MACHINE LEARNING REPORT

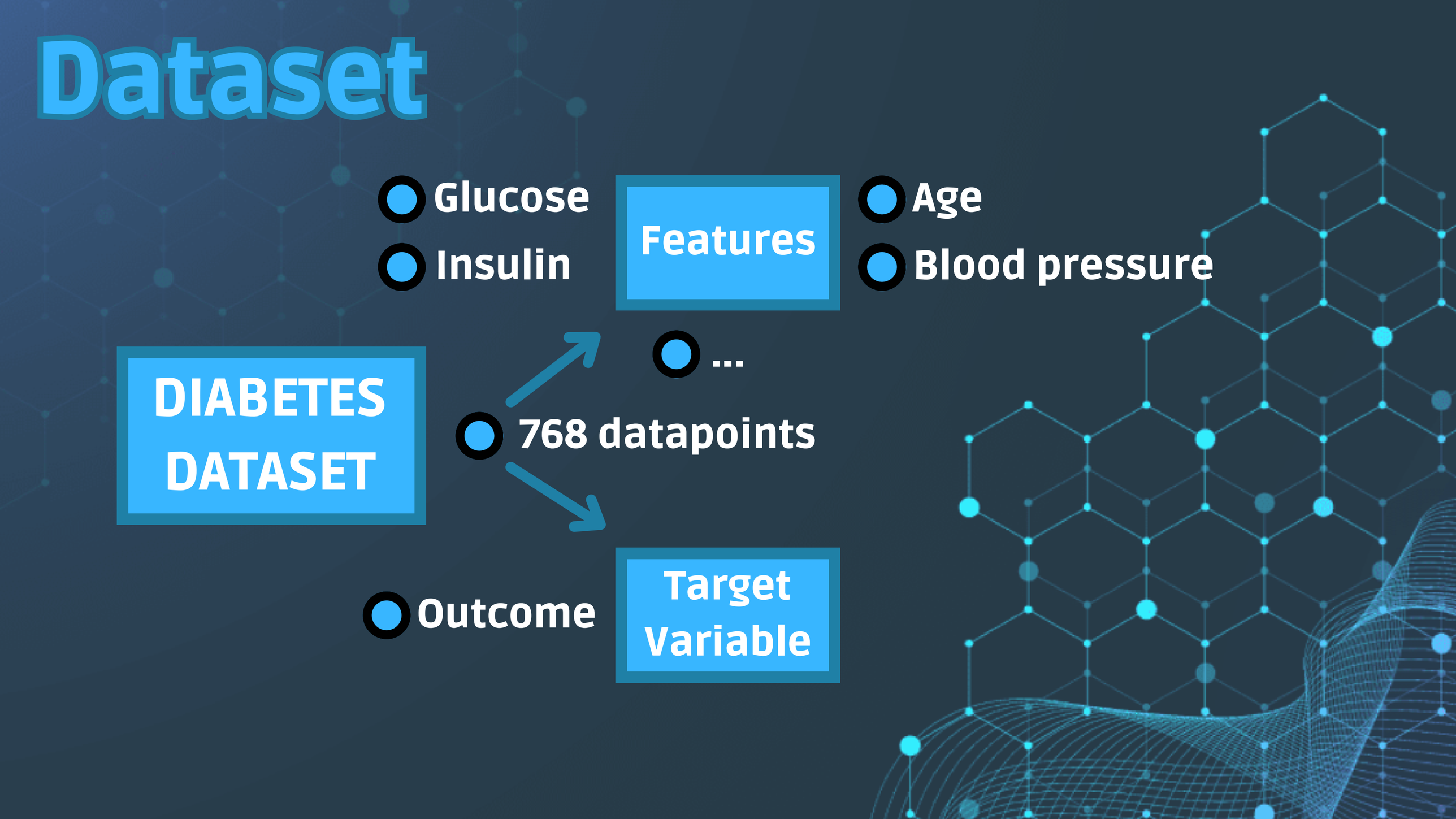
Intelligent Systems 2023-2024

INDEX

CLASSIFICATION

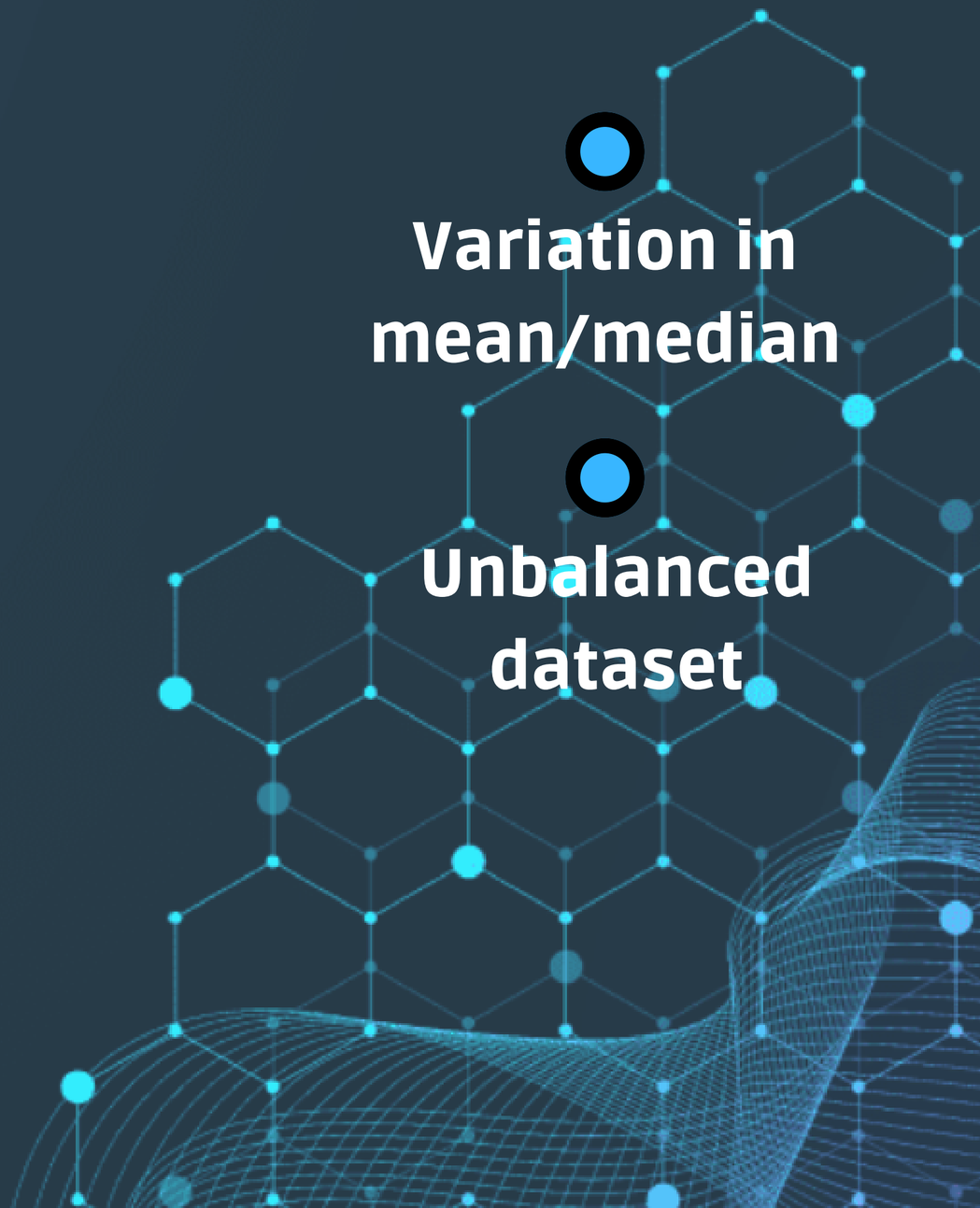
IS A PERSON SUFFERING FROM DIABETES ?

