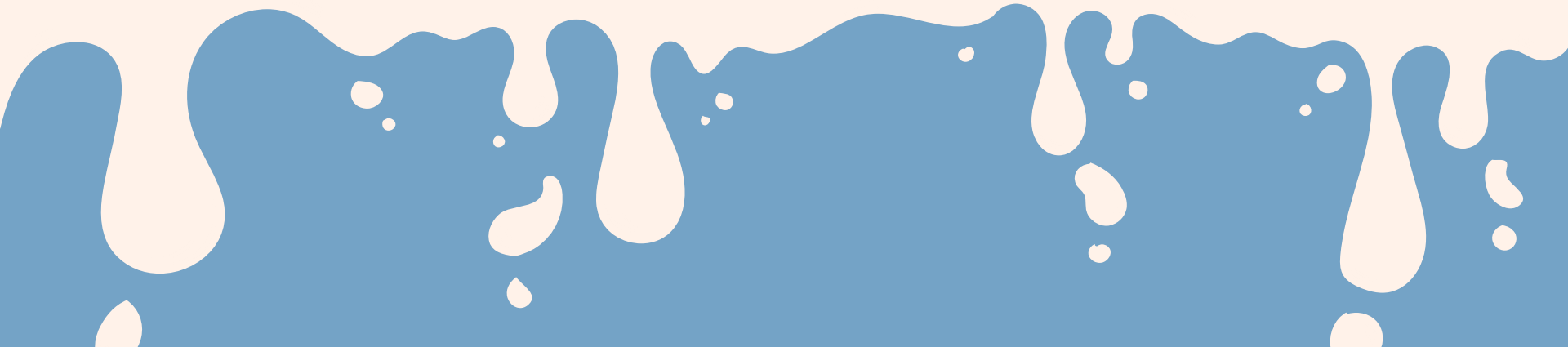
Milk Quality Prediction

Group 2
Kulaphong Jitareerat
Nijiati Abulizi



Introduction:

- **Exploratory Data Analysis**
- **Methods**
  - Modelling Flow
  - Data Preprocessing
- **Performance Evaluation**
- **Recommendation**

Data Set

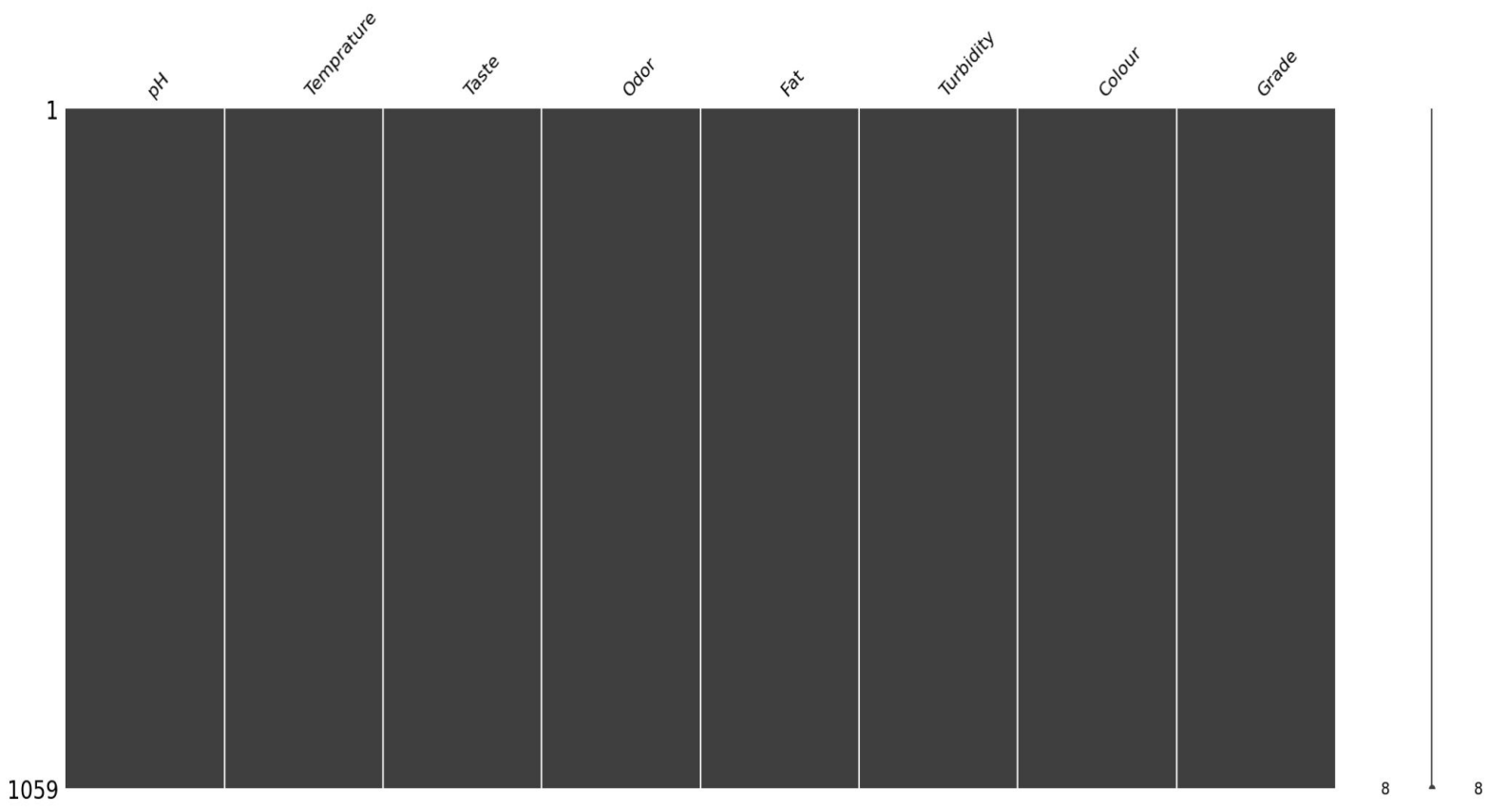
Quantitative

- **pH** (3 to 9.5)
- **Temperature** (34'C to 90'C)
- **Color** (240 to 255)

