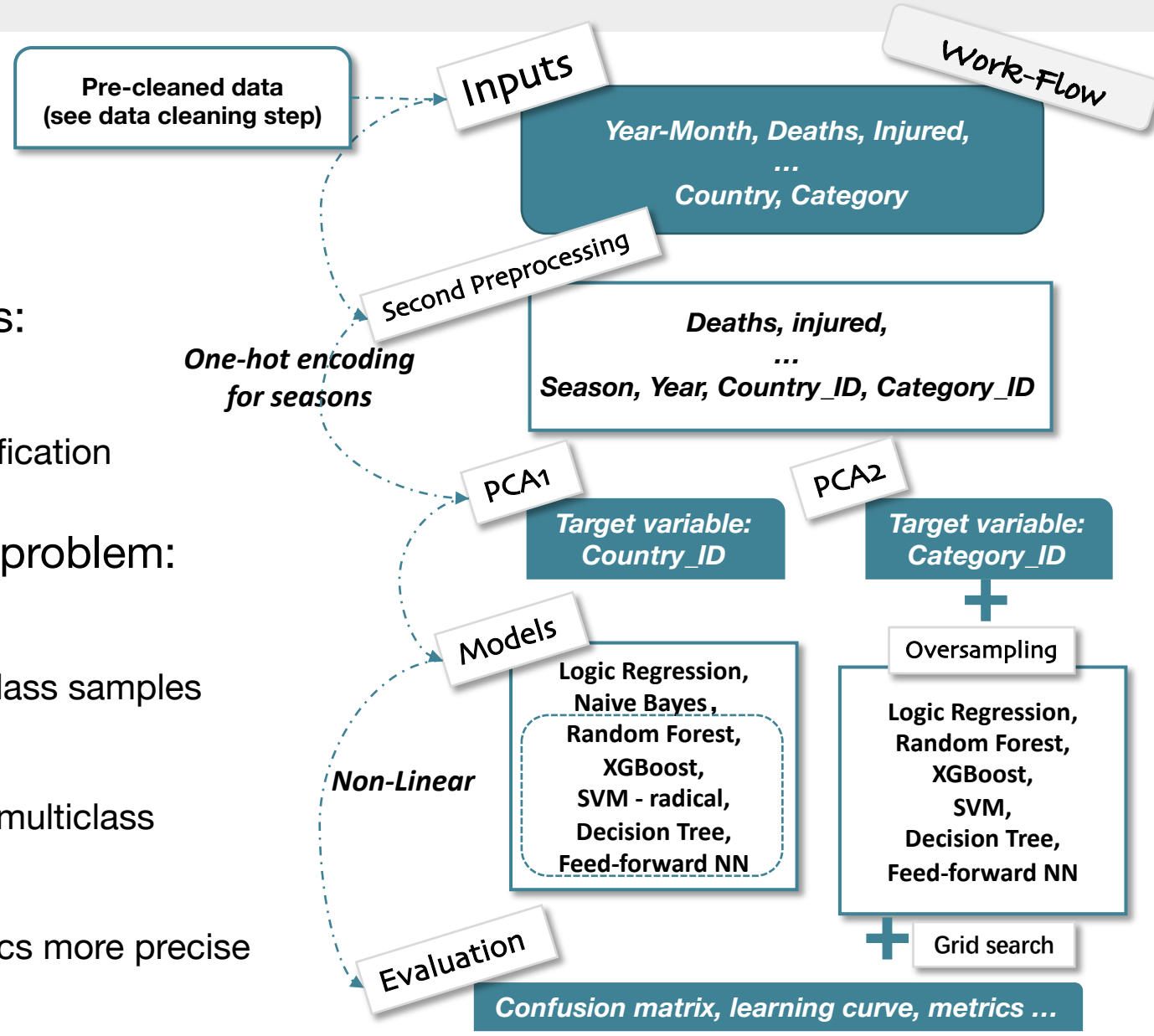
Overview of Classification

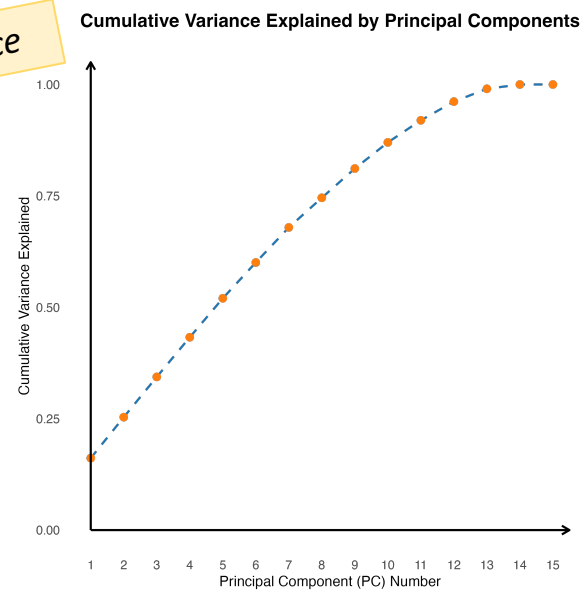
- **Two** separate classification analyses:
  - predict country's continents - *binary* problem
  - predict disaster's types - *multiclass* problem
- **Similar** Challenges appear in two problems:
  - dirty data, full of missing/wrong values
  - Imbalanced data, especially for multiclass classification
- **Various** techniques applied to tackle each problem:
  - Pre-processing disaster categories
  - SMOTE to generate synthetic data for minority class samples
  - PCA technique applied to reduce dimensionality
  - grid search to optimize the hyperparameters for multiclass problem
  - 5-Fold cross-validation applied to evaluate metrics more precise



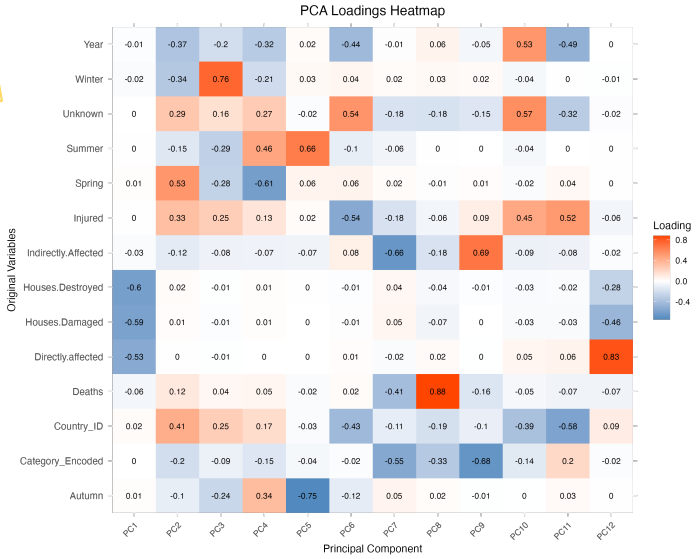
Binary classification

- ➔ **One** target with two label:
  - South America(1) and Africa(2)
- ➔ First **12** principal components were selected
  - capture 99% of the variance
  - loading plot reveals the relationship with original data [1]
- ➔ Various models are trained:
  - Logistic Regression: A simple linear model
  - Random Forest: $mtry = \sqrt{n}$
  - SVM: radial kernel, cost = 1, gamma = 0.1
  - XGBoost: evaluate logarithmic loss during training
  - Decision Tree: CART
  - FFNN: Feedforward Architecture with Single Hidden Layer
  - Naïve Bayes (as a backup)