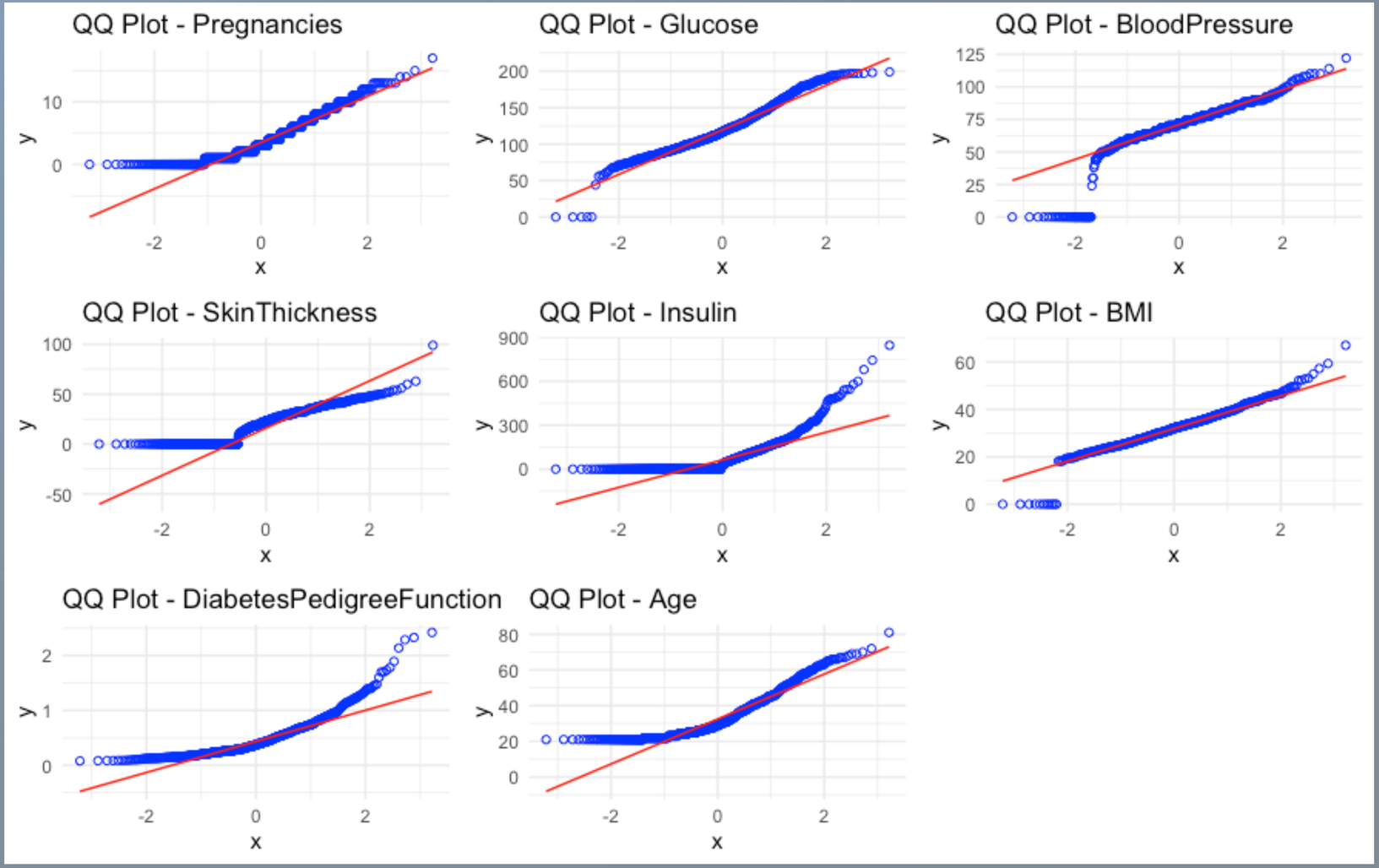
Dataset


Glucose
Insulin
Features
Age
Blood pressure
...
DIABETES DATASET
768 datapoints
Outcome
Target Variable

Data overview