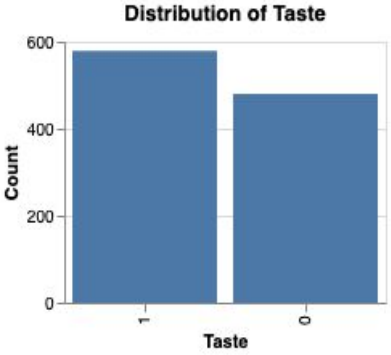
Qualitative

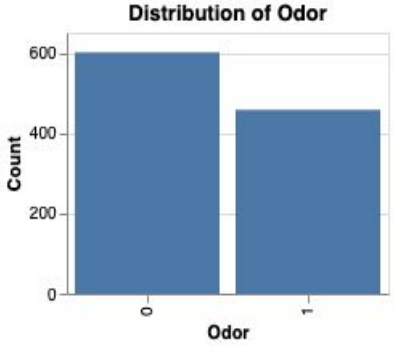
- **Taste** (0/Bad or 1/Good)
- **Odor** (0/Bad or 1/Good)
- **Fat** (0/Low or 1/High)
- **Turbidity** (0/Low or 1/High)
- **Grade** (Target: Low/Bad or Medium/Moderate, High/Good) / **Target**

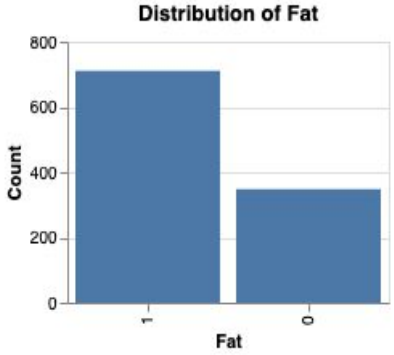
No Missing Data Points Found

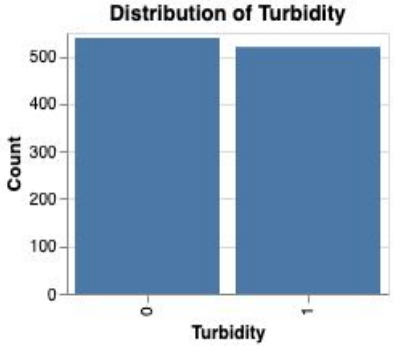


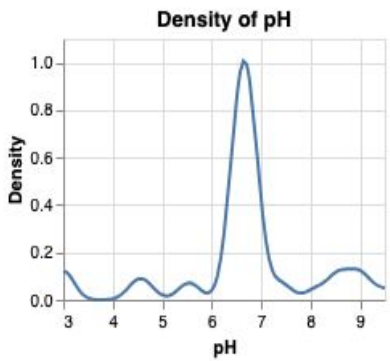
Categorical and Numerical Data Distributions