Cumulative Variance



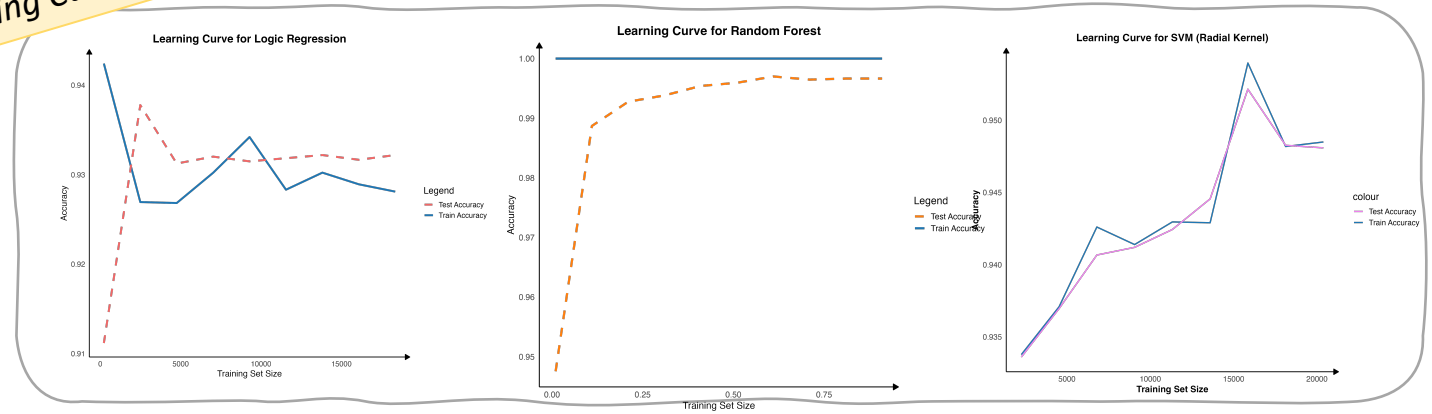
PCA Loading



Binary classification

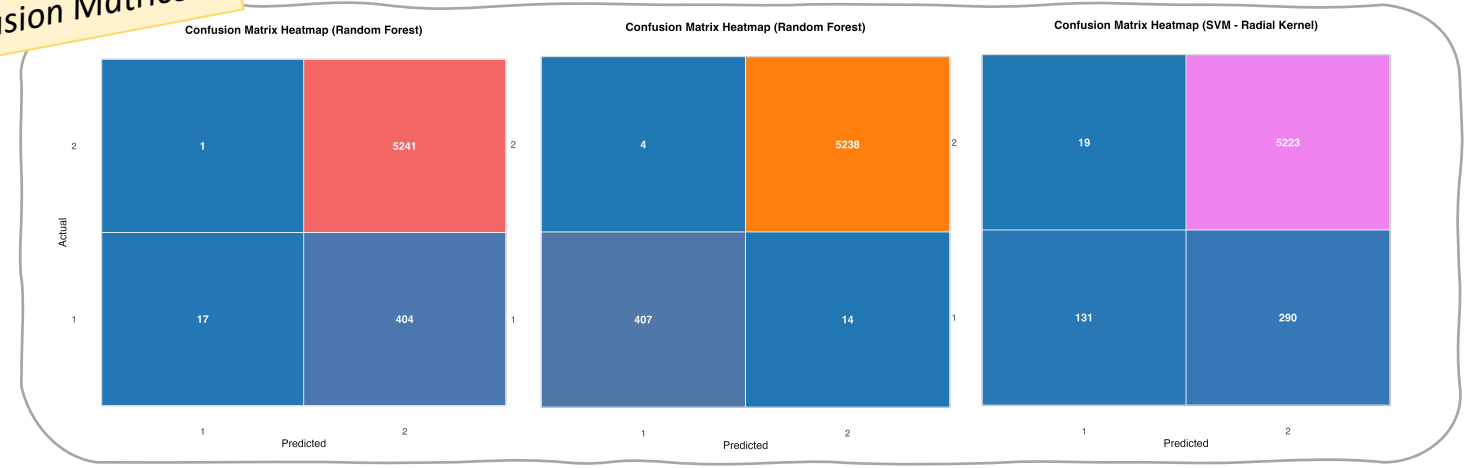
- ➔ Logistic Regression
  - Fluctuating Test Accuracy
  - Possible Underfitting at Small Sizes
  - Low sensitivity and recall
  - Model tend to predict as majority category
- ➔ Random Forest
  - Stable Test Accuracy
  - Consistency Across Sizes
  - Minimal Errors in two classes
  - Consistent Performance for two classes
- ➔ SVM - Radial Kernel
  - Converging Accuracy
  - High Variability at Small Sizes
  - Potential Optimal Training Size Range Identified
  - Higher Misclassification in Class 2
  - High False Negitive Events

