**Foundation program training: Data Science Comprehensive package**

**Capstone Project 1: Crop Recommendation System – 15 Marks**

**Project Statement:**

The agricultural sector plays a vital role in the economy, and optimizing crop selection based on environmental and soil conditions can significantly improve yield and farmer income. This project aims to develop a robust crop recommendation system using the provided dataset, which contains information on various environmental factors (Temperature, Humidity, Rainfall, PH), soil nutrients (Nitrogen, Phosphorous, Potassium, Carbon), soil type, and the corresponding crop.

**You are required to implement the following steps:**

**Phase 1: Exploratory Data Analysis (EDA)**

1. **Initial Data Inspection** 
   * Load the “Agriculture\_dataset.csv” into a panda DataFrame.
   * Display the first 10 rows of the DataFrame to get a glimpse of the data.
2. **Univariate and Bivariate Analysis:**
   * **Count Plots:** Generate count plots for the categorical features (Soil and Crop) to visualize the distribution of different soil types and crop categories. Analyze and interpret the findings.
   * **Box Plots:** Create box plots to visualize the distribution of each numerical feature (Temperature, Humidity, Rainfall, PH, Nitrogen, Phosphorous, Potassium, Carbon) for each different Crop. Analyze these plots to understand the relationship between environmental/soil factors and crop types, identifying potential variations and central tendencies.
3. **Outlier Detection and Handling:**
   * Identify potential outliers in the numerical features using appropriate visualization techniques (e.g., box plots).
   * Discuss the potential reasons for these outliers in the context of agricultural data.
   * Implement a suitable strategy to handle the detected outliers (e.g., capping, removal, or leaving them as is with justification). Clearly document the chosen approach and the reasoning behind it.
4. **Missing Value Analysis:**
   * Check for missing values in the DataFrame using appropriate methods
5. **Correlation Analysis:**
   * Calculate the correlation matrix for the numerical features.
   * Visualize the correlation matrix using a heatmap.
   * Interpret the heatmap to identify any strong positive or negative correlations between the features. Discuss the potential implications of these correlations for crop growth.

**Phase 2: Statistical Inference**

1. **Analysis of Variance (ANOVA) Hypothesis Testing:**
   * Formulate a null and an alternative hypothesis to test if there is a statistically significant difference in the mean values of a specific numerical feature (e.g., Rainfall) across different Crop categories.
   * Perform a one-way ANOVA test for the chosen numerical feature and the Crop variable.
   * State the assumptions of ANOVA and briefly check if they are likely to be met by the data (no formal testing is required, just a conceptual check based on your EDA).
   * Interpret the results of the ANOVA test (p-value and F-statistic).
   * Based on the ANOVA results, conclude whether there is a statistically significant difference in the mean of the chosen numerical feature across different crop types.

**Phase 3: Machine Learning Model Building and Evaluation**

1. **Data Preprocessing for Machine Learning:**
   * Identify the features (independent variables) and the target variable (Crop).
   * Perform appropriate encoding for the categorical features (Soil). Choose a suitable encoding technique (e.g., one-hot encoding) and justify your choice.
   * Split the dataset into training and testing sets using an appropriate ratio (e.g., 80:20). Ensure stratification based on the Crop column to maintain class proportions.
   * Consider scaling the numerical features. Choose a suitable scaling method (e.g., StandardScaler, MinMaxScaler) and justify your choice. Apply the scaling to the training and testing sets.
2. **Model Training:**
   * Train the following classification models on the training data:
     + **Random Forest Classifier:** Train a Random Forest Classifier model with appropriate hyperparameters (you can start with default values or perform a basic grid search for better performance).
     + **Support Vector Machine (SVM):** Train an SVM classifier (you can start with a linear kernel or experiment with others).
     + **Decision Tree Classifier:** Train a Decision Tree Classifier model.
3. **Performance Evaluation:**
   * Use the trained models to predict the Crop on the testing data.
   * Evaluate the performance of each model using the following metrics:
     + Accuracy
     + Precision
     + Recall
     + F1-Score
   * Generate a classification report for each model, which includes these metrics for each crop category.
4. **Results Tabulation and Comparison:**
   * Create a table summarizing the performance metrics (Accuracy, Precision, Recall, F1-Score - you can report the macro or weighted average) for all three models on the testing data.
   * Compare the performance of the different models and discuss their strengths and weaknesses in the context of this crop recommendation task.
   * Conclude which model performs best based on the evaluation metrics and provide reasons for its superior performance.

**Phase 4: Conclusion**

1. **Conclusion :**
   * Summarize the key findings from your EDA, statistical inference, and machine learning model building phases.
   * Discuss the implications of your results for developing a crop recommendation system.