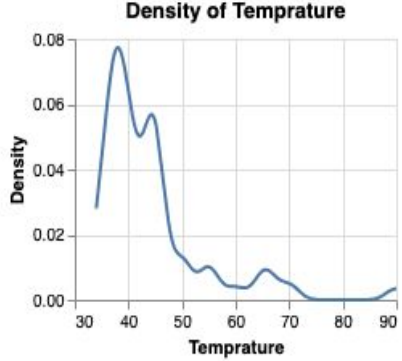


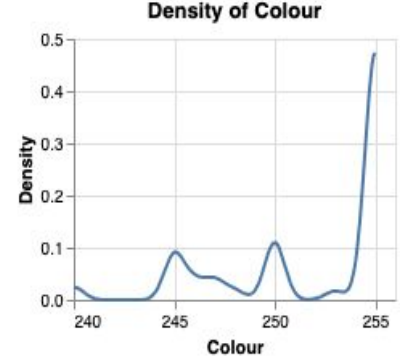




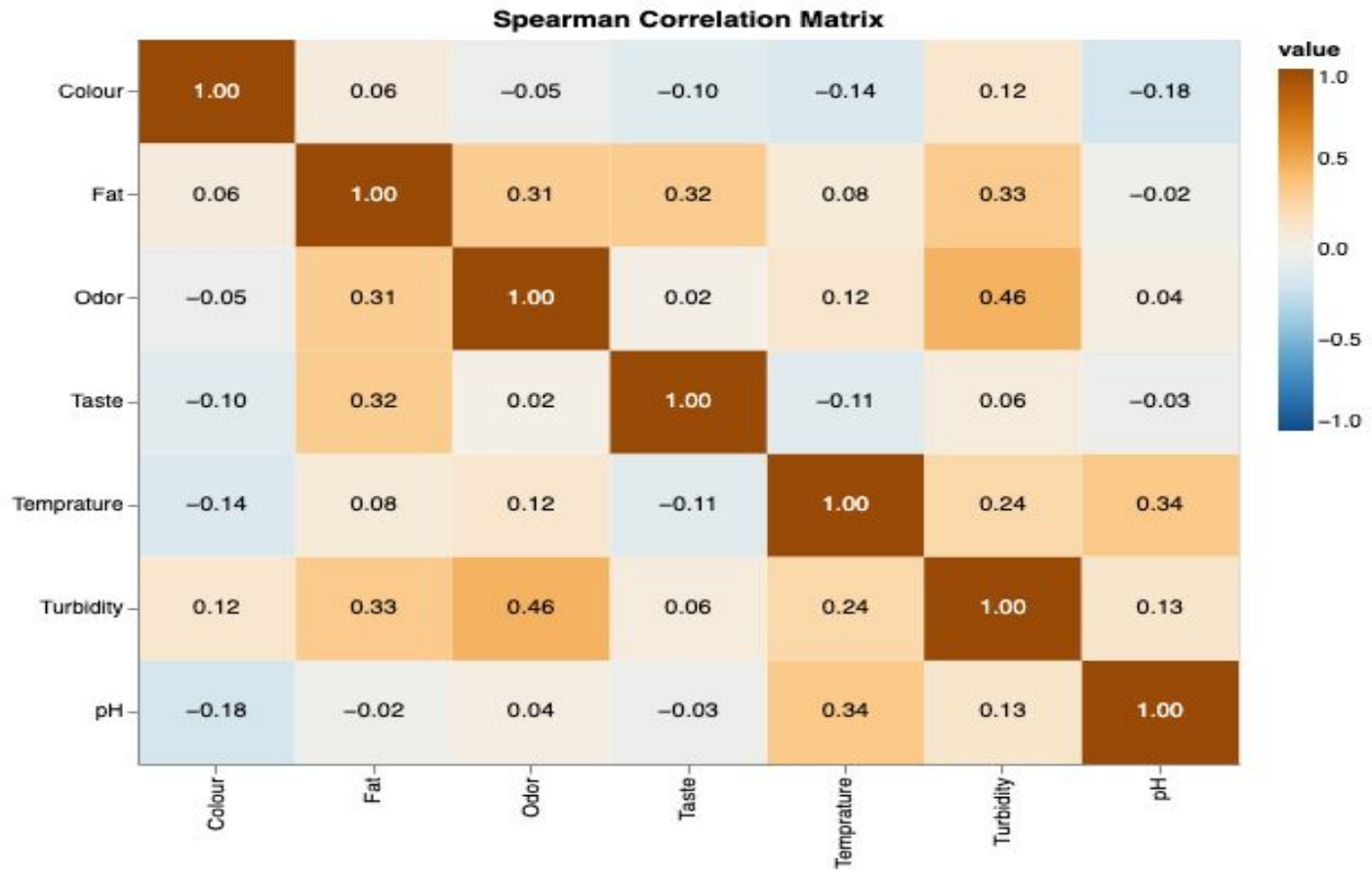








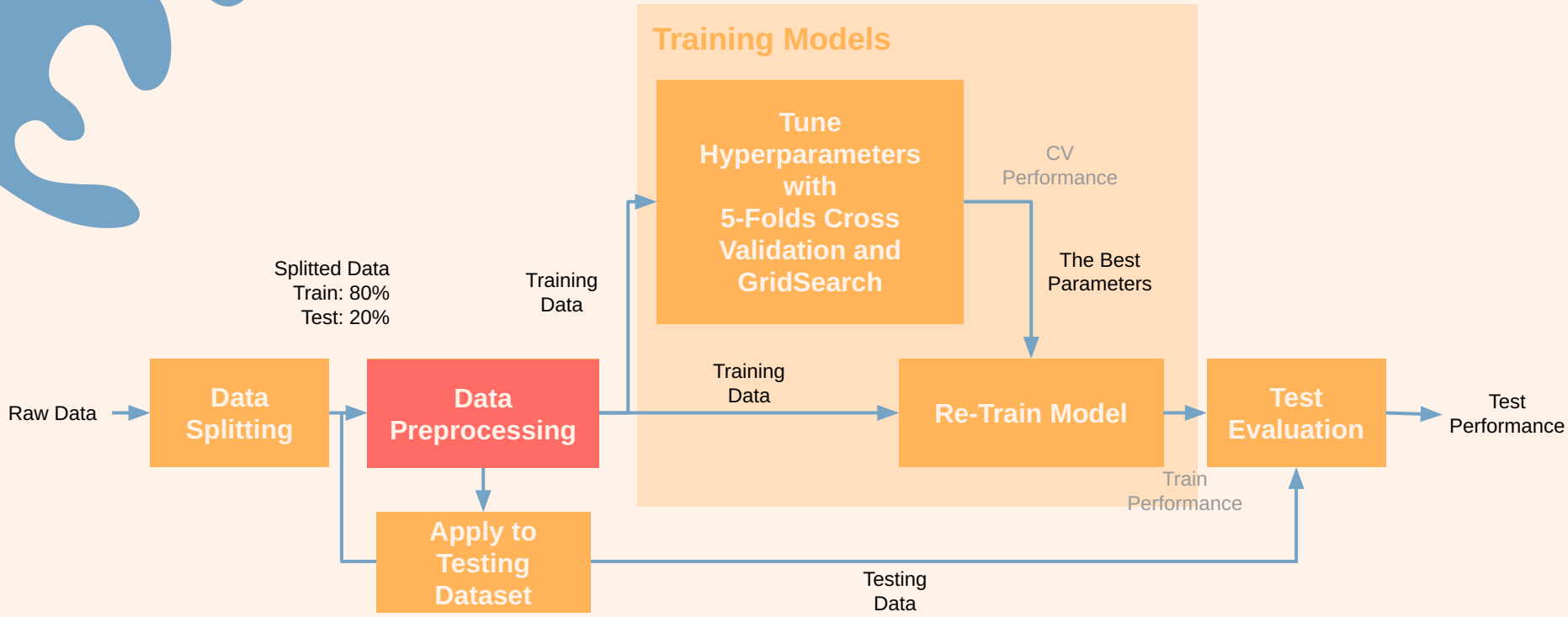
Turbidity is Correlated with Odor