Learning Curves



Logic Regression | Random Forest | SVM-radical kernel

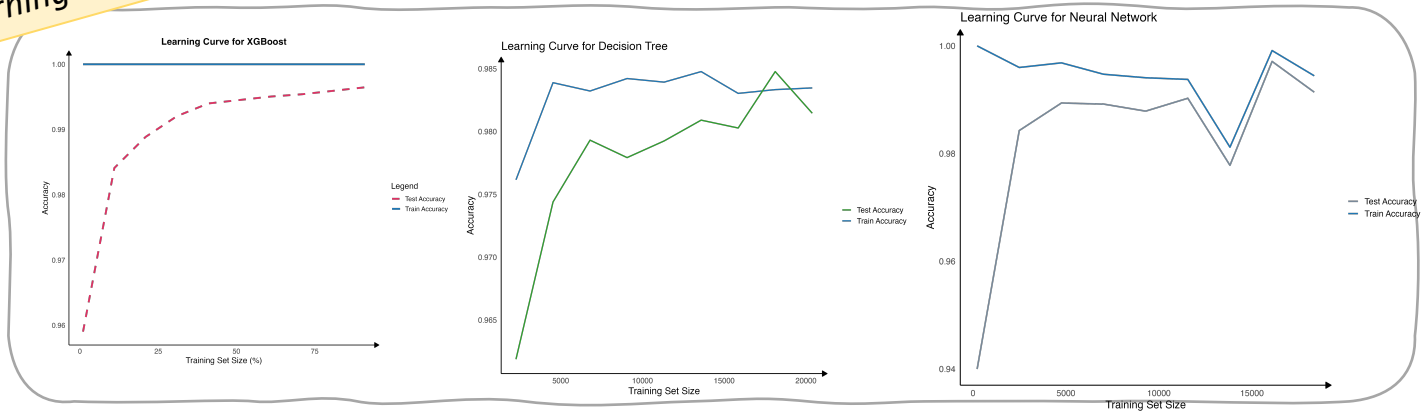
Confusion Matrices



Binary classification

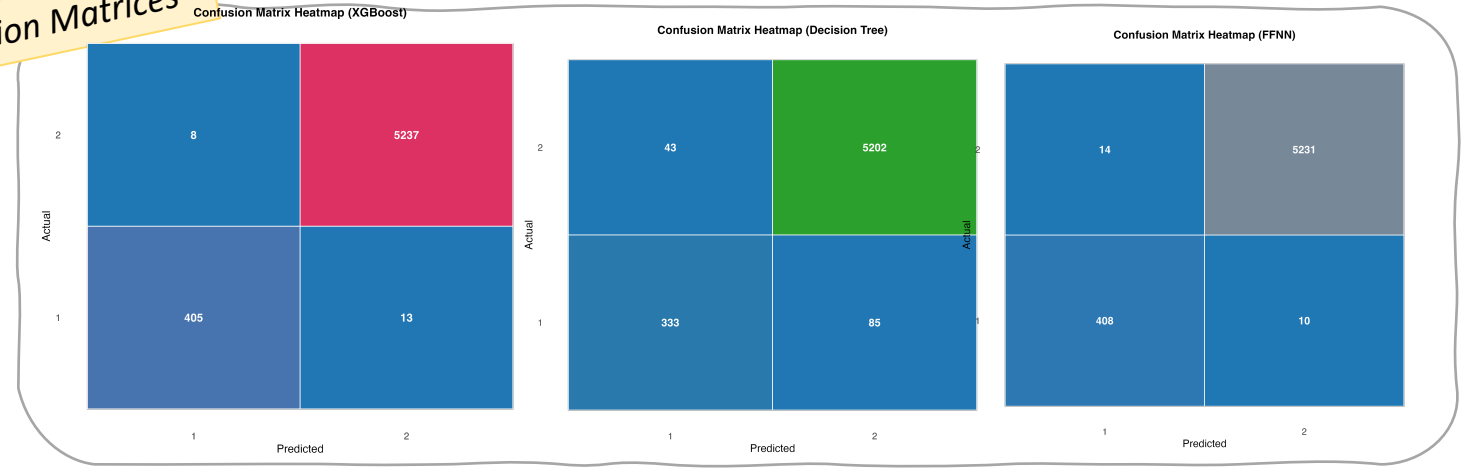
- XGboost
  - Log Loss showes convergence as the number of training rounds increases
  - The model generalizes better over time
  - Lower false negatives and false positives
- Decision Tree
  - test and training accuracies stabilize as the size grows
  - Decision trees can overfit on smaller datasets but perform more consistently with larger data
  - Higher false negatives and false positives
- Feed-forward NN
  - The accuracy shows fluctuations, likely due to the nonlinear training process of neural networks
  - Generally good for both class1 and class2
  - Acceptable false negatives and false positives

