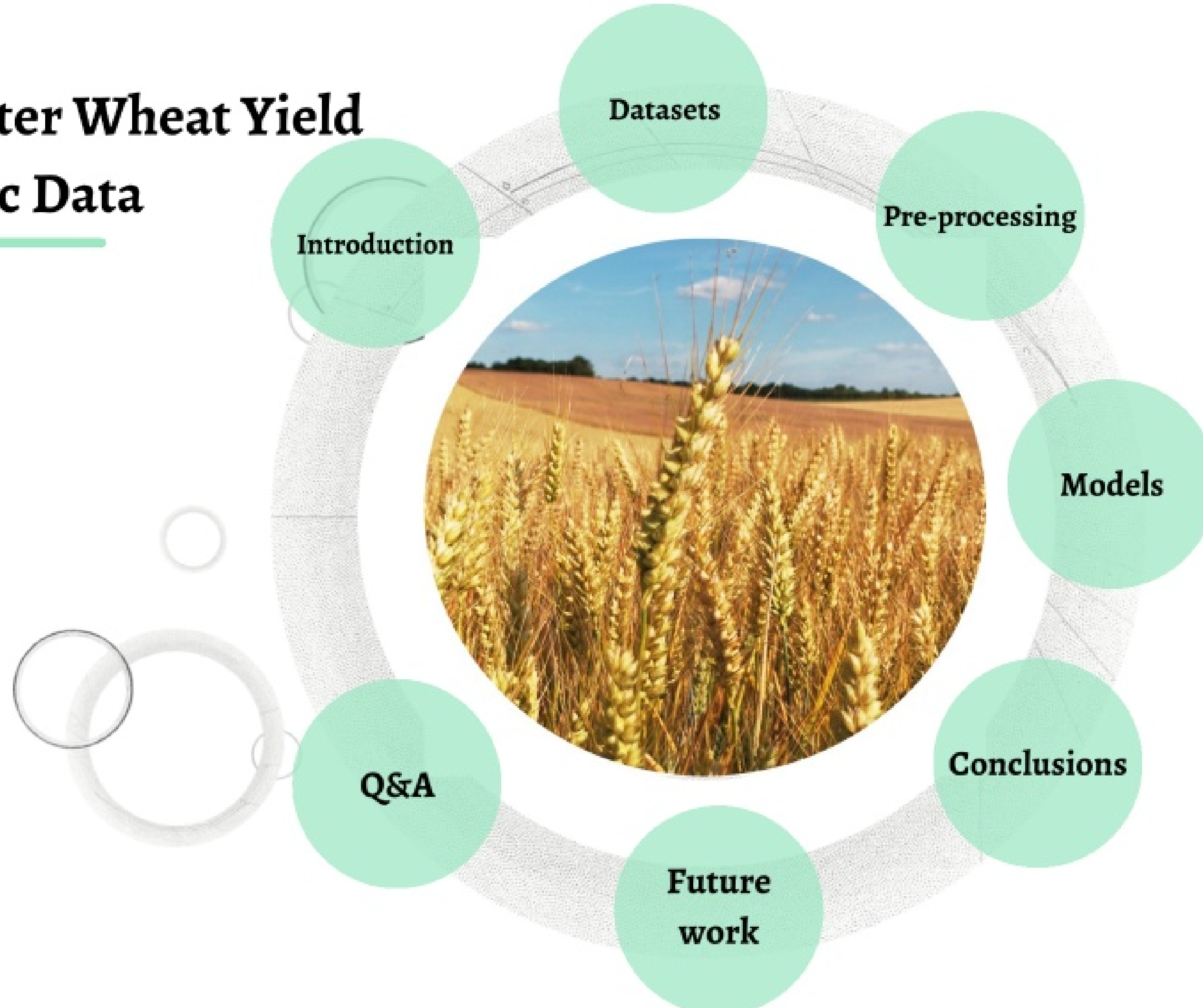
Prediction of Winter Wheat Yield Loss using Climatic Data

Maria Gil Rodriguez



Introduction

- Annual Winter Wheat Yield : Amount of grain harvested by unit of area in a given year (in tonnes per hectare)

- Depends on the characteristics of the region and the climatic conditions. Values vary greatly between regions and years

- Important to accurately predict yield loss.
  - Harvest planning
  - Management of stocks
  - Strategic information in international markets.

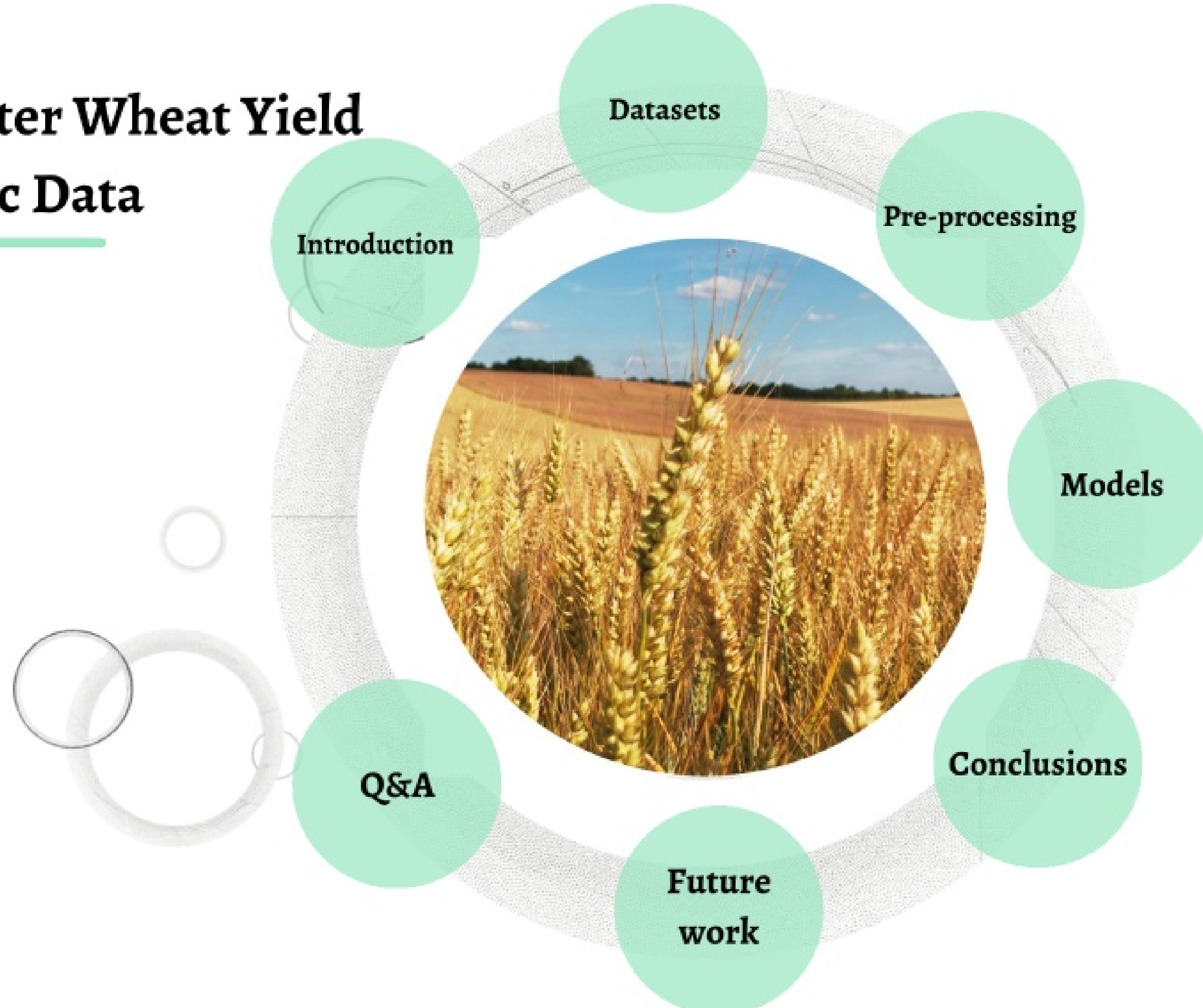
Objective

Objective

The objective of this capstone is to develop tools to classify as accurately as possible the wheat yield loss in France.