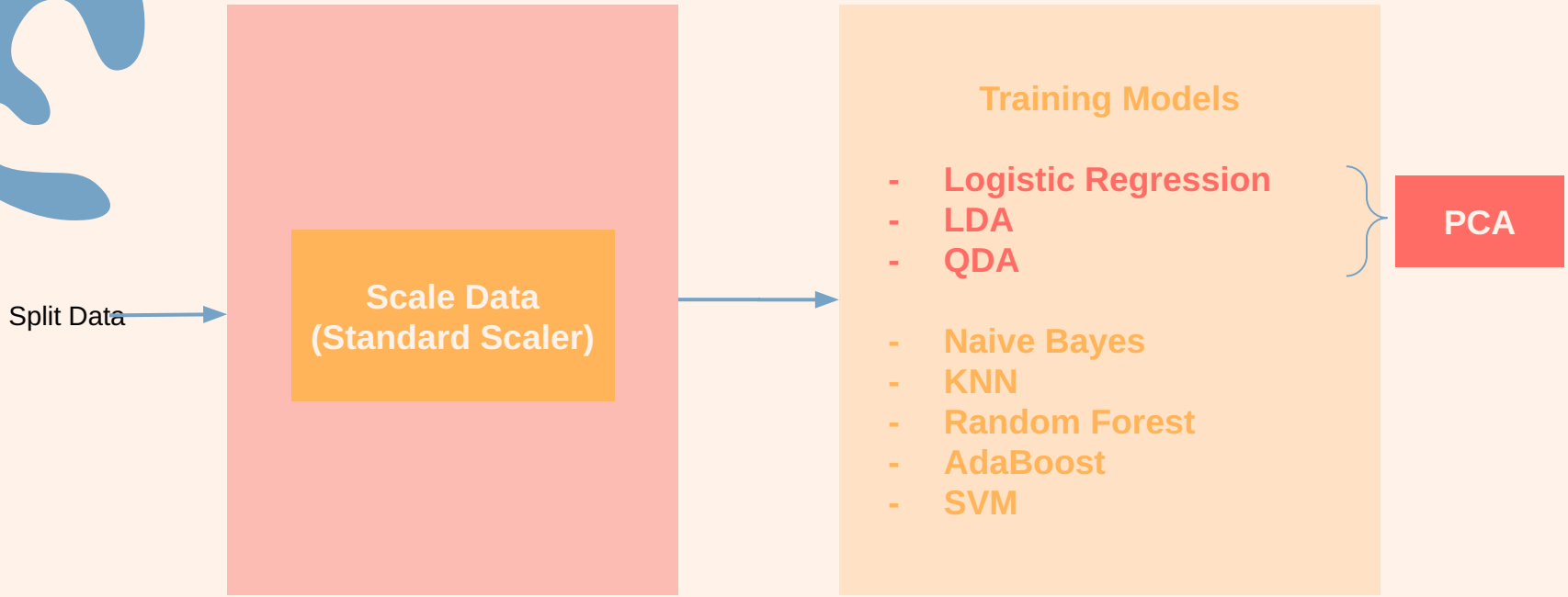
Check Multicollinearity

Variable	VIF
Turbidity	1.426035
Fat	1.368393
Odor	1.359784
Taste	1.190560
Temprature	1.126956
pH	1.110007
Colour	1.096751