Learning Curves



XGboost | Decision Tree | Feed-forward NN

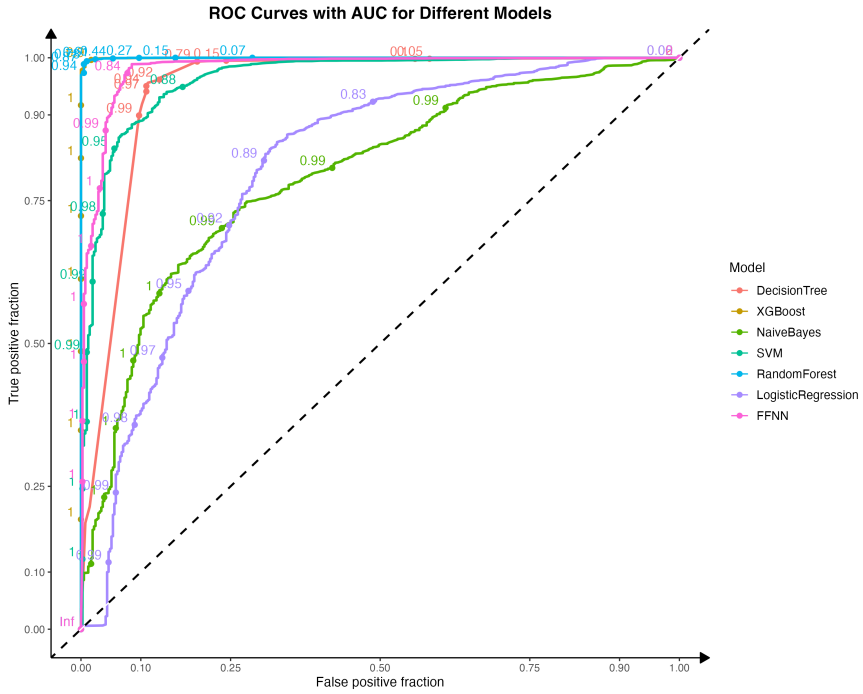
Confusion Matrices



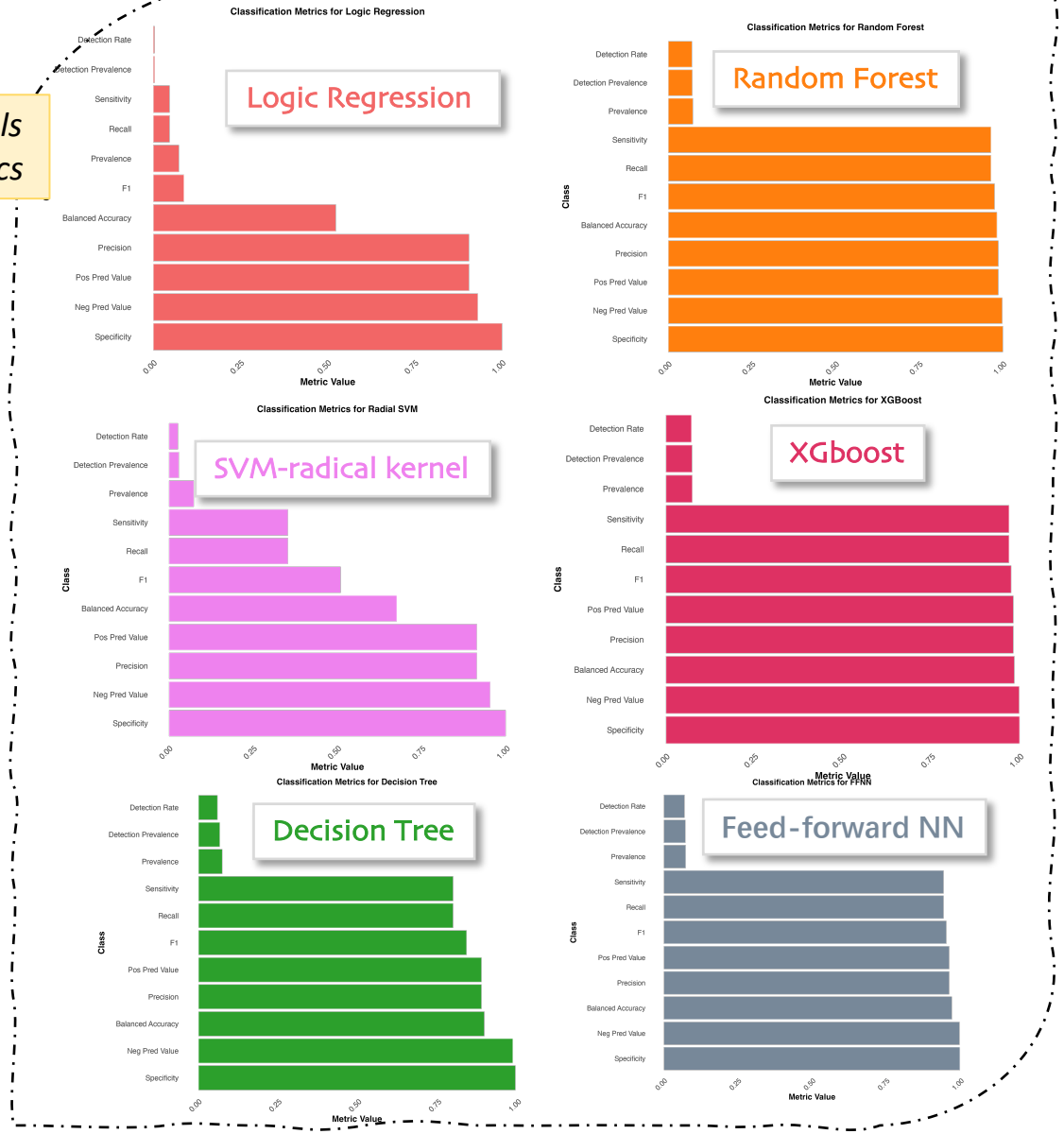
Binary classification – Model Performances

➔ **ROCs and metrics** used to evaluate models

ROC-AUCs



Models Metrics



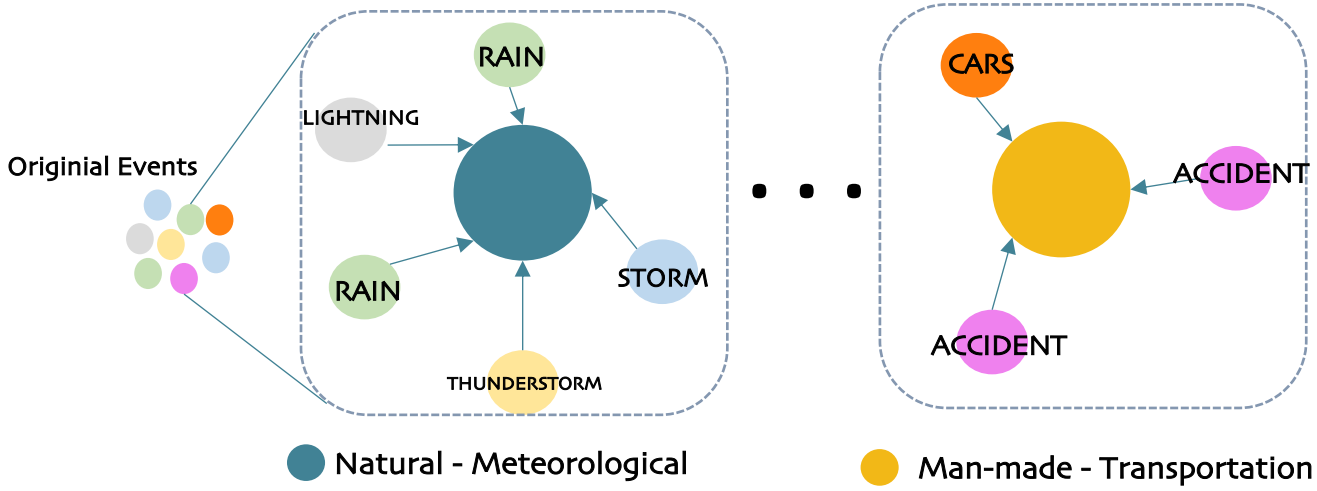
➔ Seven different models for the binary classification problem

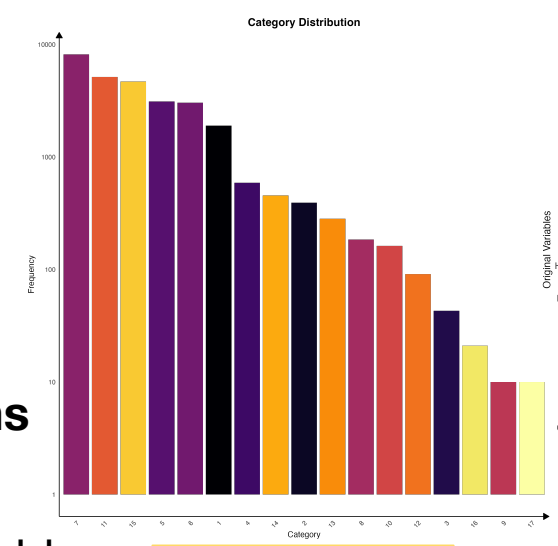
(including Naïve Bayes as a backup)

- Top-performing: Random Forest ~ XGboost
- Intermediate-performing models: FFNN > Decision Tree > SVM
- Lower-performing models: Logic Regression ~ Naïve Bayes