**Capstone Project 2: Web Traffic Prediction – 15 Marks**

**Project Statement:**

Accurate forecasting of web traffic is crucial for effective resource allocation, capacity planning, and ensuring a seamless user experience. This project aims to develop a time series forecasting model to predict future web traffic based on historical data. The provided dataset “web traffic.csv” contains time-stamped web traffic counts, which will be used to analyze traffic patterns and build a predictive model. The student will implement the following steps:

**Phase 1: Time Series Preprocessing and Exploration**

1. **Data Loading and Initial Inspection [5 Marks]:**
   * Load the “web traffic.csv” dataset into a panda’s DataFrame.
   * Convert the 'Timestamp' column to a datetime format.
   * Set the 'Timestamp' column as the index of the DataFrame.
2. **Time Series Visualization :**
   * Plot the 'TrafficCount' against the 'Timestamp' to visualize the raw time series.
   * Decompose the time series into its trend, seasonal, and residual components using an appropriate method (e.g., seasonal decomposition using moving averages or the seasonal\_decompose function from the statsmodels library).
   * Plot the decomposed components (trend, seasonality, residual) separately and interpret the observed patterns. Identify any apparent trends, seasonality, or irregular fluctuations.
3. **Autocorrelation Analysis:**
   * Plot the Autocorrelation Function (ACF) of the 'TrafficCount’ series.
   * Plot the Partial Autocorrelation Function (PACF) of the 'TrafficCount' time series.
   * Analyze the ACF and PACF plots to identify potential autoregressive (AR) and moving average (MA) components in the time series. Explain how the patterns in the ACF and PACF plots relate to the underlying time series characteristics.
4. **Stationarity Check:**
   * Explain the concept of stationarity in time series analysis and its importance for modeling.
   * Perform the Augmented Dickey-Fuller (ADF) test to check for stationarity of the 'TrafficCount' time series.
   * State the null and alternative hypotheses of the ADF test.
   * Interpret the results of the ADF test (test statistic and p-value).
   * If the time series is not stationary, apply appropriate differencing to make it stationary. Repeat the ADF test on the differenced series to confirm stationarity.
   * Clearly state the order of differencing required (d value).

**Phase 2: SARIMA Model Building and Evaluation**

1. **SARIMA Model Identification:**
   * Based on the analysis of the ACF and PACF plots (from Phase 1) and the stationarity analysis (including the order of differencing), determine the initial values for the SARIMA model parameters:
     + p (autoregressive order)
     + d (differencing order) - from Phase 1
     + q (moving average order)
     + P (seasonal autoregressive order)
     + D (seasonal differencing order)
     + Q (seasonal moving average order)
     + s (seasonal period) - Determine the seasonality (e.g., daily, weekly).
   * Explain the reasoning behind your choice of initial parameter values, linking them to the observed time series characteristics, ACF/PACF patterns, and seasonality.
   * Consider a range of plausible values around your initial estimates.
2. **Model Training and Selection:**
   * Split the time series data into training and testing sets. A common split is 80% for training and 20% for testing but justify your choice. Ensure the split maintains the time series order.
   * Train several SARIMA models with different combinations of the parameters (p, d, q, P, D, Q, s) identified in the previous step. You can use a grid search or iterative approach to explore different parameter combinations.
   * For each trained model, record the AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion) value.
   * Select the SARIMA model with the lowest AIC or BIC value as the best model. Justify your model selection based on the information criterion used. Explain the trade-off between model fit and complexity that AIC/BIC balances.
3. **Model Evaluation:**
   * Use the selected SARIMA model to generate predictions on the testing set.
   * Calculate the following evaluation metrics to assess the model's performance:
     + Mean Squared Error (MSE)
     + Root Mean Squared Error (RMSE)
     + Mean Absolute Error (MAE)
   * Plot the original time series, the training data, and the predicted values on the testing set in a single plot. Visually assess the model's fit and predictive accuracy.
   * Interpret the evaluation metrics and the plot. Discuss the model's strengths and weaknesses in forecasting web traffic.
4. **Residual Analysis [5 Marks]:**
   * Calculate the residuals (the difference between the actual and predicted values) from the selected SARIMA model on the training data.
   * Plot the residuals over time.
   * Create a histogram or density plot of the residuals.
   * Plot the ACF and PACF of the residuals.
   * Perform the Ljung-Box test to check if the residuals are white noise (i.e., randomly distributed with no autocorrelation).
   * Interpret the results of the residual analysis. Discuss whether the residuals exhibit any patterns or autocorrelation, which would indicate that the model is not capturing all the information in the time series.

**Phase 3: Conclusion**

1. **Conclusion:**
   * Summarize the key steps taken in the project, from data preprocessing and exploration to model building and evaluation.
   * State the final SARIMA model parameters and its performance on the testing data.