Prediction of Winter Wheat Yield Loss using Climatic Data

Maria Gil Rodriguez



Datasets

CLAND Challenge
- Training set
- Blind test set

Metadata

Distributions

Correlations

Datasets

CLAND Challenge

- Training set
- Blind test set

Metadata

Distributions

Correlations

Metadata

- 94 Departments
- 58 years
- Climatic data: months Sep-Jun
  - Potential Evapotranspiration (mm/day)
  - Solar Radiation (W/m^2)
  - Precipitation: monthly values (mm/day), # rainy days
  - Temperatures: Max (C), min (C), # days with extreme values
- Yield loss: 1 = loss 0 = no loss

Metadata

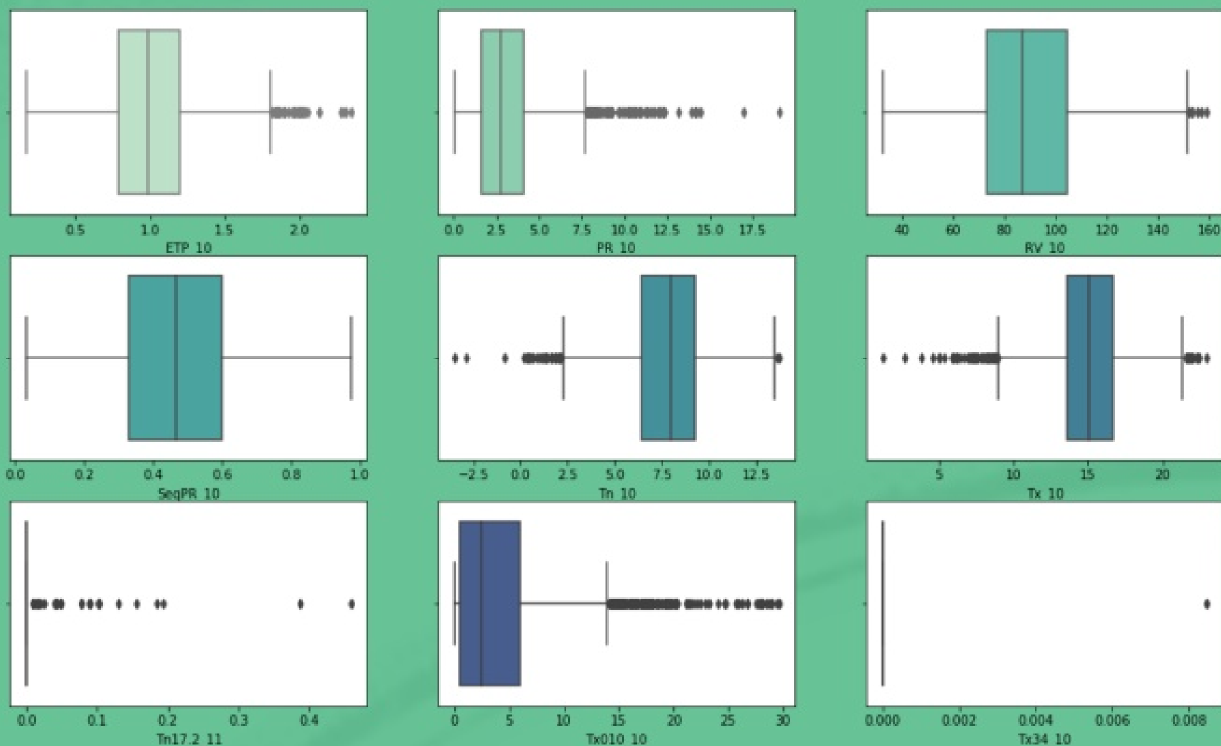
Metadata

For each year and Department:

- Potential Evapotranspiration (mm/day):
  ETP_9, ETP_10, ETP_11, ETP_12, ETP_1, ..., ETP_6
- Precipitation: monthly values (mm/day) and # rainy days:
  PR_9, PR_10, PR_11, PR_12, PR_1, ..., PR_6
  SeqPR_9, SeqPR_10, SeqPR_11, ..., SeqPR_6
- Solar Radiation (W/m2):
  RV_9, RV_10, RV_11, RV_12, RV_1, ..., RV_6
- Temperatures: Max (C), min (C), # days with extreme values
  Tx_9, Tx_10, Tx_11, Tx_12, Tx_1, ..., Tx_6
  Tn_9, Tn_10, Tn_11, Tn_12, Tn_1, ..., Tn_6
  Tx34_9, Tx34_10, Tx34_11, ..., Tx34_6 — # days with daily maximum T > 34 C
  Tx010_9, Tx010_10, Tx010_11, ..., Tx010_6 — # days with daily maximum T between 0 and 10 C
  Tn17.2_9, Tn17.2_10, Tn17.2_11, ..., Tn17.2_6 — # days with daily minimum T < -17 C

Distributions

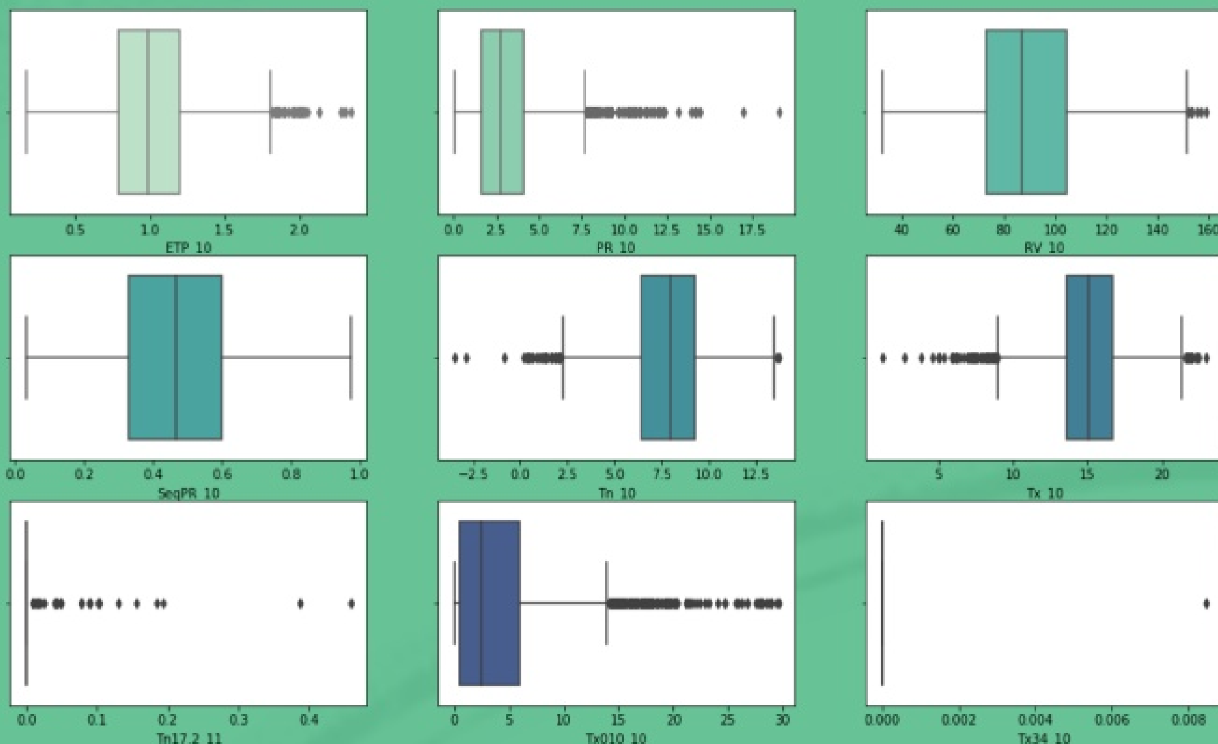
Boxplots of climatic variables in October (except for Tn17.2, which corresponds to November)



Noise,
noise,
noise...