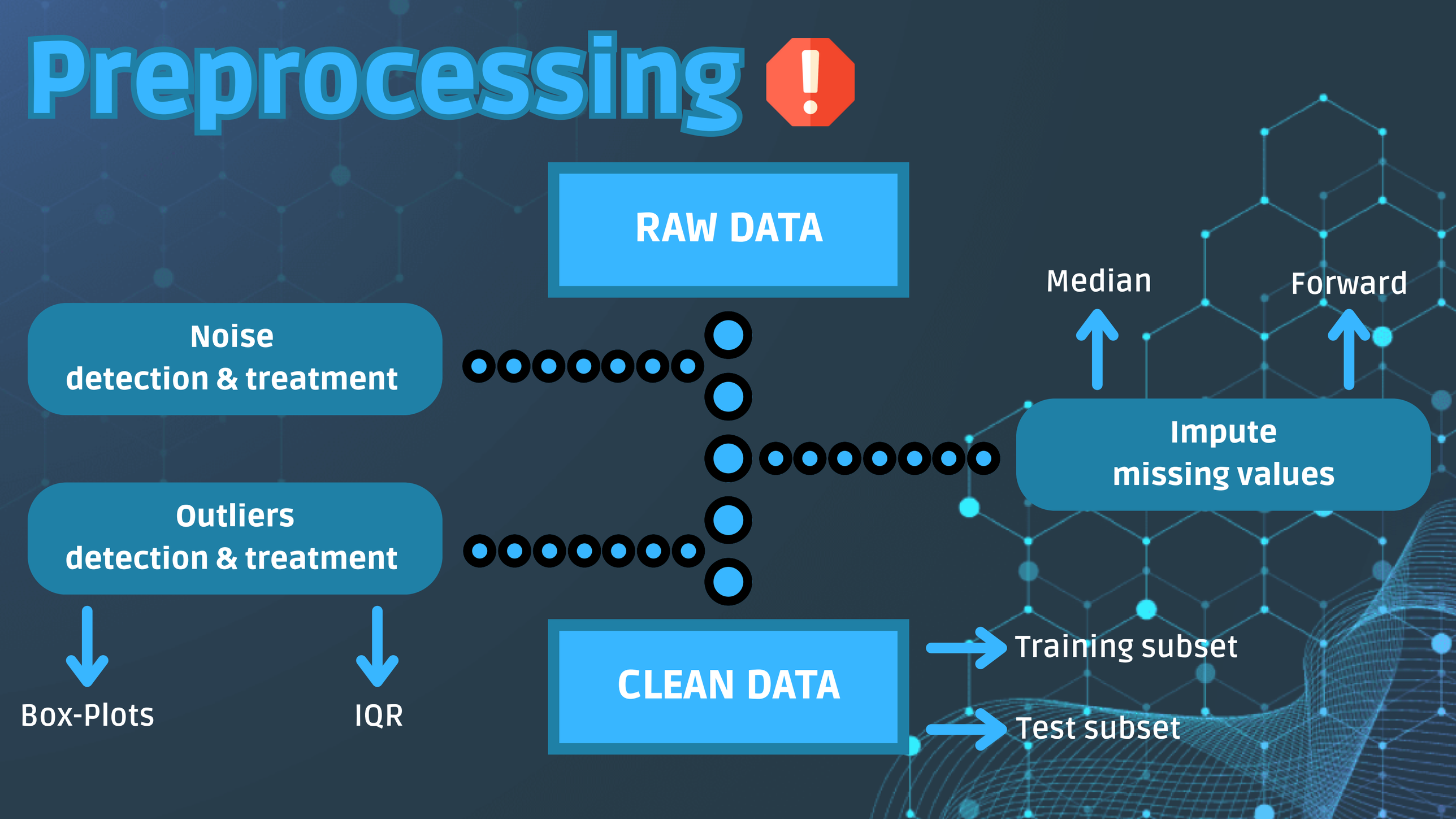
- Skewness
- Range of variability
- Variation in mean/median
- NO normal distribution
  - Saphiro Test
  - QQ-Plots
- Range of dispersion
- Unbalanced dataset