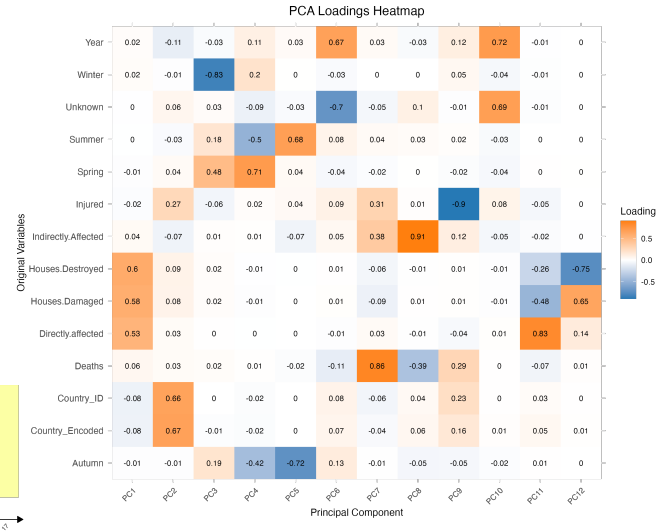
Multi-classification

- **Simplifying** Disaster Categorization
  - **Reason 1:** Different countries may use varied terms for the same type of disaster.
  - **Reason 2:** Insufficient data in some categories can hinder model performance.
- Oversampling technique applied **(SMOTE)**
  - **Reason:** extremely unbalanced in some classes
- **Various** algorithms are applied
  - Multi nom-logic
  - Random Forest
  - SVM - radical
  - FNN
  - KNN
  - XGBoost
  - Decision tree
- **Grid searches are applied for top-performing algorithms** **(back up slides)**
- **PCA** loading plots can reveal relationship with original variables





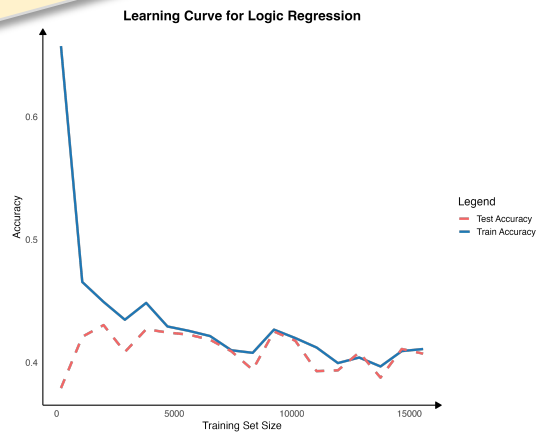
Class Frequency

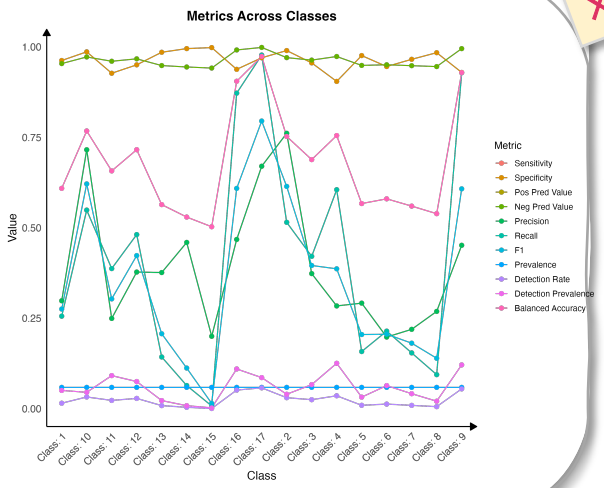


PCA Loading

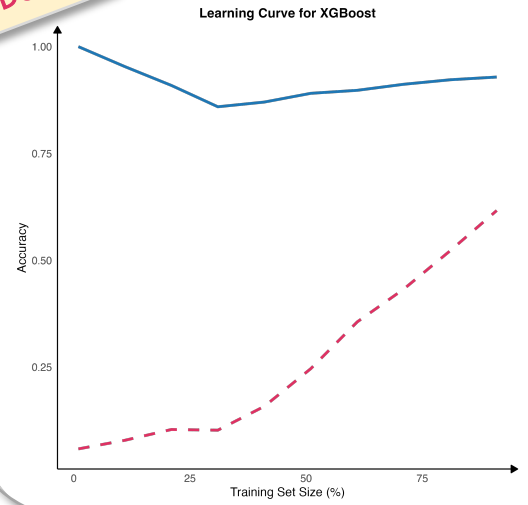
Multi-classification

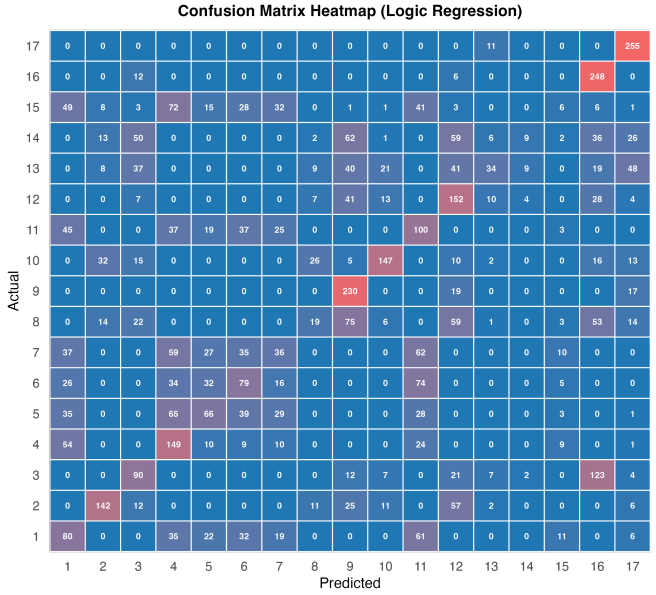
Logic Regression

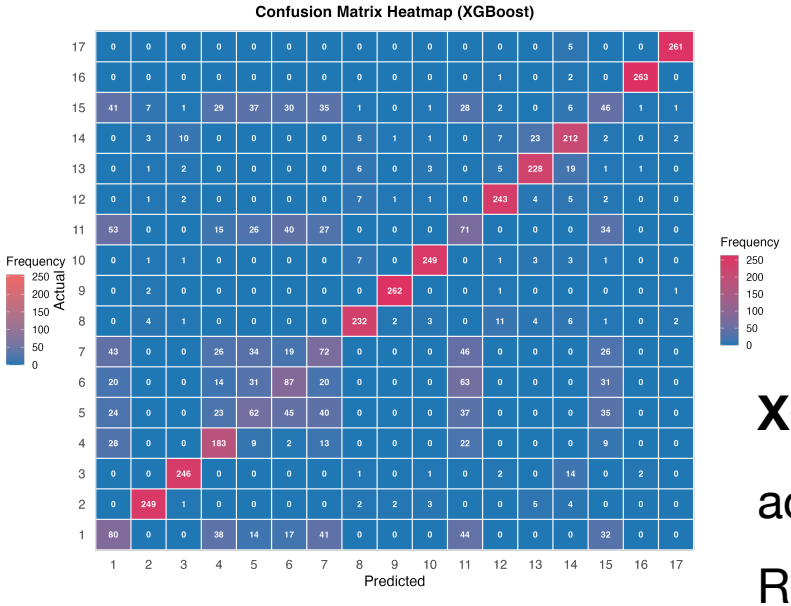




XGboost







- **Logic Regression**:

Good performance in Class ***16, 17, 9***

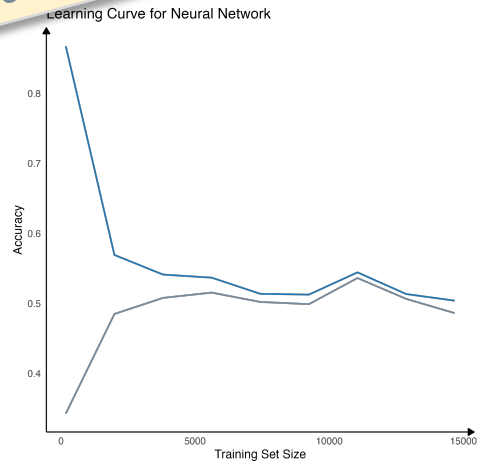
- **XGboost**:

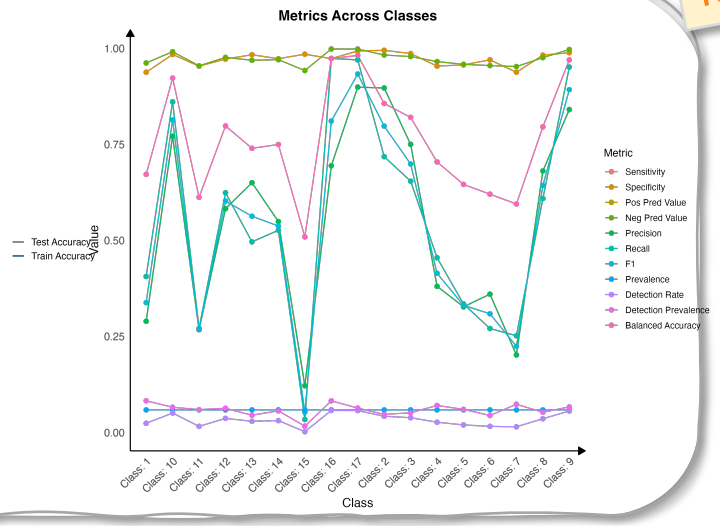
Good performance in Class ***10, 12, 13, 14, 16, 17, 2, 3, 8, 9***

XGboost shows more stable and has higher accuracy and F1 score compared to Logic Regression, but can be improved with more data.