Distributions

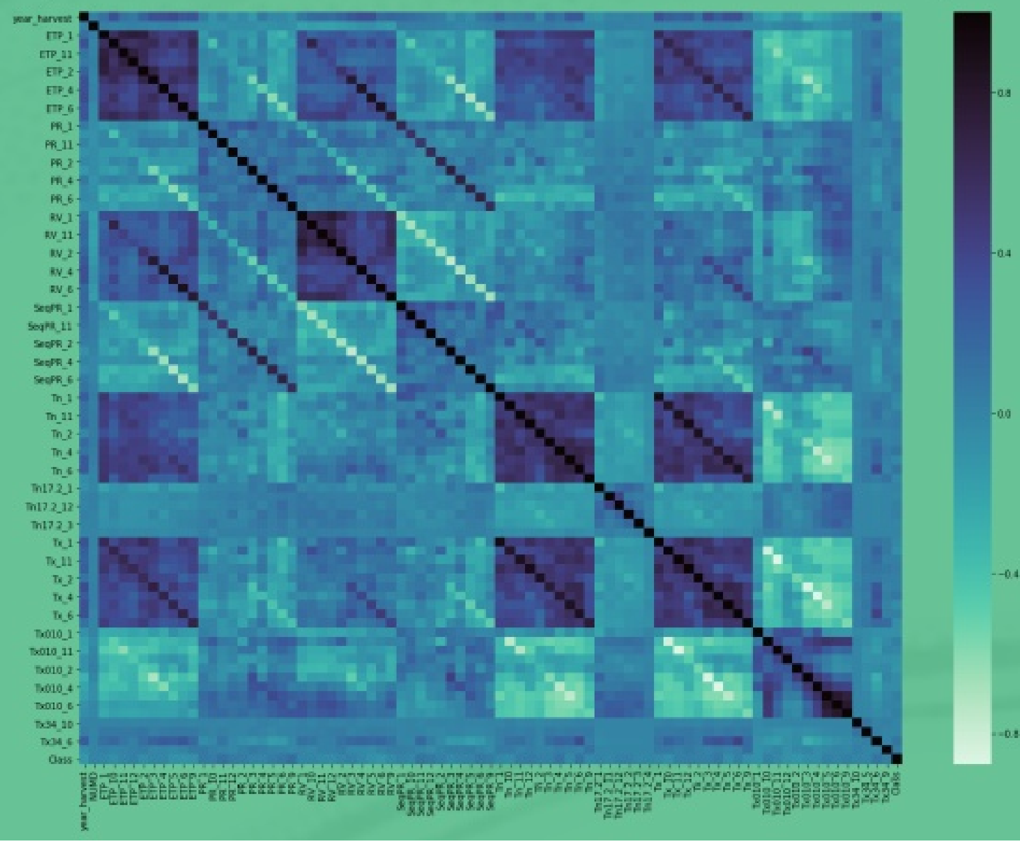
Boxplots of climatic variables in October (except for Tn17.2, which corresponds to November)



Noise, noise, noise...

Can't remove outliers!

Correlations



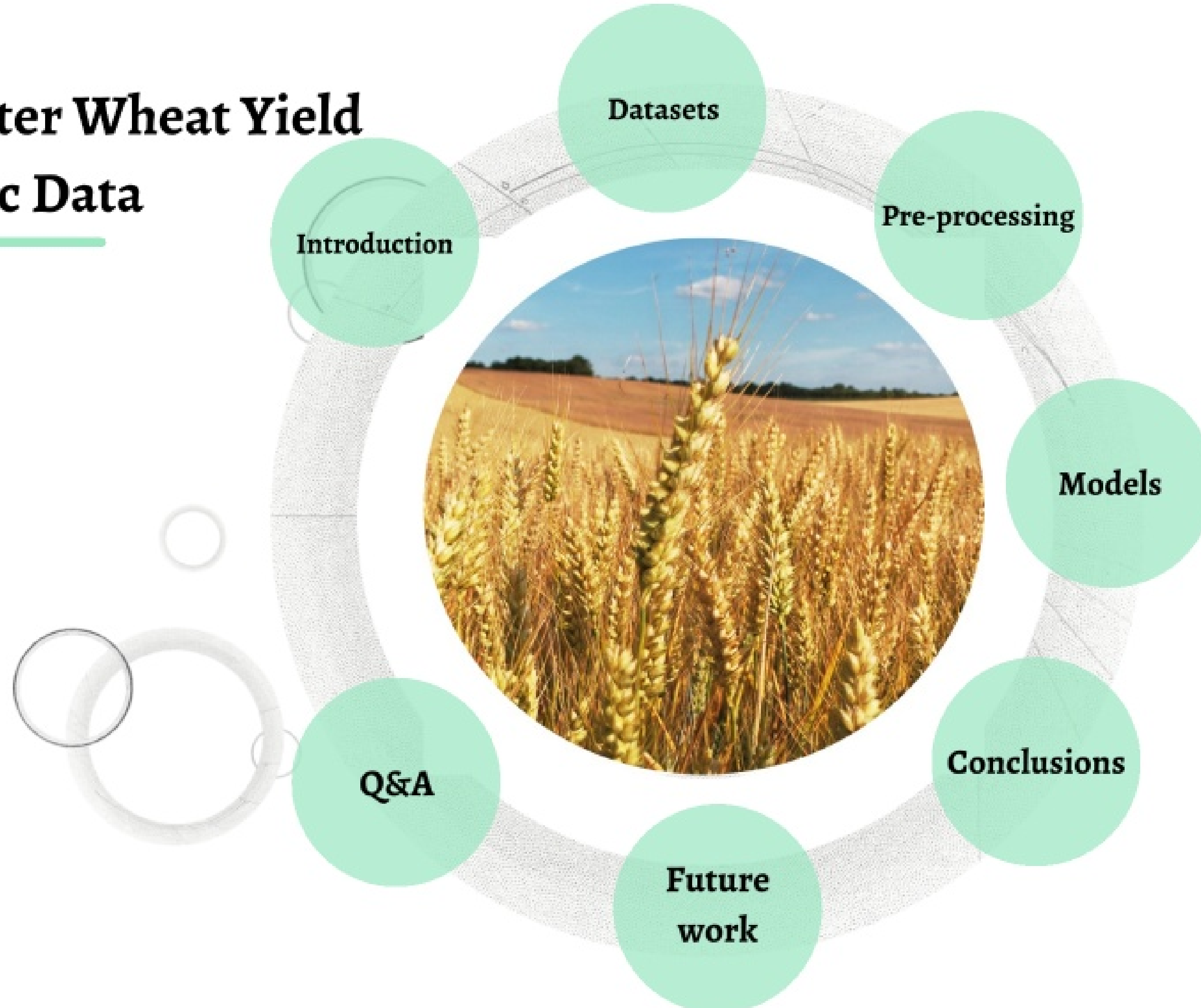
Variables for one month: strong positive and negative correlations