Variable	p-value
Pregnancies	< 2.2e-16
Glucose	1.986e-11
BloodPressure	< 2.2e-16
SkinThickness	< 2.2e-16
Insulin	< 2.2e-16
BMI	1.842e-15
DiabetesPedigreeFunction	< 2.2e-16
Age	< 2.2e-16

Preprocessing



Model's implementation

RANDOM FOREST CLASSIFIER

- n° of trees
- tree deep
- class weight

TRAINING SUBSET

LOADING . . .

SUPPORT VECTOR MACHINE

- kernel
- cost
- class weight

Model's evaluation

RANDOM FOREST CLASSIFIER

...

TEST SUBSET

Sensitivity - 67.50%

Specificity - 86.11%

Accuracy - 74.48%

Precision - 98.36%

F1 score - 87.59%

Model's evaluation

TEST SUBSET

SUPPORT VECTOR MACHINE

Sensitivity - 75.83%

Specificity - 87.50%

Accuracy - 80.21%

Precision - 91.00%

F1 score - 82.72%

RANDOM FOREST CLASSIFIER

SUPPORT VECTOR MACHINE

High precision → accuracy of positive predictions

High specificity → Patient are correctly identified neagtive

SUPPORT VECTOR MACHINE

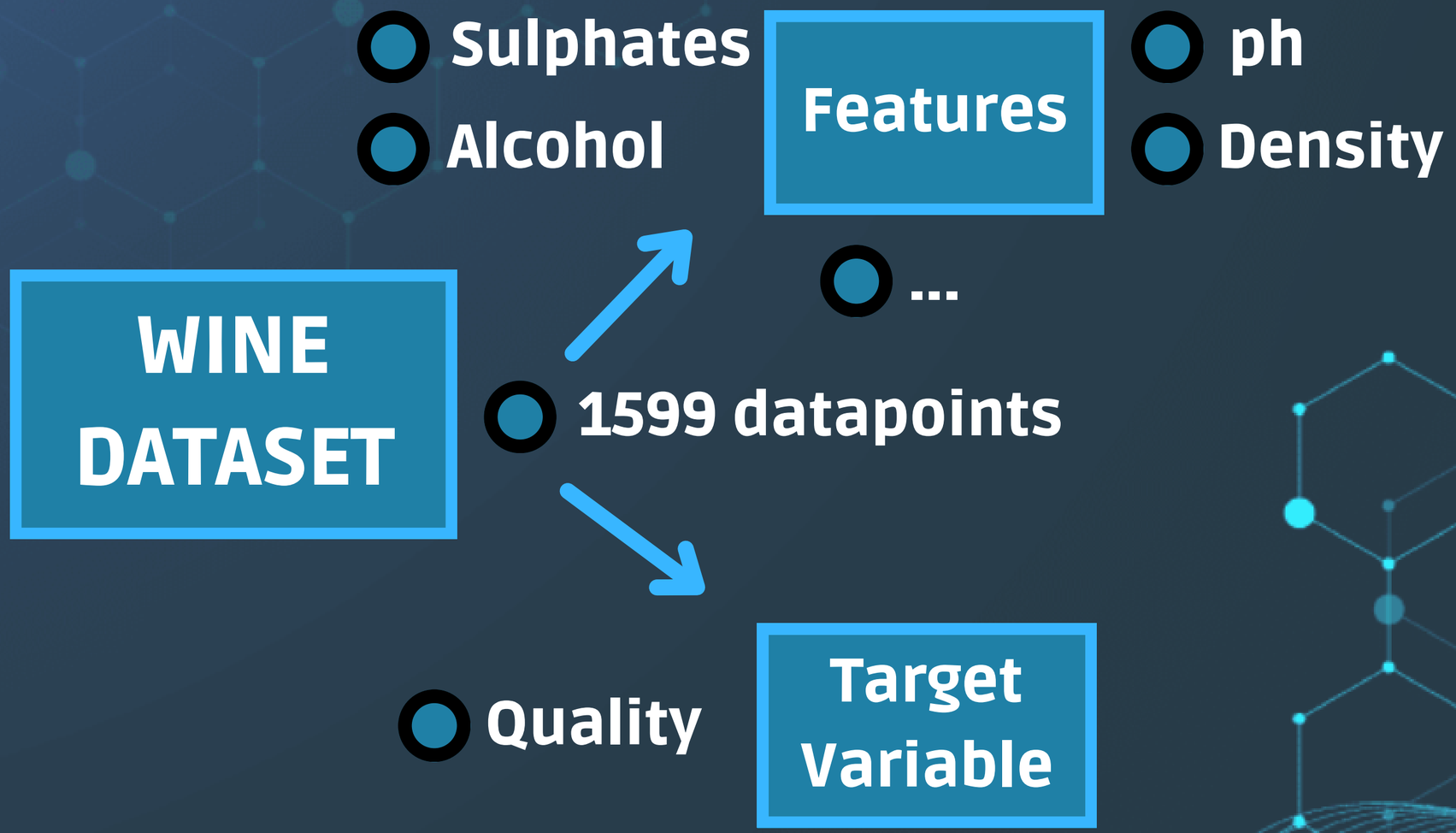
High accuracy → Patient are correctly identified

High sensitivity → Patient are correctly identified positive

REGRESSION

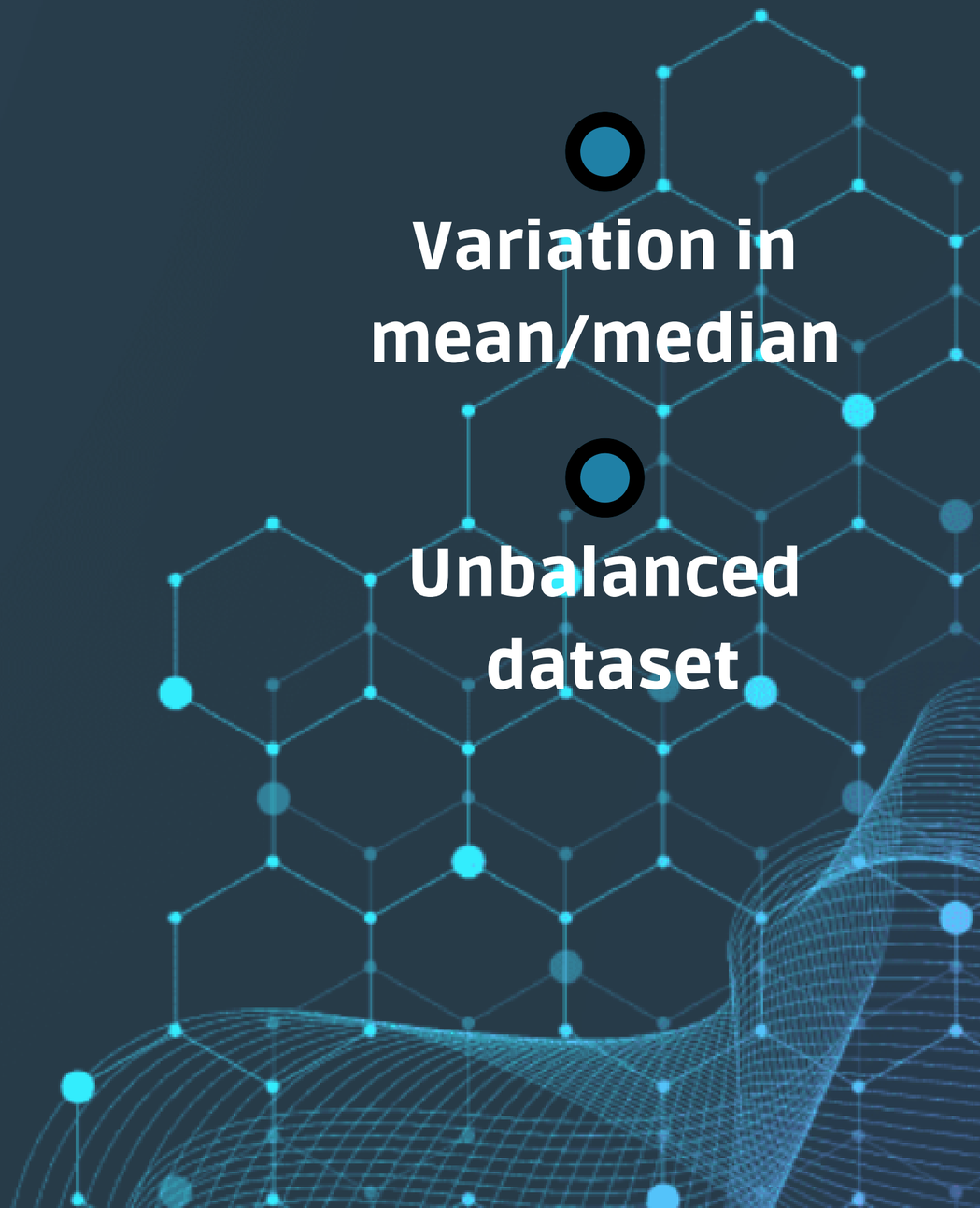
HOW GOOD IS THIS WINE ?