Method: Modelling Flow

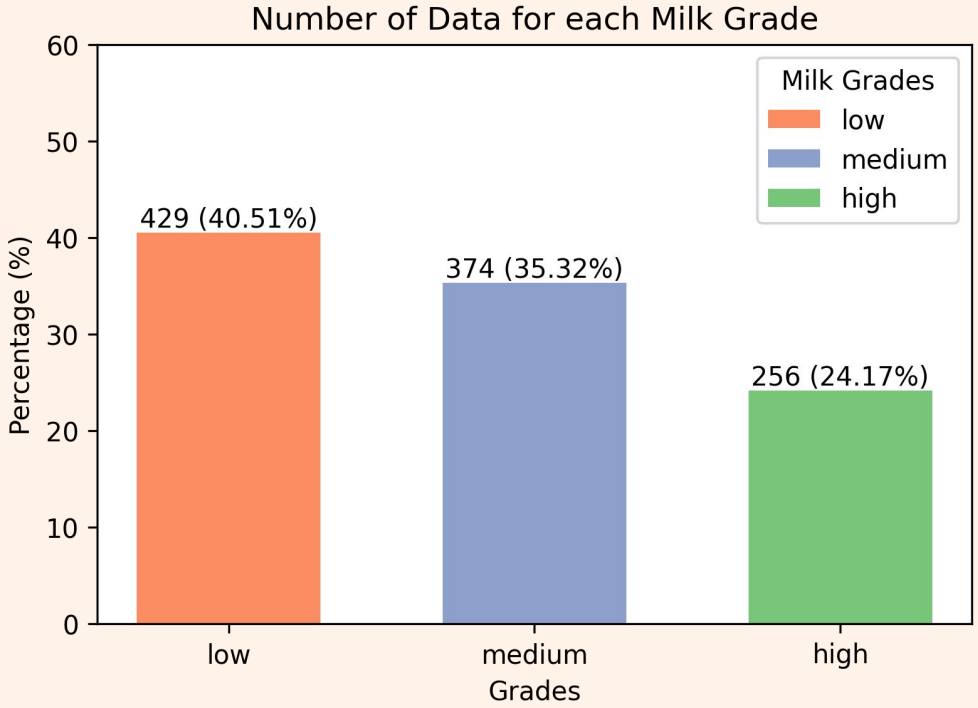


Data Scaling and PCA Transformation for Sensitive Models



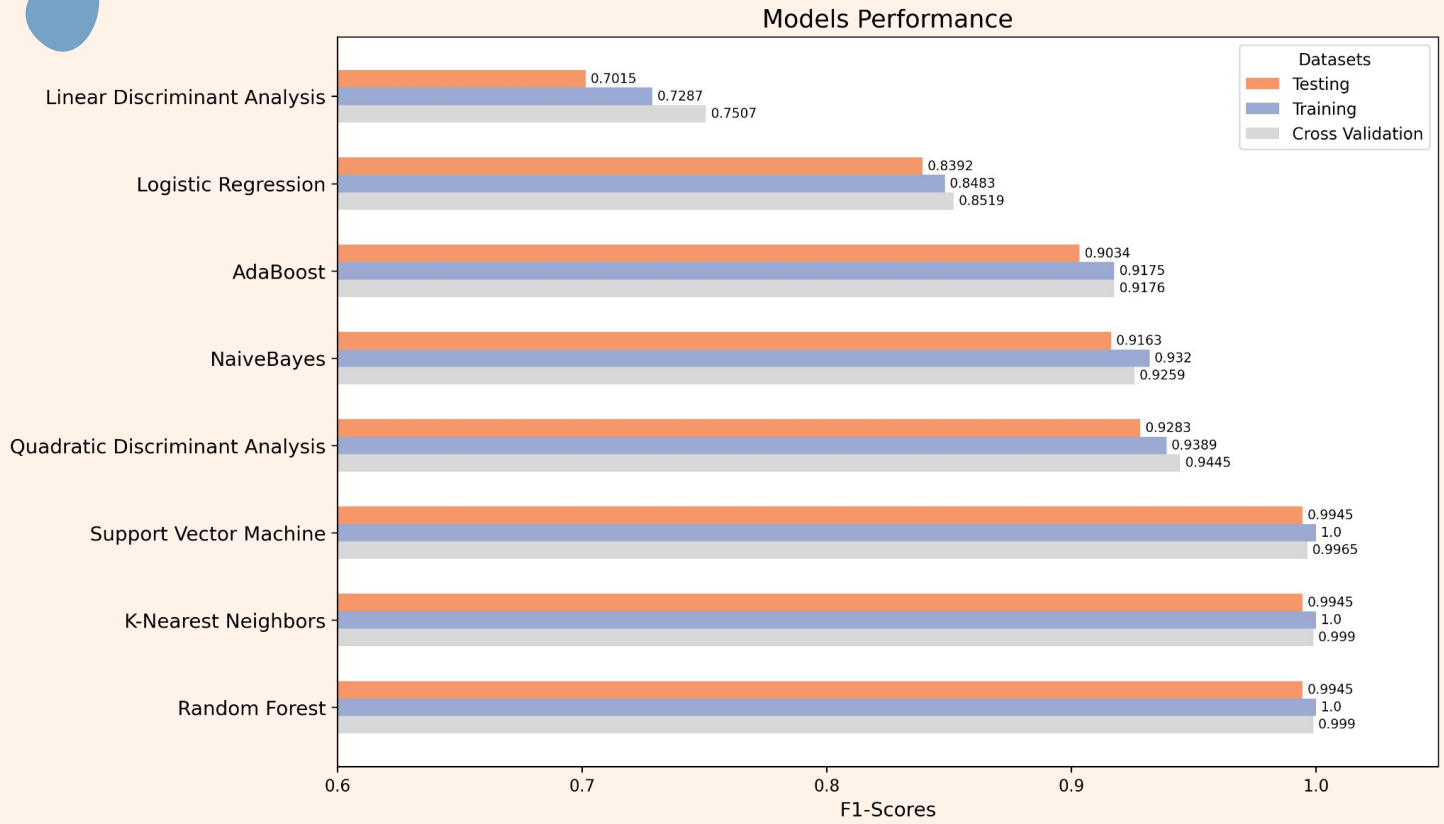
Algorithm	Multicollinearity	Scaling	Reasons for Multicollinearity	Reasons for Scaling
Quadratic Discriminant Analysis	High	High	Assumes normality and independence; highly sensitive to correlated predictors.	Involves calculation of covariance matrices; benefits from consistent feature scales.
Logistic Regression	High	High	Linear model; sensitive to correlated predictors; assumption of independence.	Log-odds interpretation; coefficients sensitive to feature scales.
Linear Discriminant Analysis	High	High	Assumes normality and independence; sensitive to multicollinearity.	Involves covariance matrices; benefits from consistent feature scales.
Support Vector Machine	Low	High	Kernel-based; sensitive to distance metrics; benefits from scaled features.	Relies on distance metrics; improves convergence and performance with scaled features.
Naive Bayes	Low	Low	Conditional independence assumption; less affected by correlated features.	Probability-based; robust to feature scales, but may not benefit from scaling.
AdaBoost	Low	Low	Boosting ensemble; relatively robust to multicollinearity.	Weak learners compensate for issues; scaling may not be critical.
K-Nearest Neighbors	Moderate	High	High sensitivity to distances; affected by feature correlations.	Distance-based algorithm; requires normalized feature scales.
Random Forest	Moderate	Moderate	Ensemble method; less sensitive, but may still be affected by correlated features.	Decision trees are less sensitive to feature scales; minimal impact.