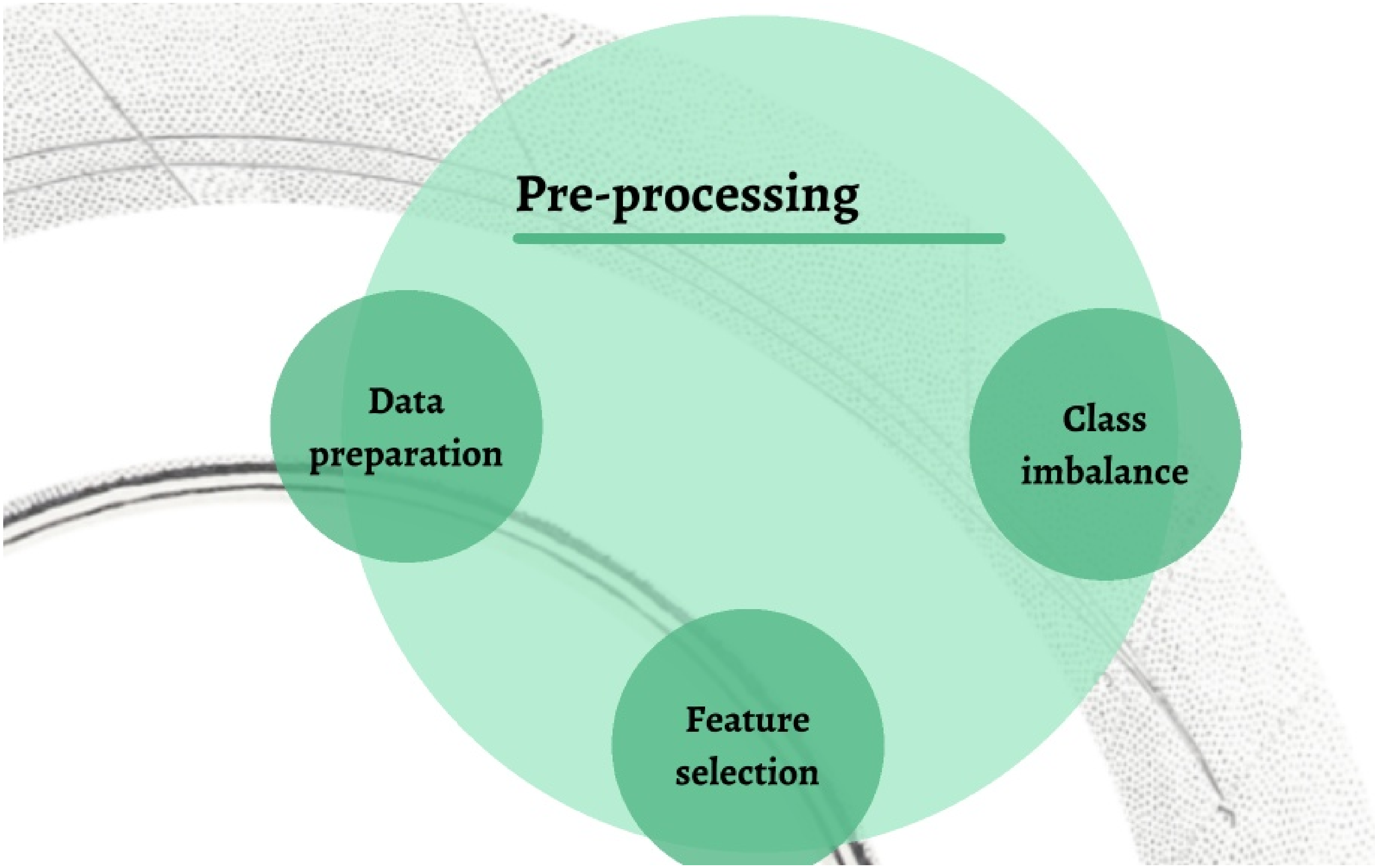
↓

Difficult to drop without losing information

Prediction of Winter Wheat Yield Loss using Climatic Data

Maria Gil Rodriguez




Pre-processing
Data preparation
Class imbalance
Feature selection

Data preparation

- **Data Cleaning**

  Delete columns of straight zeros

  ↓

  80 features and 3571 instances

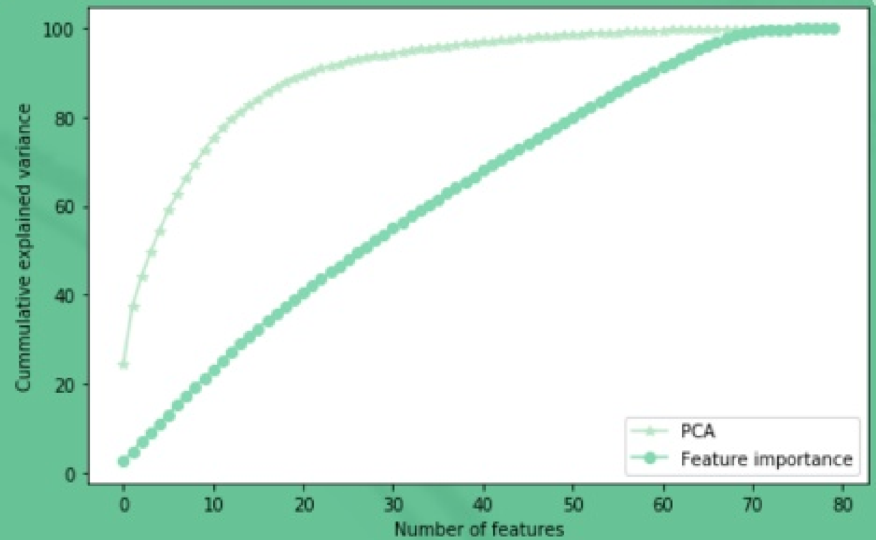
- **Splitting**

  Random: 75% train - 25% test

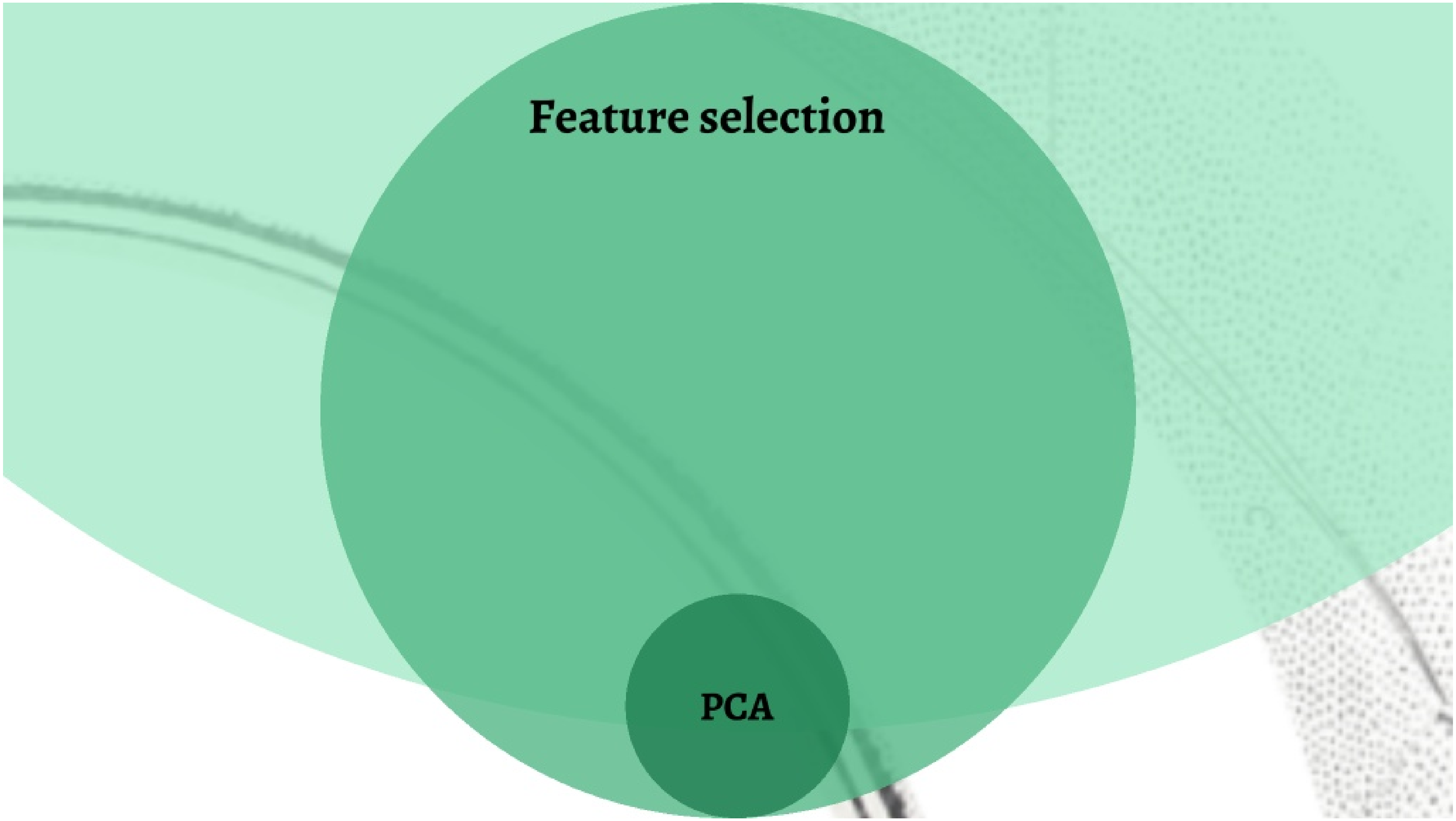
  (Stratified splitting didn't work well)