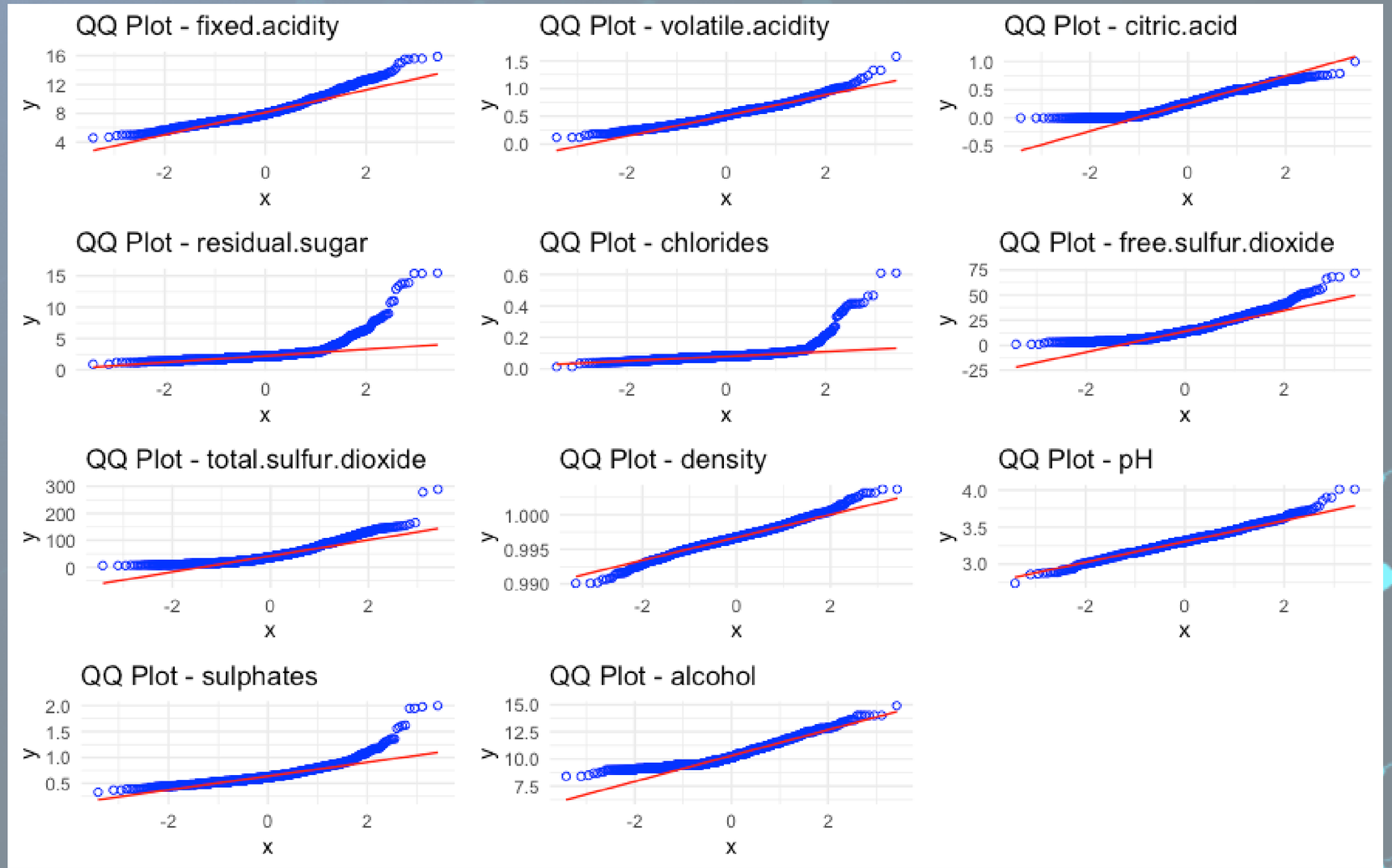
Dataset


WINE DATASET
1599 datapoints
Features
Sulphates
Alcohol
ph
Density
...
Target Variable
Quality

