best due to class imbalance and non-linear patterns.

**Logistic Regression**

**Q:** How does Logistic Regression work?  
**A:** It predicts probabilities using the logistic (sigmoid) function and classifies based on a threshold.

**Q:** Why use Logistic Regression as a baseline?  
**A:** It’s simple, interpretable, and provides a benchmark for comparison.

**K-Nearest Neighbors (KNN)**

**Q:** How does KNN classify data?  
**A:** It finds the k nearest training points to a test point and assigns the majority class among them.

**Q:** How do you choose K?  
**A:** Using cross-validation – small K overfits, large K underfits.

**Decision Tree (DT)**

**Q:** How does a Decision Tree split nodes?  
**A:** Using metrics like Gini Impurity or Information Gain to find the best feature split.

**Q:** Main drawback of Decision Trees?  
**A:** Overfitting, solved by pruning or using ensembles like Random Forest.

**Neural Network (MLP)**

**Q:** Why use a Neural Network?  
**A:** It captures complex non-linear relationships that simpler models may miss.

**Q:** What is backpropagation?  
**A:** An algorithm to update weights by propagating error backward and minimizing loss.

**Balanced Random Forest (BRF)**

**Q:** How does BRF handle class imbalance?  
**A:** It balances each sample by under-sampling the majority class before training each tree.

**Q:** Why BRF over normal Random Forest?  
**A:** It ensures fair representation of minority classes, improving prediction for successes.

**8. Model Evaluation**

**Q:** Why not rely on Accuracy alone?  
**A:** Accuracy can mislead with imbalanced data; Precision, Recall, F1-score, and AUC-ROC give a more complete picture.

**Q:** What is an ROC curve?  
**A:** A graph showing the trade-off between True Positive Rate and False Positive Rate.

**Q:** Why Stratified K-Fold Cross-Validation?  
**A:** It preserves class distribution in each fold, ensuring fair evaluation.

**9. Challenges & Mitigation**

**Q:** What challenges did you face?  
**A:** Class imbalance, missing data, feature complexity, and overfitting.

**Q:** How did you handle class imbalance?  
**A:** Using SMOTE, class weighting, and ensemble methods.

**10. Expected Outcomes & Value**

**Q:** What do you expect as results?  
**A:** High predictive accuracy, robust performance on unseen data, key feature insights, and a deployable prediction system.

**Q:** How will this help in real life?  
**A:** Enables data-driven decisions, reduces risks, and supports planning, insurance, and investment.

**11. Technical & Future Scope**

**Q:** What technologies did you use?  
**A:** Python (Pandas, NumPy, Scikit-learn), Jupyter Notebook, Git/GitHub, and visualization tools (Matplotlib, Seaborn, Plotly).

**Q:** Future scope?  
**A:** Real-time prediction dashboard/API, integration with weather data, global scaling.

**12. General/Advanced**

**Q:** Why did you choose an ensemble model?  
**A:** Ensembles improve generalization, reduce variance, and handle imbalance effectively.

**Q:** Could deep learning do better?  
**A:** Possibly, but classical ML is more efficient and interpretable for this dataset size.