- **Normalization**

Feature selection

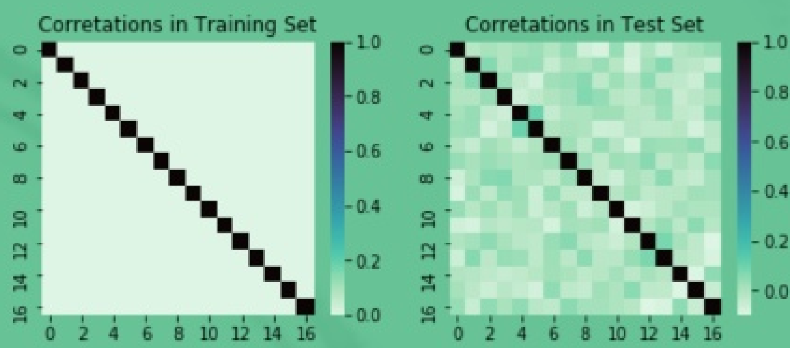


PCA


Feature selection
PCA

Feature selection

After PCA:



PCA

Why does PCA overperform?

Usually:

It is recommended to remove highly correlated variables before PCA

↓

Correlated variables point in the same direction making that component stronger

In our case:

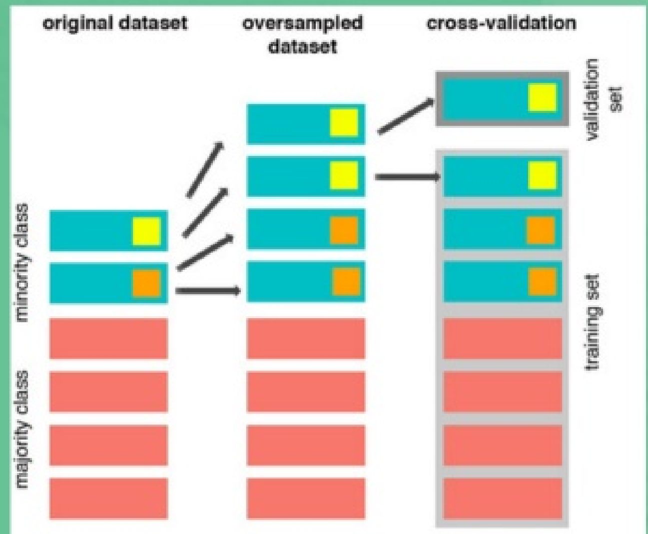
We have roughly the same number of variables for each month