Why we used “F1-Score” metric



- **Imbalance** in each grade
- Interested in both **Precision** and **Recall**

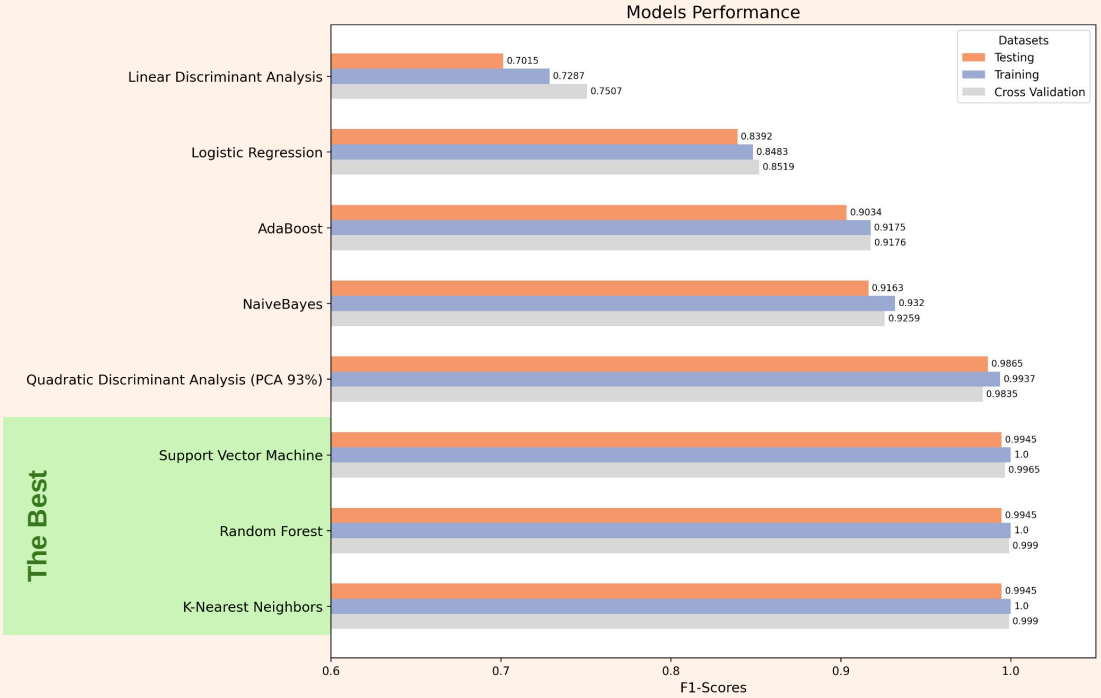
Performance of Models after Scaling



Multicollinearity Consideration for Sensitive Models (LDA, QDA, Logistic Regression)

93% of Variance Retained While Reducing Collinearity

The Best of Each Model



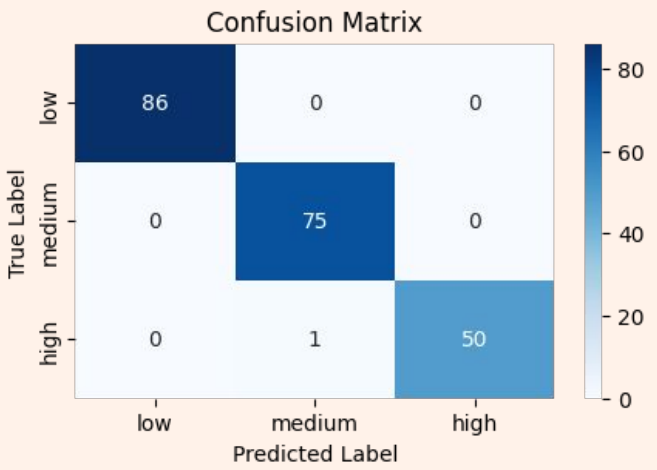
model	score_cv	score_train	score_test	best_params
K-Nearest Neighbors	0.998962	1.000000	0.994492	{'n_neighbors': 21, 'p': 1, 'weights': 'distance'}
Random Forest	0.998962	1.000000	0.994492	{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 10}
Support Vector Machine	0.996550	1.000000	0.994492	{'C': 1, 'gamma': 1, 'kernel': 'poly'}
Quadratic Discriminant Analysis (PCA 93%)	0.983481	0.993745	0.986466	{'reg_param': 0.0}
NaiveBayes	0.925870	0.931992	0.916258	{'var_smoothing': 0.1}
AdaBoost	0.917615	0.917543	0.903386	{'learning_rate': 0.01, 'n_estimators': 100}
Logistic Regression	0.851933	0.848336	0.839155	{'C': 10, 'max_iter': 50, 'penalty': 'l1', 'solver': 'liblinear'}
Linear Discriminant Analysis	0.750690	0.728673	0.701511	{'shrinkage': None, 'solver': 'svd'}

