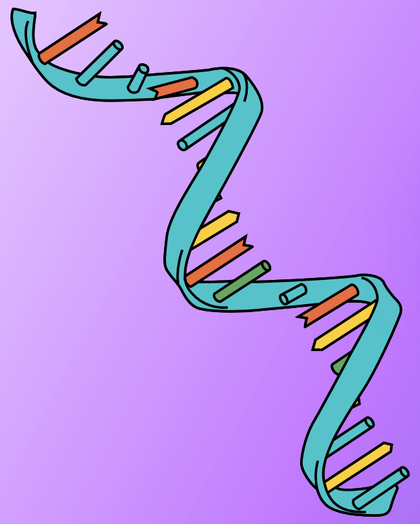
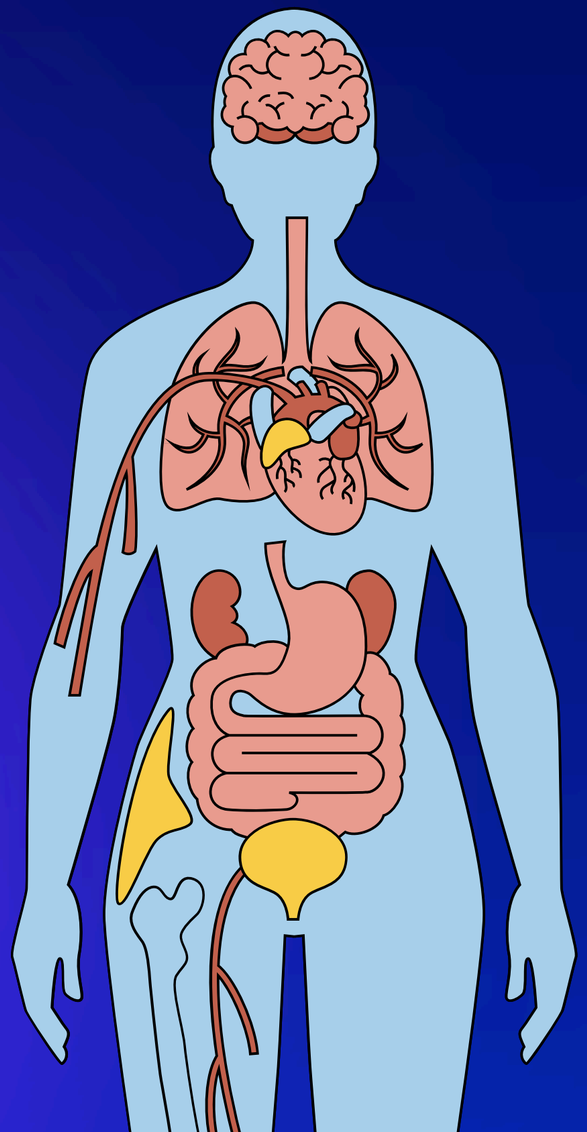
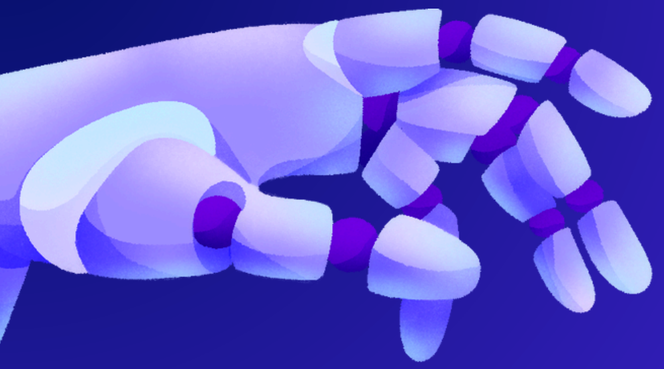
Professors:
Diego Gragnaniello
Eduardo Sommella
Vicky Caponigro



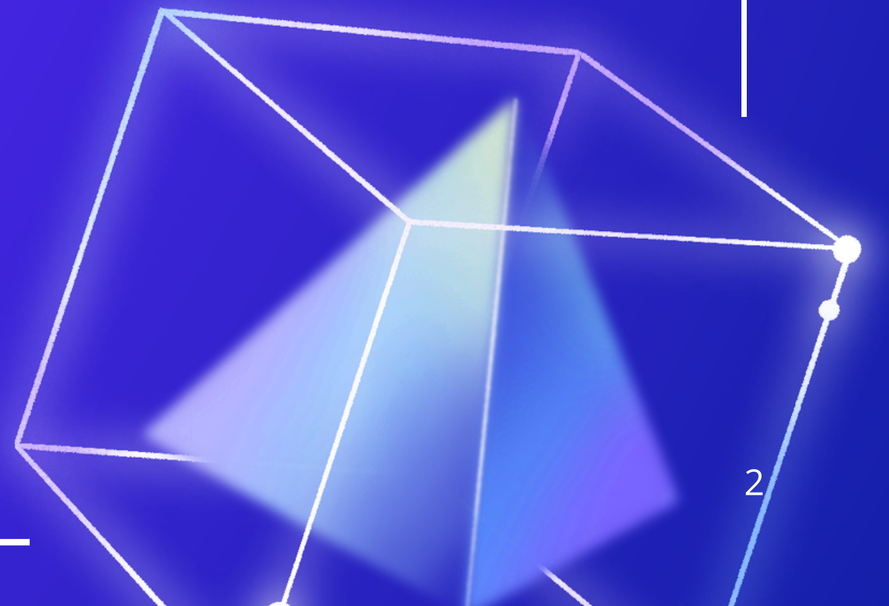
ARTIFICIAL INTELLIGENCE FOR OMICS DATA ANALYSIS



Bruno Salvatore mat. 0623200054
Apicella Mario mat. 0623200060



INDEX



PROJECT OVERVIEW

The primary objective of this project is to analyze and classify metabolomics datasets to identify biomarkers and gain insights into the metabolic profiles of two distinct classes: CHD (Cardiac Heart Defects) CTRL (Control Group). By employing advanced machine learning techniques and feature selection methods, the project aims to enhance classification accuracy while ensuring robust and biologically meaningful results.