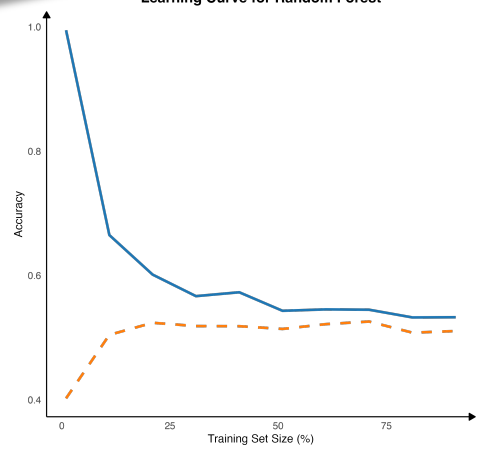
Multi-classification

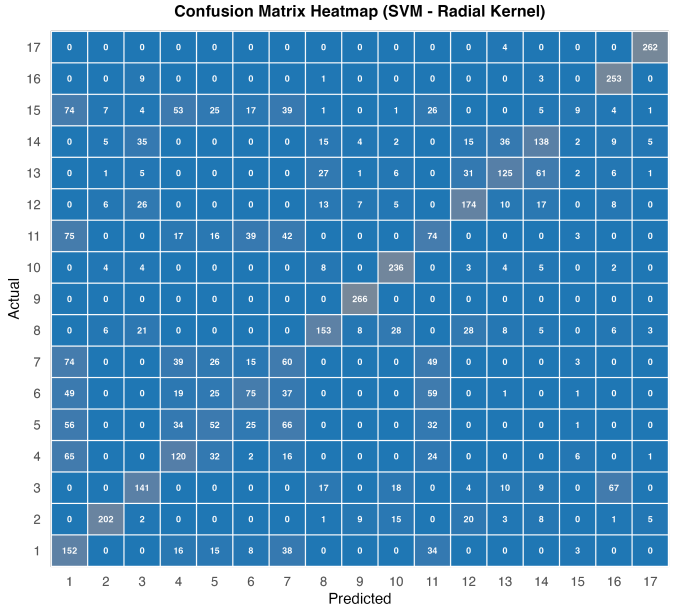
Feed-forward NN

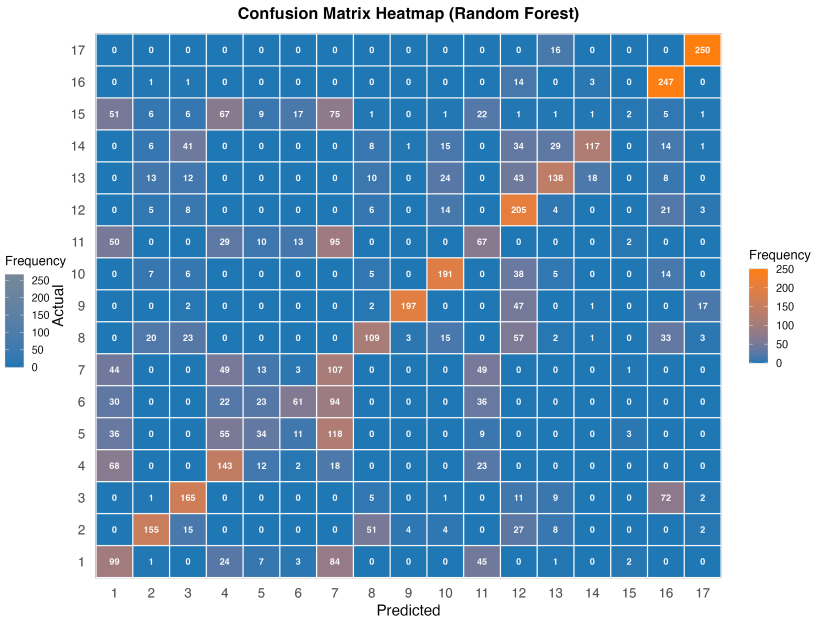




Random Forest







- **FNN:**

Good performance in Class ***10, 16, 17, 9***

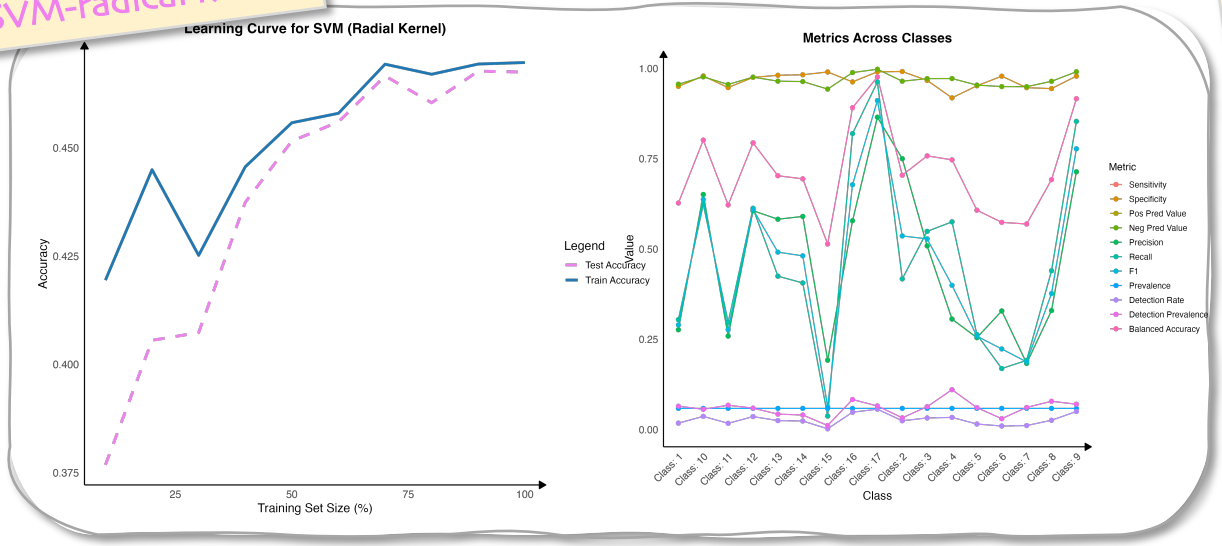
- **Random Forest:**

Good performance in Class ***16, 17, 9***

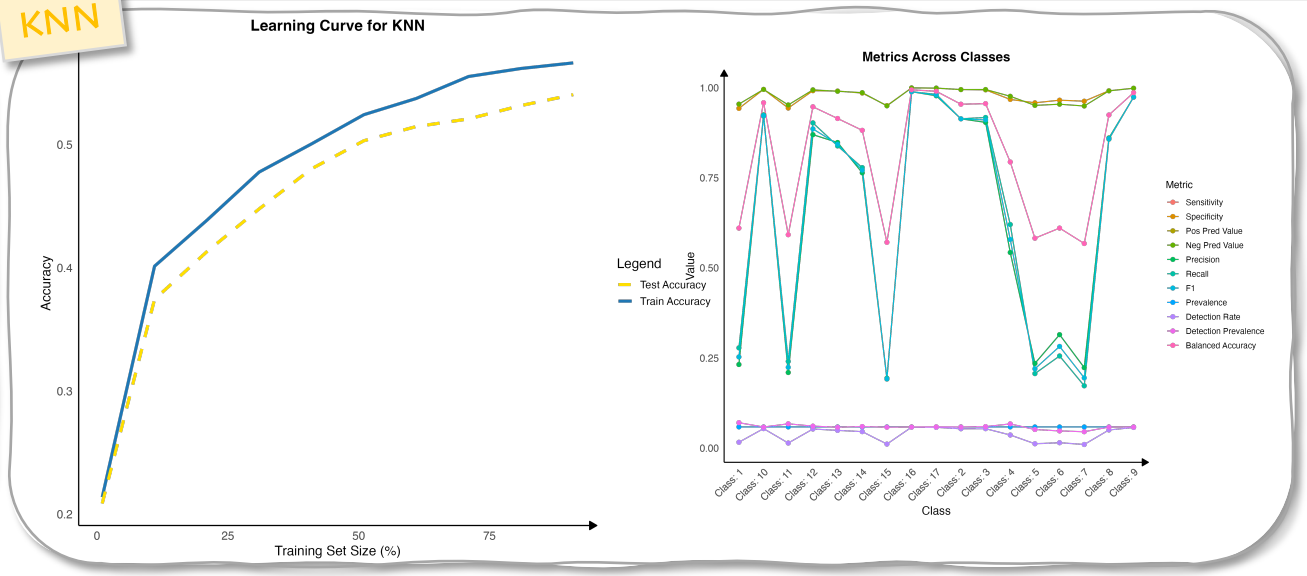
- *All learning curves converged indicating* ***no over-fitting*** *in two models*