Data overview

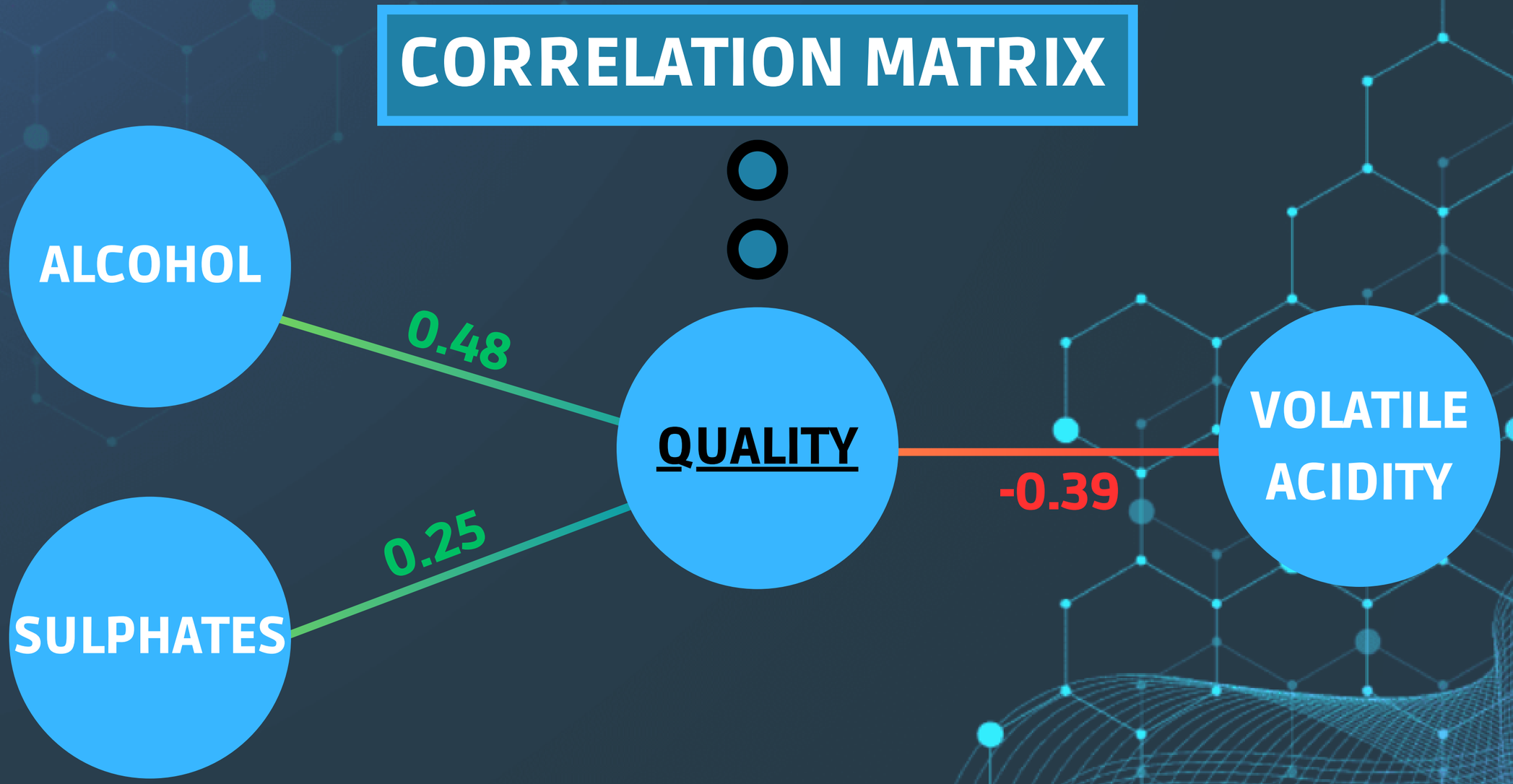
- Skewness
- Range variability
- Variation in mean/median
- NO normal distribution
  - Saphiro Test
  - QQ-Plots
- Range of dispersion
- Unbalanced dataset



