In this project, you will work with a Panic Disorder Detection dataset that contains patient demographic information, symptoms, medical history, lifestyle factors, and mental health assessment results. The goal is to analyze the dataset, preprocess the data, resolve class imbalance, extract meaningful features, and evaluate machine learning models for predicting panic disorder.

Follow the step-by-step instructions below to complete the project.

Step 1: Data Exploration & Understanding (15 points)

(1.1) Load the dataset (panic_disorder_detection.csv) using pandas.
(1.2) Print basic information about the dataset (.info(), .describe(), etc).
(1.3) Check for missing values and duplicate entries.
(1.4) Perform an initial visualization of categorical variables (Gender, Symptoms, Social Support, etc.). Use plotting library to use (e.g., matplotlib, seaborn)

- Use countplot or barplot for categorical variables.
- Possibly group by Panic Disorder Diagnosis to see how variables like Gender, Symptoms, Social Support differ across classes.

(1.5) Identify class imbalance in the target variable (Panic Disorder Diagnosis).

- Print the distribution of classes.
- Visualize the imbalance using a bar chart.

Deliverable: A well-documented section with exploratory data analysis (EDA) and plots. EDA is a crucial first step in any data science project, as it helps uncover patterns, spot anomalies, understand feature distributions, and gain valuable insights that guide data cleaning, feature selection, and modeling decisions.

Step 2: Data Cleaning & Transformation (10 points)

(2.1) Handle missing values (either drop or impute them appropriately like mean/mode).
(2.2) Remove duplicate values (ensure label proportions remain unchanged).
(2.3) Convert categorical columns into consistent format (fix typos, lowercase, remove extra spaces).
(2.4) Convert date-related fields (if any) into proper datetime format.
(2.5) Handle outliers in numerical columns (Age, Severity, etc.). Visualize distributions first (e.g., boxplots or histograms) and possibly remove or cap/floor extreme values.
Deliverable: Explanations of transformations applied.

Step 3: Handling Class Imbalance (10 points)

(3.1) Analyze the imbalance ratio in Panic Disorder Diagnosis.
(3.2) Apply at least two class imbalance handling techniques:

- Oversampling using SMOTE (Synthetic Minority Oversampling Technique).
- Random oversampling of the minority class or random undersampling of the majority class.

- Class-weight adjustments in models (alternative to resampling).

(3.3) Compare model performance with and without class balancing.
Deliverable: A section discussing class distribution before and after applying balancing techniques. Explanation of why class balancing is important in medical diagnosis models.

Step 4: Feature Engineering (10 points)

4.1 Feature Engineering on Structured Data
(4.1.1) Encode categorical variables using one-hot encoding or label encoding.
(4.1.2) Scale numerical features (Age, Severity Score, etc.) using techniques like min-max, z-score, etc
(4.1.3) Create new features based on existing columns (e.g., Impact on Life levels). (If necessary)
Deliverable: Explanations of feature engineering applied.

Step 5: Feature Importance Analysis (15 points)

5.1 Correlation Analysis (5 points)
(5.1.1) Compute the correlation matrix to see relationships between numerical features. Plot a heatmap of the correlation matrix.
(5.1.2) Colinearity - Identify highly correlated features that might be redundant (if any).

5.2 Feature Importance from Models (10 points)
(5.2.1) Train a Random Forest model and extract feature importance scores.
(5.2.2) Train a Logistic Regression model and analyze coefficients.
(5.2.3) Visualize feature importance using a bar chart.
Deliverable: Insights on which features contribute most to predictions.

Step 6: Model Training & Evaluation (20 points)

6.1 Model Selection & Training (10 points)
(6.1.1) Split the dataset into training and test sets (80-20 or 70-30 split).
(6.1.2) Train at least three different classification models:
- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)

(6.1.3) Use cross-validation to tune hyperparameters.

6.2 Model Evaluation & Comparison (10 points)
(6.2.1) Compute evaluation metrics:
- Accuracy
- Precision, Recall, F1-score
- ROC Curve & AUC Score

(6.2.2) Compare all models before and after handling class imbalance.

(6.2.3) Visualize and compare results of different models using bar plots.
Deliverable: Trained classification models should be accompanied by a thorough performance evaluation, including a comparative analysis of different models. Emphasize how data cleaning, feature selection, and class balancing have influenced the overall model performance.

Step 7: Ethical Considerations (5 points)

(7.1) Discuss bias and fairness in mental health predictions.
(7.2) Explain potential ethical concerns when using AI for mental health assessments.
Deliverable: A short write-up (100-150 words).

Bonus (Optional)

Step 8: Implement SHAP (SHapley Additive exPlanations) to explain how each feature impacts the model's predictions.

Submission (15 points):

- (5 points) Include a Jupyter Notebook (.ipynb file) with all your clean code, outputs, and comments above each cell (notebook) for each task.
- (10 points) Submit a formal report in PDF format that clearly explains each step of the project. The report should include your findings along with relevant outputs, visualizations (plots), and results. Ensure that each step—data exploration, preprocessing, modeling, and evaluation—is thoroughly documented with clear explanations and insights.

Due Date: Monday, May 5th, 11:59 pm. Late submissions are not accepted.