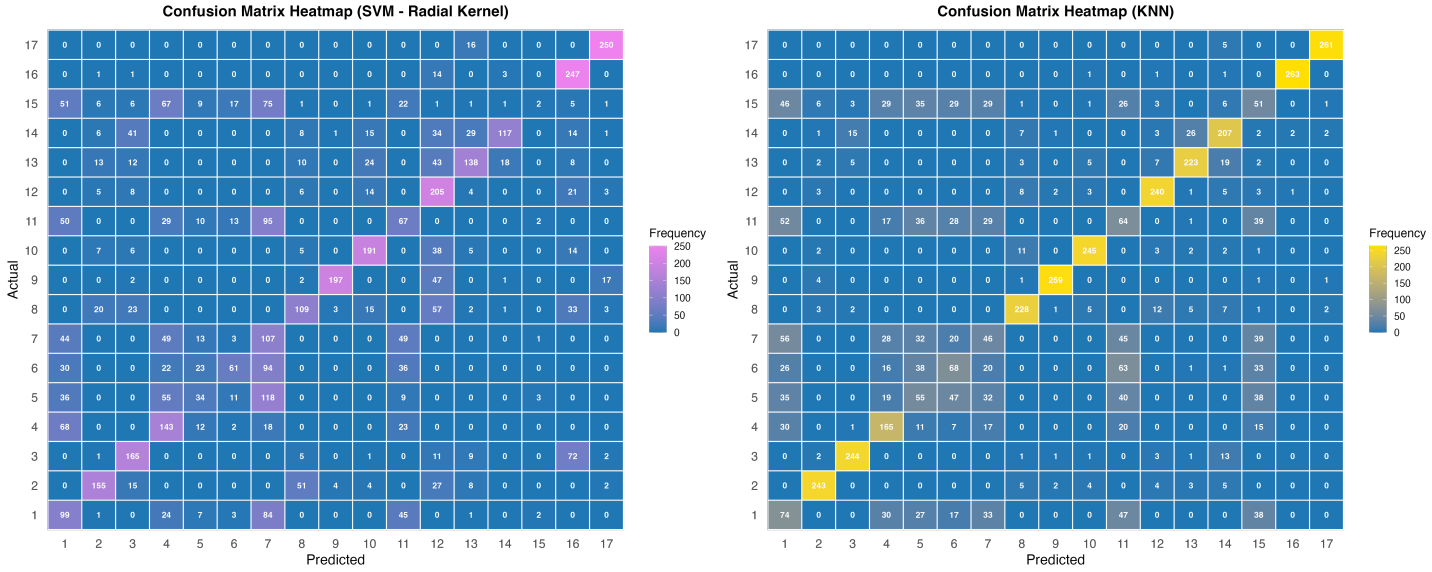
Multi-classification

SVM-radical kernel



KNN





- **SVM:**

Good performance in Class ***17, 9***

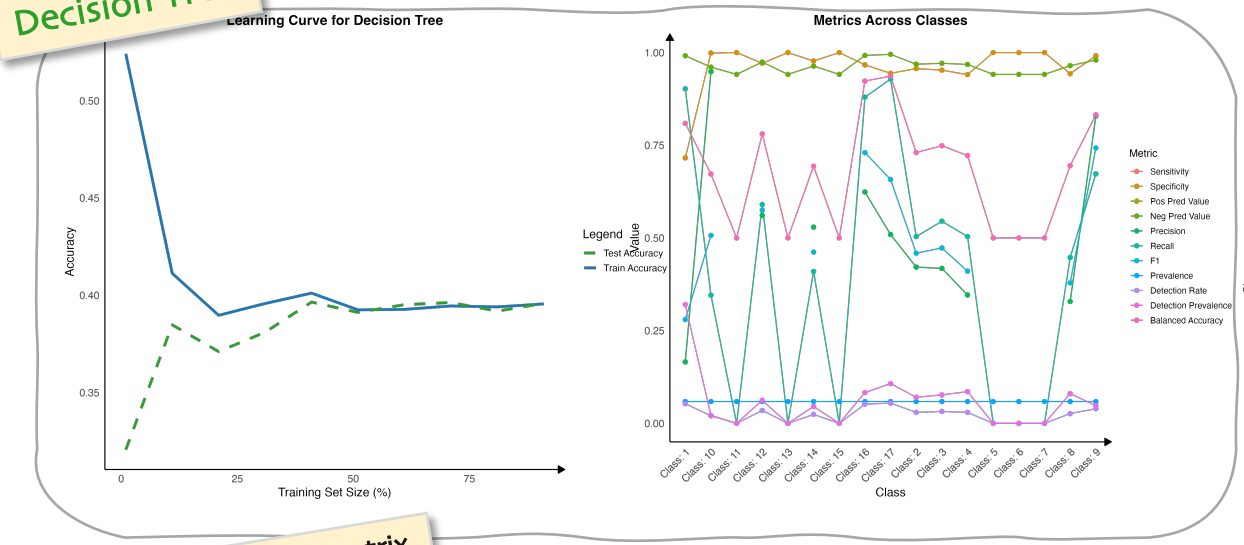
- **KNN:**

Good performance in Class ***10, 12, 13, 14, 16, 17, 2, 3, 8, 9***

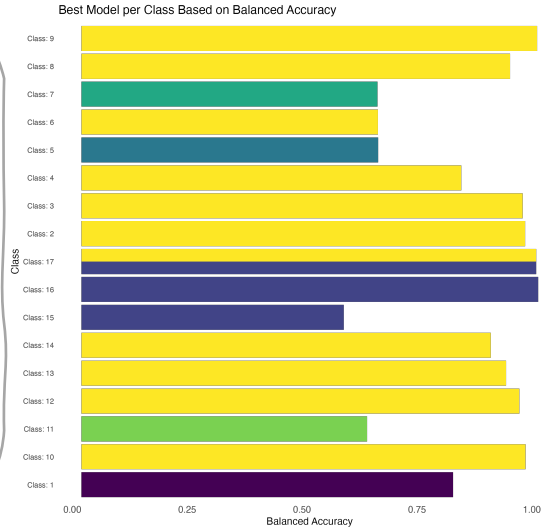
- KNN is more stable compared to SVM

Multi-classification

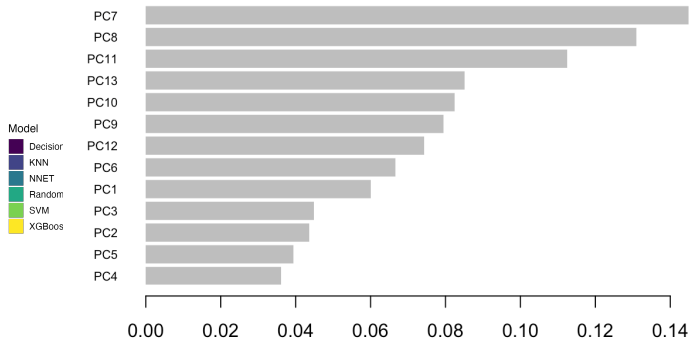
Decision Tree



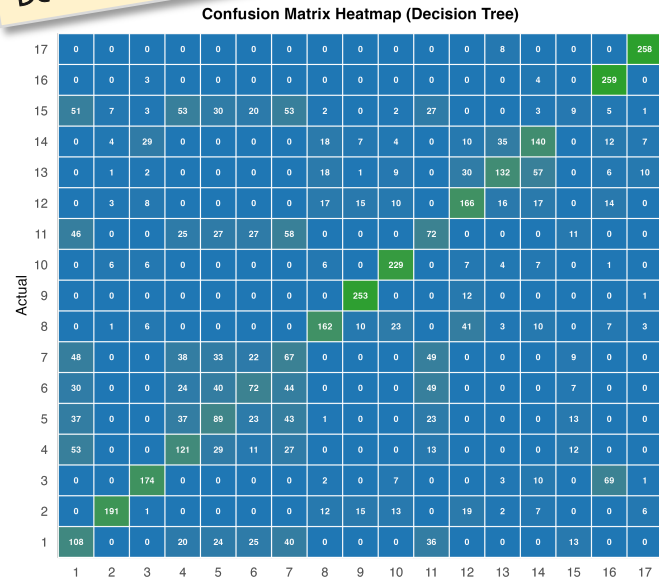
Highest Accuracy Models



Importance plot



Decision tree confusion matrix



Summary

- **With various algorithms, all classes accuracies can be larger than 50%**
- **XGBoost and KNN shows better performances**
- **XGboost is time consuming and possible overfited**
- ***Different classes are suitable for different algorithms***

Summary

- **XGBoost** is a good choice for *class **2,3,4,6,8,9,10,12,14,17***
- **KNN** performs well for *class **15, 16, 17***
- **SVM** works best for *class **11***
- **Random Forest** is suitable for *class **7***
- **FNN** is suitable for *class **5***
- **Decision Tree** is suitable for *class **1***