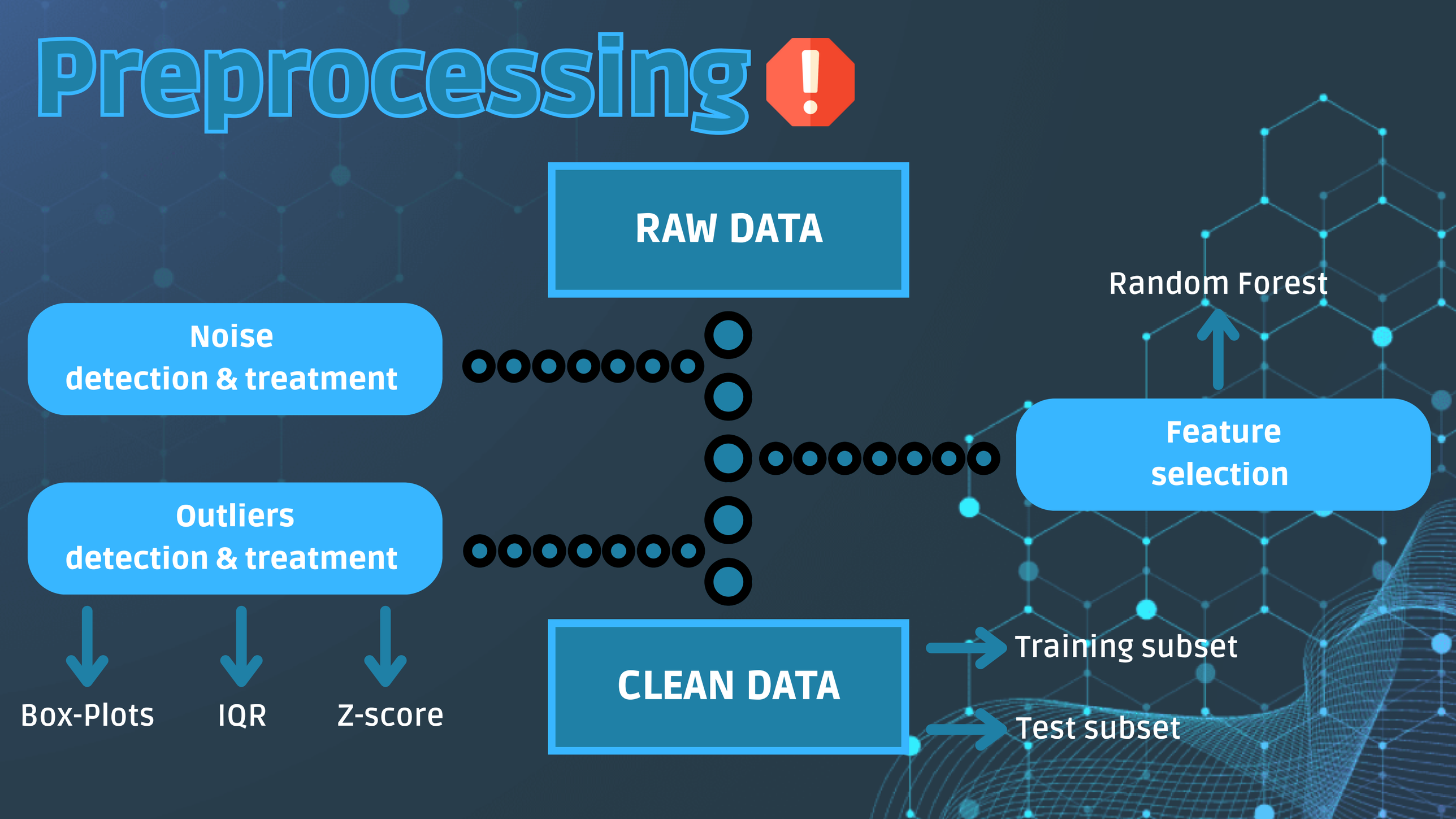


Variable	p-value
Fixed acidity	< 2.2e-16
Volatile acidity	2.693e-16
Citric acid	< 2.2e-16
Residual sugar	< 2.2e-16
Chlorides	< 2.2e-16
Free sulfur dioxide	< 2.2e-16
Total sulfur dioxide	< 2.2e-16
pH	1.712e-06
Sulphates	< 2.2e-16
Alcohol	< 2.2e-16

Features correlation



Preprocessing



Model's implementation

KNN REGRESSION

TRAINING SUBSET

RIDGE REGRESSION

Cross Validation

K parameter

Cross Validation

λ parameter

HYPERPARAMETER TUNING

Model's evaluation

TEST SUBSET

BASELINE MODEL

VS

KNN REGRESSION
K = 19

RIDGE REGRESSION
λ = 0.02

BASELINE MODEL	MAE 0.6283	RMSE 0.7165	R^2 -0.00053
KNN REGRESSION K = 19	MAE 0.4730	RMSE 0.5685	R^2 0.3729
RIDGE REGRESSION λ = 0.02	MAE 0.4730	RMSE 0.5499	R^2 0.4109

THANK YOU
FOR YOUR TIME !

Sassi Gabriele