Class Imbalance

Oversample with SMOTE while using GridSearchCV

SMOTE + cross validation: Pipeline

Wrong



Source: https://www.marcoaltini.com/

Class Imbalance

Oversample with SMOTE while using GridSearchCV

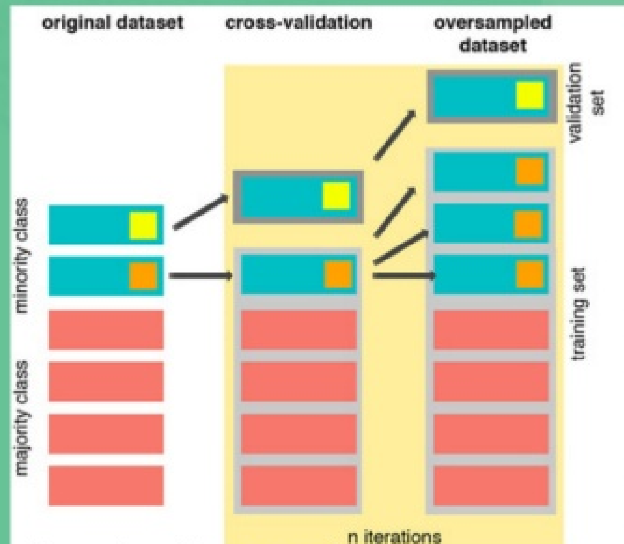
SMOTE + cross validation: Pipeline

Class Imbalance

Oversample with SMOTE while using GridSearchCV

SMOTE + cross validation: Pipeline

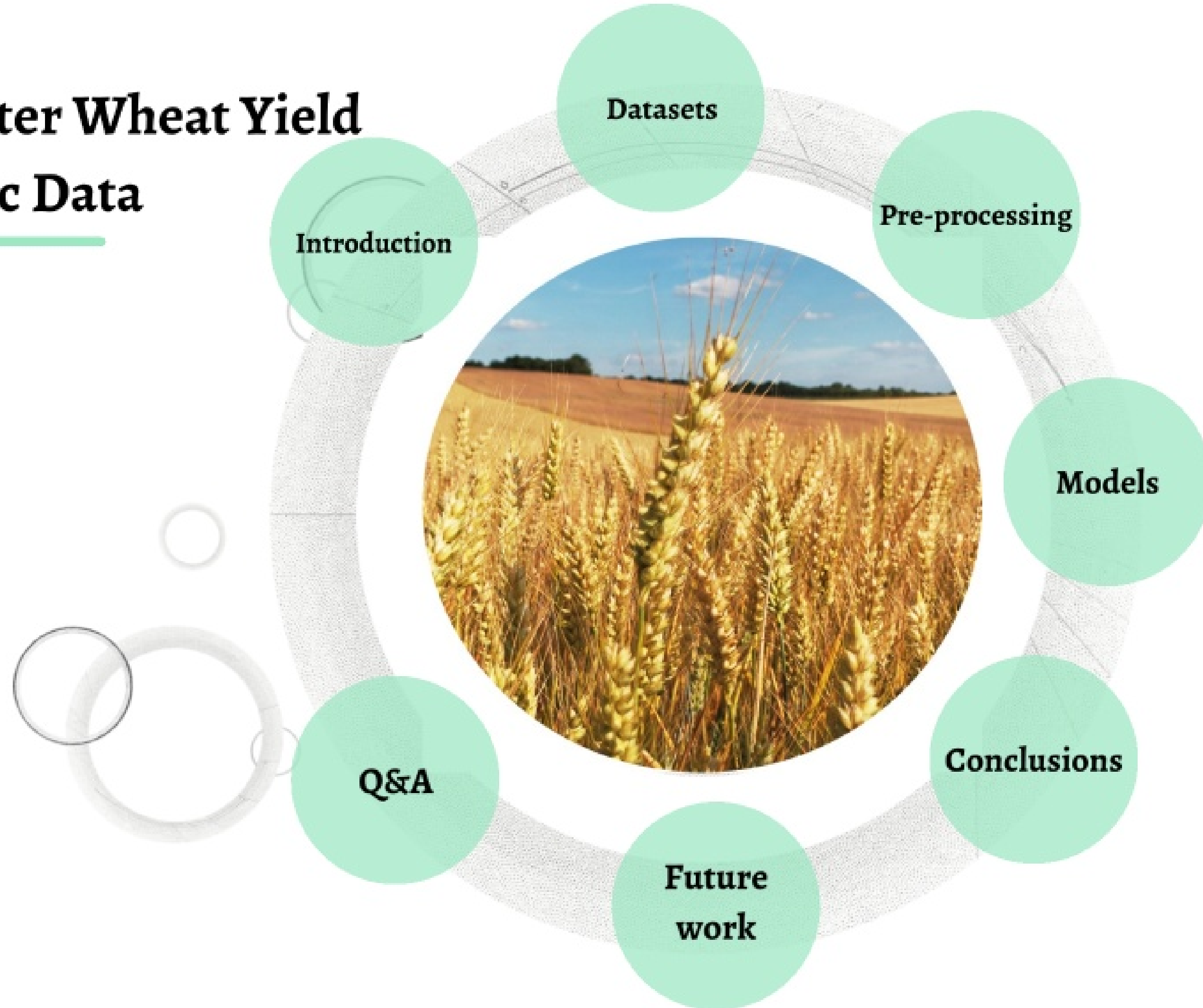
Right



Source: https://www.marcoaltini.com/

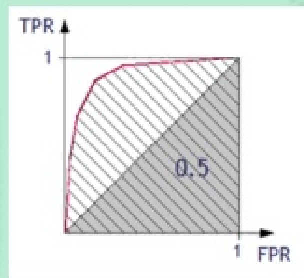
Prediction of Winter Wheat Yield Loss using Climatic Data

Maria Gil Rodriguez



Models

Metric: Area under the ROC curve



Logistic Regression

SVC

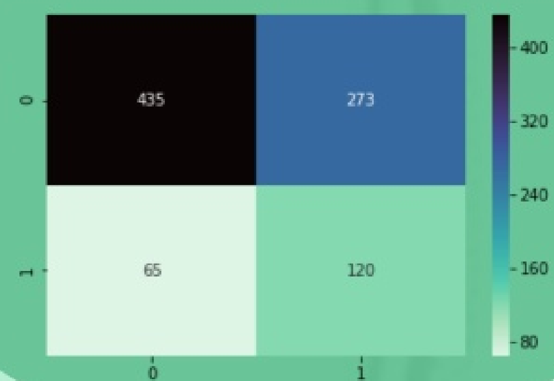
KNN

Gradient Boosting

Random Forest

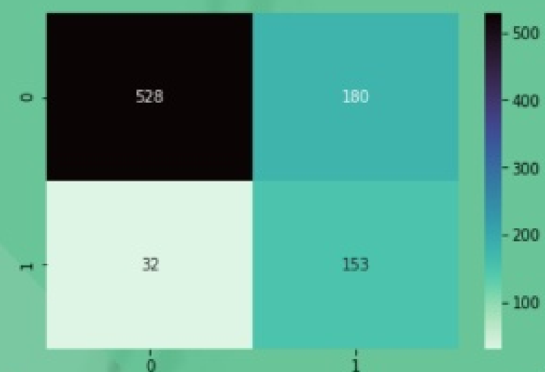
Logistic Regression

```
Train set score:                 0.68
Best cross validation score:     0.66
Test set score:                  0.66
Report:
              precision    recall  f1-score   support

           0       0.87      0.61      0.72       708
           1       0.31      0.65      0.42       185

   micro avg       0.62      0.62      0.62       893
   macro avg       0.59      0.63      0.57       893
weighted avg       0.75      0.62      0.66       893
```



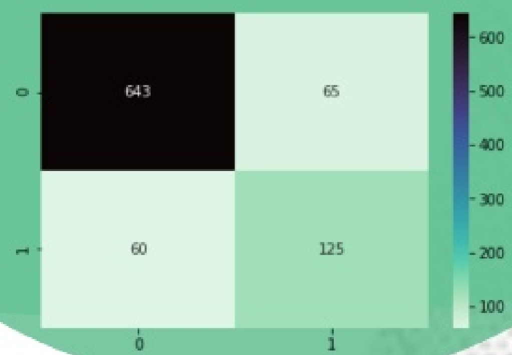
KNN

```
Train set score:              0.92
Best cross validation score:  0.86
Test set score:               0.88
Report:
              precision    recall  f1-score   support

           0       0.94      0.75      0.83       708
           1       0.46      0.83      0.59       185

   micro avg       0.76      0.76      0.76       893
   macro avg       0.70      0.79      0.71       893
weighted avg       0.84      0.76      0.78       893
```



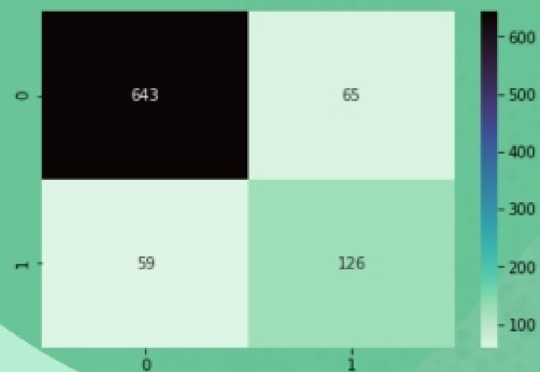
Random Forest

```
Train set score:                 1.0
Best cross validation score:     0.88
Test set score:                  0.89
Report:
              precision    recall  f1-score   support

           0       0.91      0.91      0.91       708
           1       0.66      0.68      0.67       185

   micro avg       0.86      0.86      0.86       893
   macro avg       0.79      0.79      0.79       893
weighted avg       0.86      0.86      0.86       893
```



