1. Dataset Introduction

1a. Load the dataset "train-val.csv" into your notebook.

2. Exploratory Data Analysis (EDA)

Use code cells with appropriate EDA techniques to understand the dataset and text cells to explain your findings.

Using appropriate methods, present the following information through code:

2a. The number of samples and features in the dataset.
2b. The types of features in the dataset.
2c. The labels of the features.
2d. The number of categories.
2e. How many samples belong to each category.
2f. The correlation between the data.
2g. Any other information you believe is helpful for understanding the dataset.

⚠️ During the preprocessing of the dataset (both train-val and test), **DO NOT REMOVE THE ID COLUMN** as it is required for the Kaggle competition.

3. Dataset Preprocessing

💡 Use the Column Transformer to create and apply separate transformers for numerical and categorical data.

3a. Will you use all features of the dataset to train the classifiers, or will you select some? Will you combine some features to create new features for your model?

3b. Are there any missing values? Write appropriate code to handle these values.

3c. Write code to appropriately transform categorical variables so that the classifiers you use can handle them.

3d. Write code to scale the features, if you deem it necessary.

3e. Perform all the preprocessing steps (using transformers) so that the initial dataset is "clean" and ready to be used for classifier training.

3f. After cleaning the initial training set, write the appropriate code to split it into features (X) and target (y).

💡 The column 'RainTomorrow' is the target variable the classifier should predict.

3g. Split the dataset into a training set (train set) and a validation set (validation set) with a split ratio of 70% for training and 30% for validation.

4. Model Training with Default Classifier Parameter Values

We will train the following classifiers using the training set of our dataset:

1. Naive Bayes
2. KNeighborsClassifier
3. LogisticRegression

4. MLP with one hidden layer
5. SVC
6. Decision Tree
7. Random Forest

4a. Train (fit) all seven classifiers mentioned above (show the samples from the train set along with their labels as input to each classifier) using default parameter values.

4b. Apply the trained models to the validation set, using only the samples without their labels (predict).

4c. Compare the output of each model with the corresponding labels of the validation set and evaluate their performance using the F1 score.

4d. Evaluate the overall performance of the models using a graph (e.g., histogram, bar plot) and comment on which model performed best.

5. Preparing the File for the First Kaggle Submission

5a. Use **ONLY** your best-performing model to make predictions using the test set.

5b. Save the predictions from your best model to a CSV file. The CSV file should contain two columns: the first column should contain the id from the test set file, and the second column should contain the corresponding predictions made by your best-trained model (step 5a).

6. Model Optimization with Hyperparameter Tuning

6a. For the seven classifiers, optimize their performance using grid search with cross-validation (5-fold) to find the best hyperparameters.

6b. Apply the trained models to the validation set, using only the samples without their labels (predict).

6c. Compare the output of each model with the corresponding labels of the validation set and evaluate their performance using the F1 score.

6d. Evaluate the overall performance of the models using a graph (e.g., histogram, bar plot) and comment on which model performed best.

7. Preparing the File for the Second Kaggle Submission

7a. Use **ONLY** your best-performing model to make predictions using the given test set.

7b. Save the predictions from your best model to a CSV file. The CSV file should contain two columns: the first column should contain the id from the test set file, and the second column should contain the corresponding predictions made by your best-trained model (step 7a).