Compare PCA vs Non-PCA

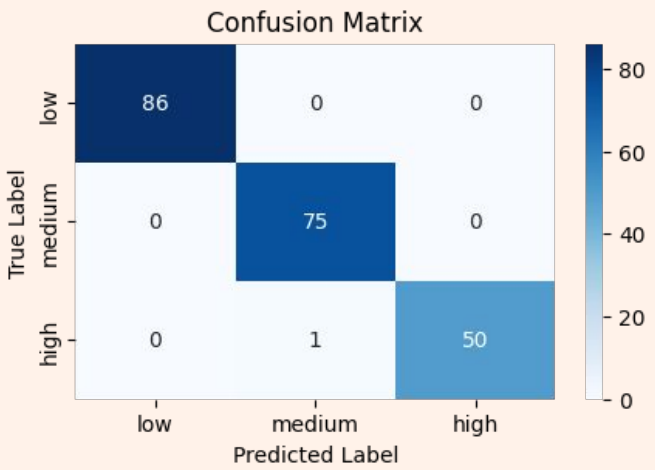
model	score_cv	score_train	score_test	best_params
Random Forest (PCA 93%)	0.997913	1.000000	0.994773	{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
Support Vector Machine (PCA 93%)	0.996550	1.000000	0.994492	{'C': 1, 'gamma': 1, 'kernel': 'poly'}
Support Vector Machine	0.996550	1.000000	0.994492	{'C': 1, 'gamma': 1, 'kernel': 'poly'}
Random Forest	0.998962	1.000000	0.994492	{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 10}
K-Nearest Neighbors	0.998962	1.000000	0.994492	{'n_neighbors': 21, 'p': 1, 'weights': 'distance'}
K-Nearest Neighbors (PCA 93%)	0.998962	1.000000	0.989199	{'n_neighbors': 33, 'p': 2, 'weights': 'distance'}
Quadratic Discriminant Analysis (PCA 93%)	0.983481	0.993745	0.986466	{'reg_param': 0.0}
AdaBoost (PCA 93%)	0.943173	0.975955	0.975131	{'learning_rate': 0.1, 'n_estimators': 50}
Quadratic Discriminant Analysis	0.944525	0.938947	0.928274	{'reg_param': 0.010101010101010102}
NaiveBayes	0.925870	0.931992	0.916258	{'var_smoothing': 0.1}
AdaBoost	0.917615	0.917543	0.903386	{'learning_rate': 0.01, 'n_estimators': 100}
NaiveBayes (PCA 93%)	0.899199	0.889880	0.864846	{'var_smoothing': 0.001}
Logistic Regression	0.851933	0.848336	0.839155	{'C': 10, 'max_iter': 50, 'penalty': 'l1', 'solver': 'liblinear'}
Logistic Regression (PCA 93%)	0.799986	0.798774	0.748371	{'C': 0.001, 'max_iter': 50, 'penalty': None, 'solver': 'saga'}
Linear Discriminant Analysis	0.750690	0.728673	0.701511	{'shrinkage': None, 'solver': 'svd'}
Linear Discriminant Analysis (PCA 93%)	0.736745	0.714251	0.693585	{'shrinkage': 'auto', 'solver': 'lsqr'}

The incorrect prediction

Confusion Matrix of Testing dataset
Random Forest

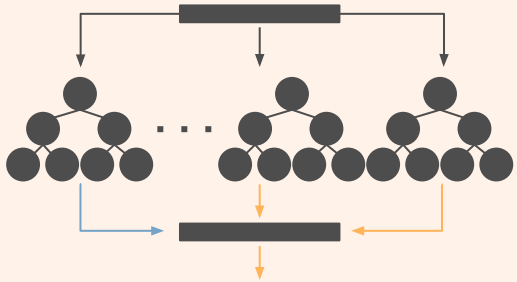


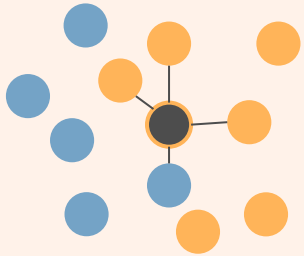
Confusion Matrix of Testing dataset
KNN



index	pH	Temprature	Taste	Odor	Fat	Turbidity	Colour	Grade	pred_test
1	-0.020104	-0.800605	-1.108579	1.132704	-1.446112	1.010683	0.288216	2	1

