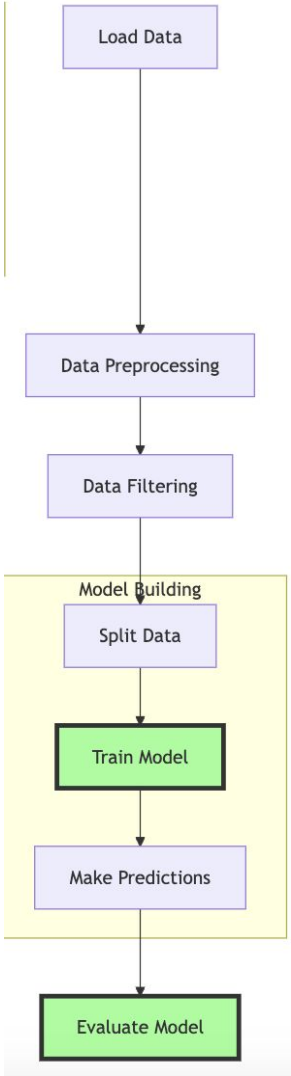
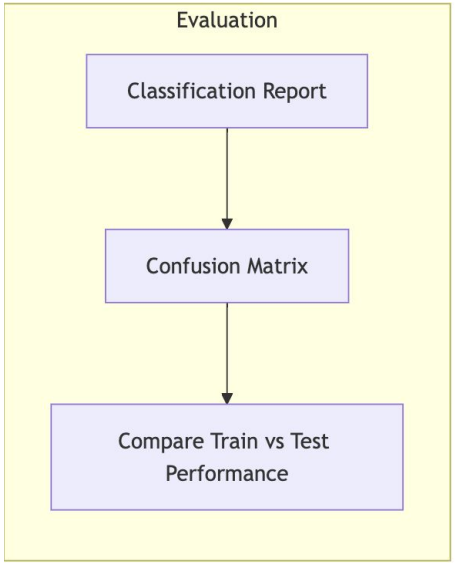
Week 9
Discussion lab 8
Machine Learning

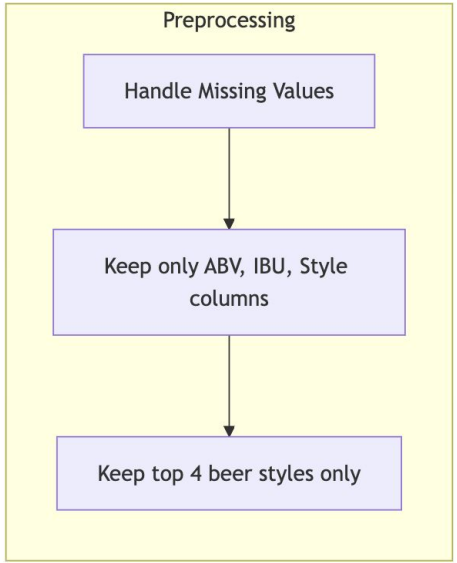
Deadlines

DUE DATES

- Quiz 7 is due Monday, May 19
- Checkpoint #2: EDA is due Wednesday, May 28
- Discussion lab 8 is due Friday May 30







1. Data preprocessing: Handle missing values and extract ABV, IBU features
2. Filter data to keep only top 4 most common beer styles
3. Split data into training (80%) and test (20%) sets
4. Train SVM model and generate predictions
5. Evaluate model using classification reports and confusion matrices for both training and test sets

Part I: Data, Wrangling, & EDA

1. Analyze missing values(**.isnull().sum(axis=0)**)

	Name	ABV	IBU
0	Beer1	5.0	45.0
1	Beer2	NaN	60.0
2	Beer3	7.5	NaN
3	None	4.8	NaN

	Name	ABV	IBU
0	False	False	False
1	False	True	False
2	False	False	True
3	True	False	True

```
Name    1
ABV     1
IBU     2
dtype: int64
```

2. Remove rows with missing values in style, abv, ibu(**dropna(subset=[])**)

3. Merge beer and brewery datasets(left join)
   - Why left join?
   - How does left join works?
4. Filter dataset to keep only top 4 styles (**.value_counts()[:].index.tolist()**)

Part II : Prediction Model

1. **Extract features (X: ABV, IBU) and labels (Y: Style)**

```
data_x = beer_df[['abv','ibu']]
data_y= np.array(beer_df['style'])
```

2. **Split data into train/test sets**

```
train_X = data_x[:num_training]
train_ Y = data_y[num_training:]
test_X = data_x[:num_training]
test_ Y = data_x[num_training:]
```

3. Train SVM model and generate predictions

```
beer_clf = train(train_X, train_ Y)

beer_clf.predict(train_X)
beer_clf.predict(test_X)
```

Part III : Model Assessment

1. Generate classification reports (precision, recall, f1-score)

2. Create confusion matrices

3. Compare training vs testing performance
   Train accuracy vs test accuracy

4. Analyze where model performs well/poorly
   F1 score
5. Evaluate potential overfitting