Gradient Boosting

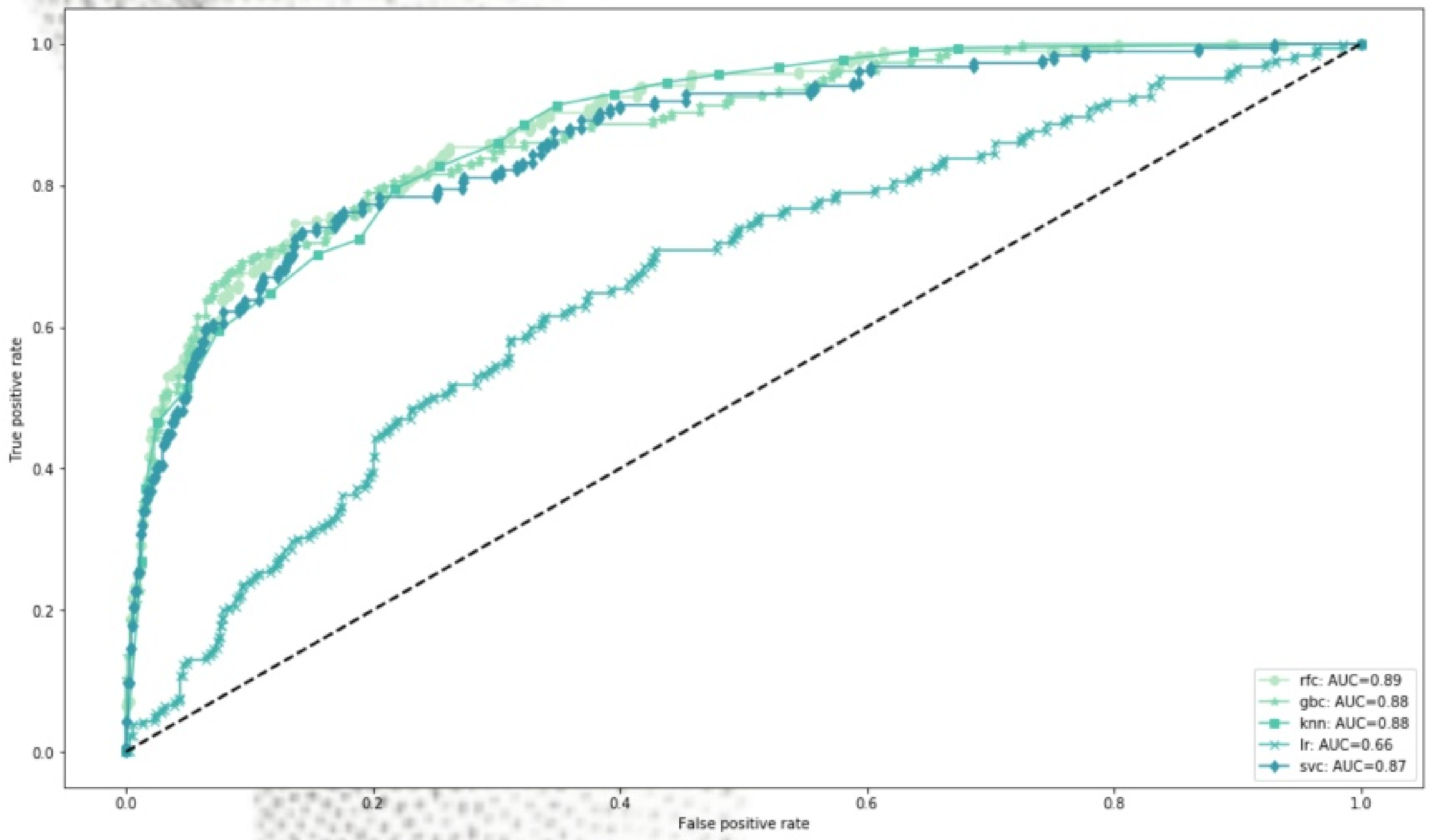
```
Train set score:                  1.0
Best cross validation score:      0.87
Test set score:                   0.88
Report:
              precision    recall  f1-score   support

           0       0.92      0.91      0.91       708
           1       0.66      0.68      0.67       185

   micro avg       0.86      0.86      0.86       893
   macro avg       0.79      0.79      0.79       893
weighted avg       0.86      0.86      0.86       893
```



SVC

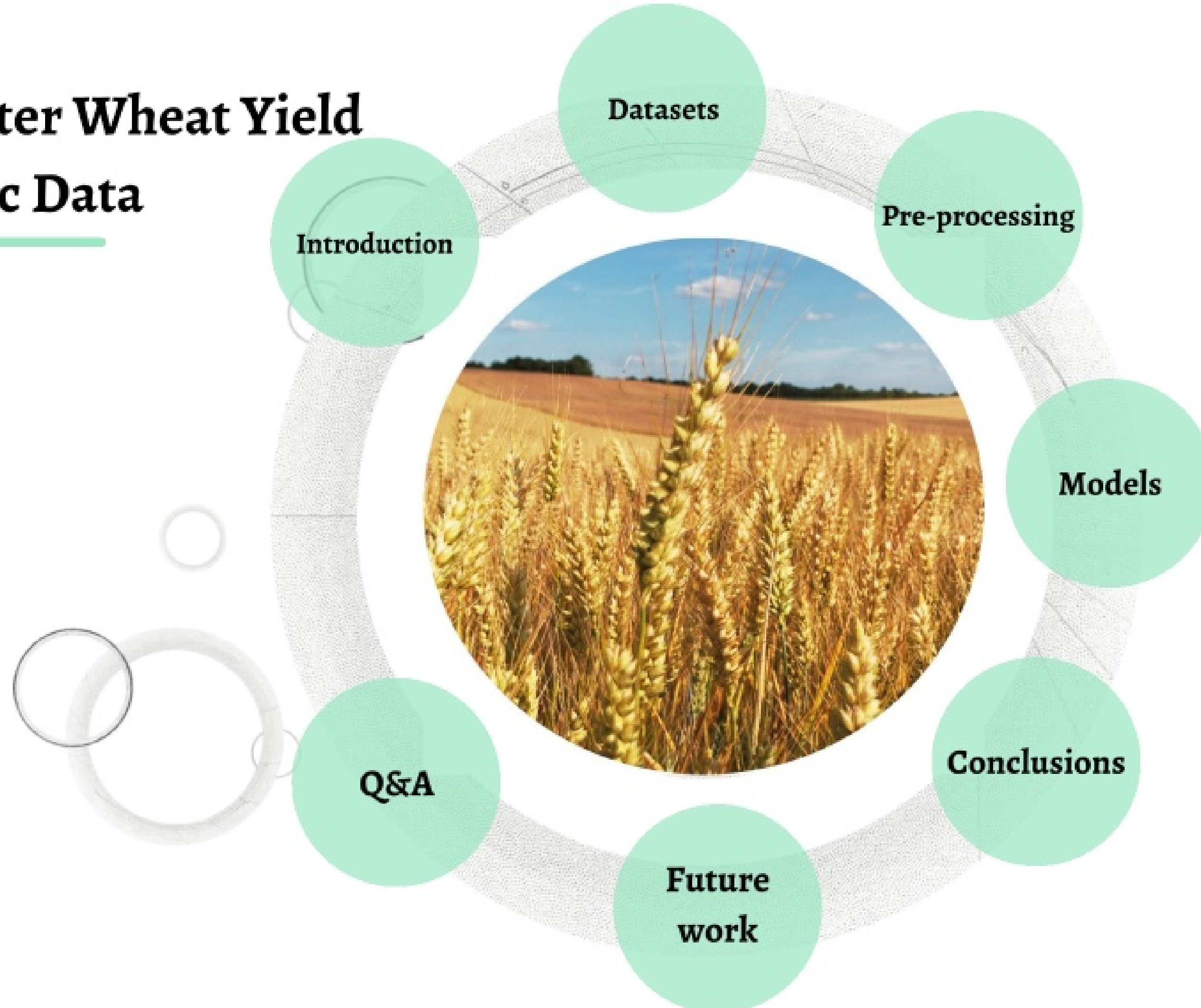
```
Train set score:                  0.96
Best cross validation score:      0.85
Test set score:                   0.87
Report:
              precision    recall  f1-score   support

           0       0.93      0.84      0.88       708
           1       0.54      0.74      0.63       185

   micro avg       0.82      0.82      0.82       893
   macro avg       0.73      0.79      0.75       893
weighted avg       0.85      0.82      0.83       893
```