Recommendation





Random Forest

Pros

- Insights into the importance of each feature
- Robustness against overfitting

Cons

- Interpretability might still be a challenge due to the ensemble of trees

K-Nearest Neighbors

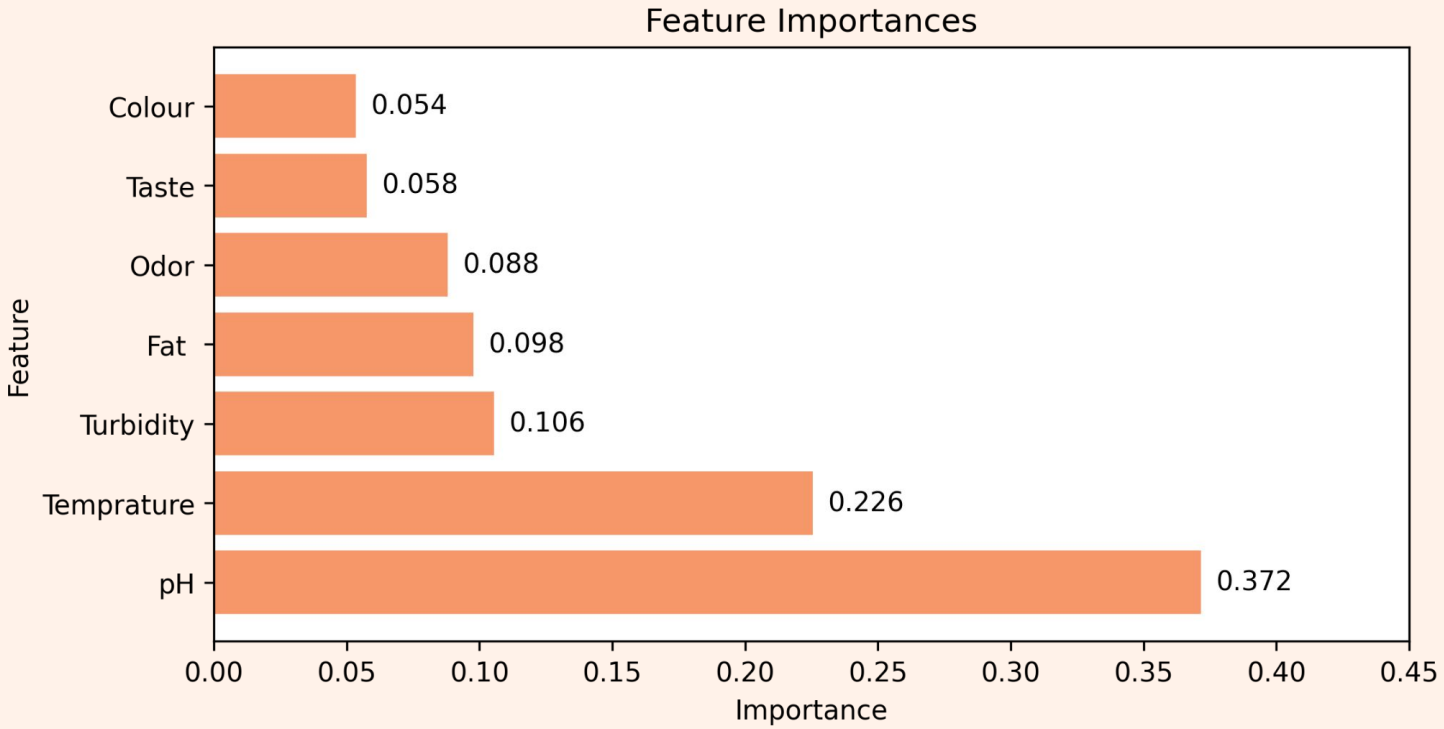
Pros

- Simple and quick to implement
- Interpretability

Cons

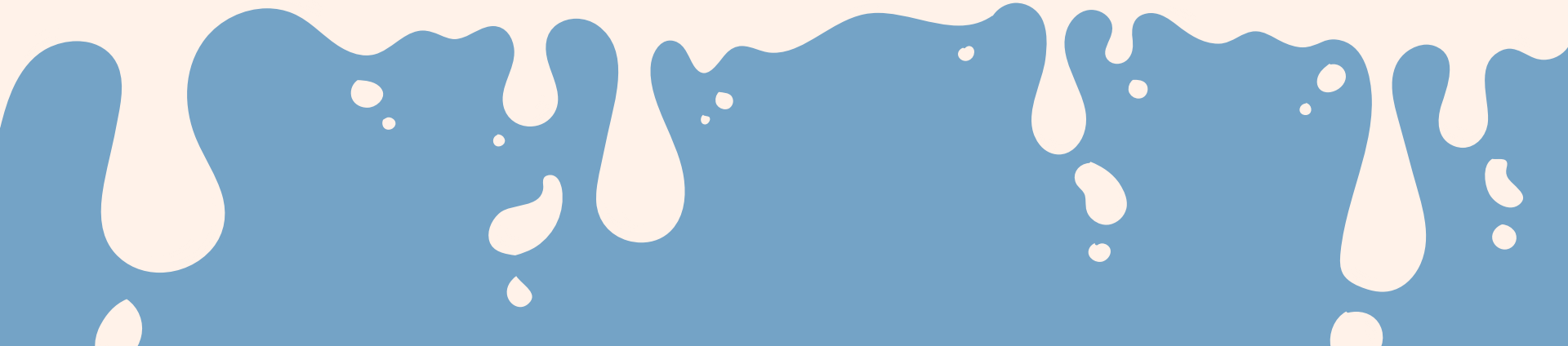
- Scalability as the dataset grows
- Sensitive to outliers

Insight from Random Forest Model



Thank You

Questions?




slidesgo