Model Comparison: ROC Curves



Prediction of Winter Wheat Yield Loss using Climatic Data

Maria Gil Rodriguez



Conclusions

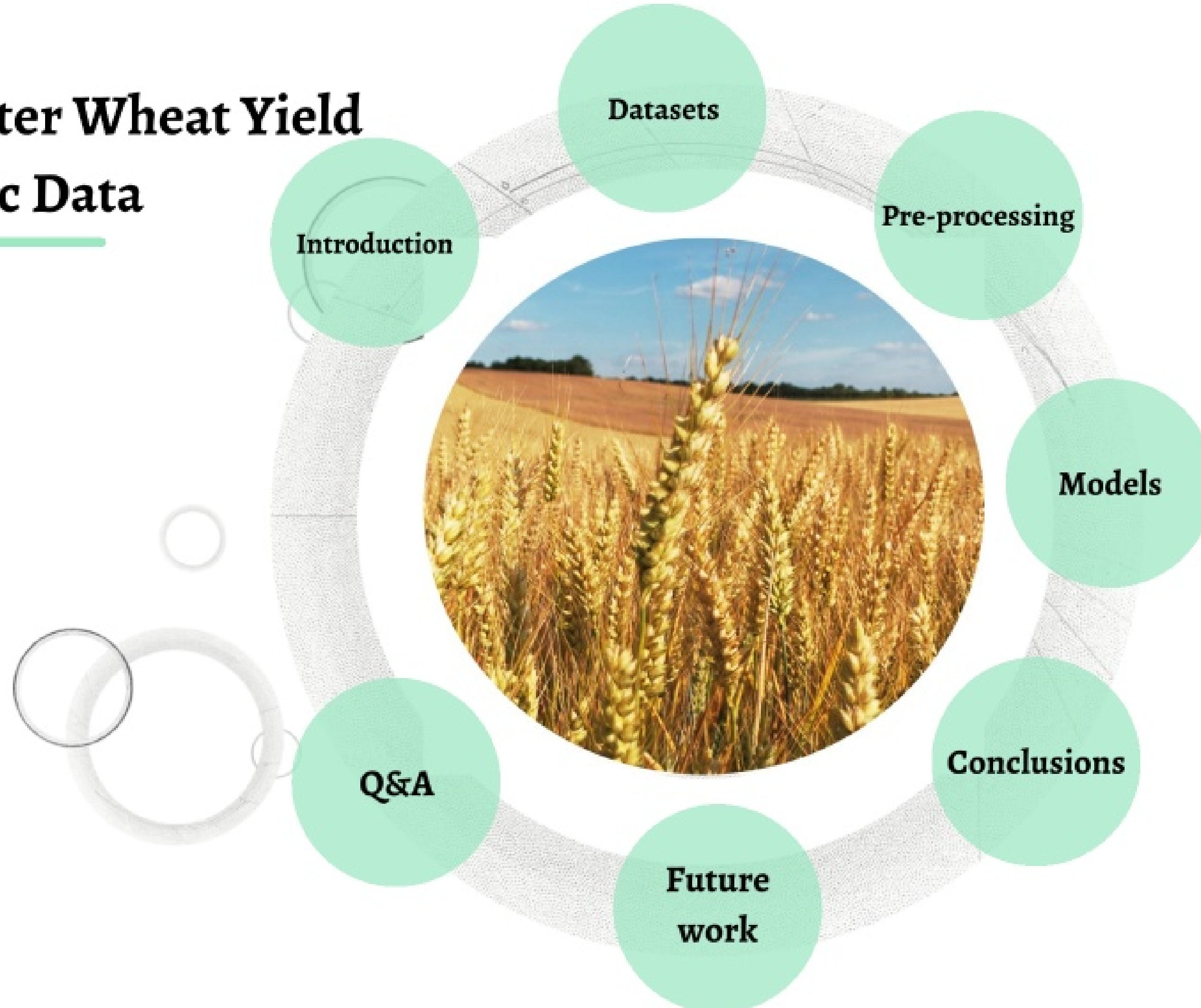
- Climate variables are relatively good predictors of wheat yield loss in France. The predictions are valuable in order to plan harvests, manage stocks, optimize contracts and operate in international markets.

- Random Forest was the best model, with an area under the ROC curve of 0.89.

- The results would be much better with a less noisy data. That could be achieved by working with local data (vs generalized for an entire Department) or using the data of several stations for one Department.

Prediction of Winter Wheat Yield Loss using Climatic Data

Maria Gil Rodriguez



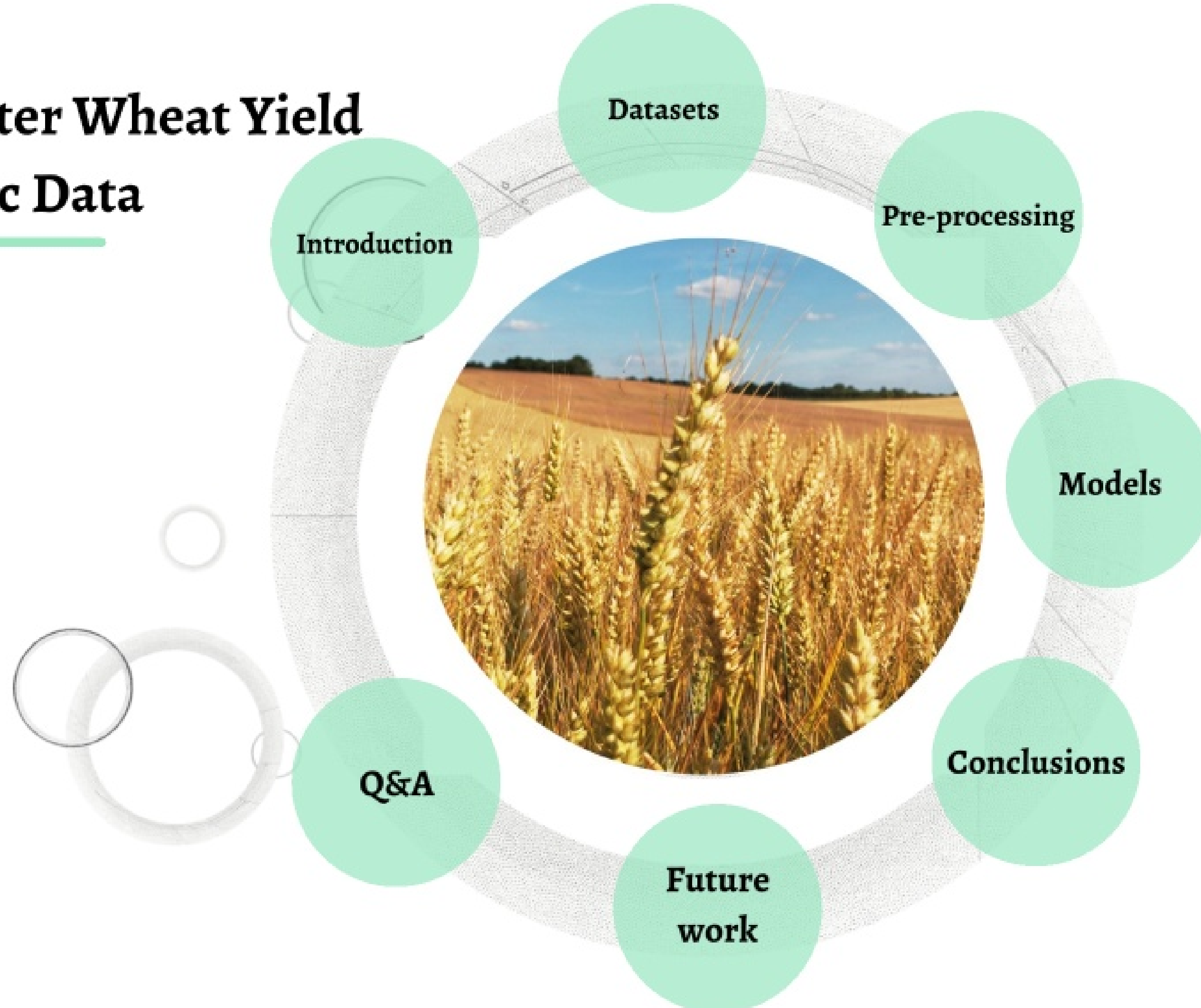
Future work

- France has 5 different climates. Thus, performing some clustering before using our ML models would be ideal and most likely improve the results.

- There is also some information in the NUMD (number of department) variable that might be worth to explore.

Prediction of Winter Wheat Yield Loss using Climatic Data

Maria Gil Rodriguez




Q & A

Prediction of Winter Wheat Yield Loss using Climatic Data

Maria Gil Rodriguez