The ultimate goal is to provide a robust pipeline for biomarker discovery and classification in metabolomics studies, supporting medical research and potential clinical applications. This workflow can be adapted to other datasets and diseases, ensuring its broad applicability in precision medicine.

DATASET: METABOLOMICS ESI+ METABOLOMICS ESI-

Datasets Description

Datasets are organized initially with samples in columns and variables (metabolites) in rows. Each dataset includes:

- A **mass-to-charge ratio (m/z)** column representing the mass spectrometry measurements.
- A **metabolite name row** corresponding to each m/z value.
- Columns with **samples**
- **Additional metadata columns** in negative dataset (e.g., AF, AG, AH) specific to metabolomics analysis, which were excluded from downstream analyses.

Sample Classes:

- **QC (Quality Control)**
- **CHD (Cardiac Heart Defects)**
- **CTRL(Controls)**

THE CONTEXT: CONGENITAL HEART DISEASE(CHD)

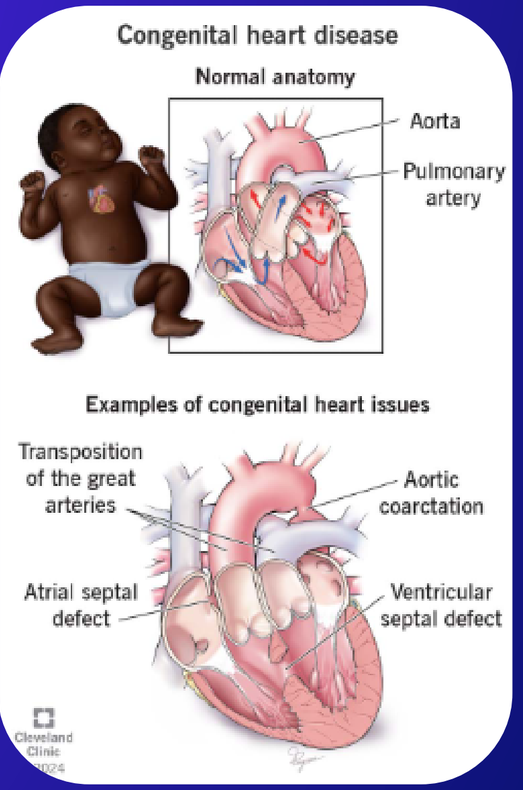
CHD

Congenital heart disease (CHD) is the most common congenital anomaly, affecting approximately 0.63–0.8% of live births in Europe. It encompasses a diverse range of structural heart and vascular malformations, varying in anatomy, clinical presentation, and severity.

Metabolomic profiling is well-suited for studying CHD because it provides insights into the biochemical pathways and metabolic disruptions associated with the disease

Benefits:

- Identifies potential biomarkers for early diagnosis.
- Monitors disease progression and treatment efficacy.
- Enhances understanding of CHD pathophysiology at the molecular level.



MATERIALS AND METHODS

Data Preprocessing:
- Normalizations: PQN and TIC.
- PCA for dimensionality reduction.
- Multiblock creation and autoscaling for combined data analysis.

Classification Models:
- SVM: Optimized with Grid Search and LOOCV.
- Random Forest: Leveraged feature importance.
- Logistic Regression: Linear classification.

Feature Selection:
- SHAP for SVM and Logistic Regression.
- Random Forest's inherent importance metric..

Validation:
- Grid Search CV: Iterative hyperparameter optimization.
- LOOCV: Robust individual sample evaluation.

Univariate Analysis:
- Heatmaps of mean intensity for shared important features.
- Pairplots for principal component distributions.

Tools:
- Python: Pandas, NumPy, scikit-learn, SHAP.
- Visualizations: Matplotlib, Seaborn.
- Google Colab for computation.

PROJECT WORKFLOW

The following pipeline has been repeated for 3 kind of normalization in order to compare model response

01

PreProcess Raw Data

02

Statistical Analysis

03

Multiblock on splitted Classes:

- Sum PCA
- Outliers removal
- Split Train and test set

04

Find model Parameters:

- Raw Data Cleaning
- Split Train and test set
- Preprocessing
- Frobenius Norm
- Multiblock
- Autoscaling
- Find and saveModel Parameters

05

Find most important features:

- preprocess cleaned raw datas
- Frobenius Norm
- Multiblock
- Autoscaling
- Load models parameters
- test Models on Dataset
- Find most important features

06

Fit models on most important features

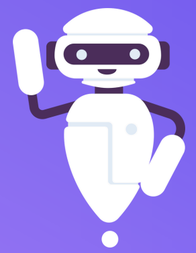
- Consider only most important features fior raw data
- Split train and test set
- Preprocess
- Frobenius Norm
- Multiblock
- Autoscaling
- Fit Models on most importan features

07

Other Normalization methods

01

PRE PROCESS RAW DATAS

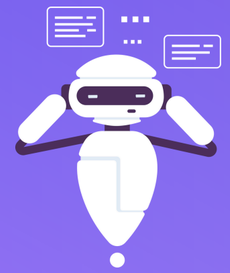


DATA IMPORT

Performing the import and the transposition of the dataset cleaning useless rows in metabolomics_negative

PRE PROCESSING

- Data Visualization
- Normalization with PQN(first used)
- Handling missing values
- Visualize scatter plots and histograms
- PCA



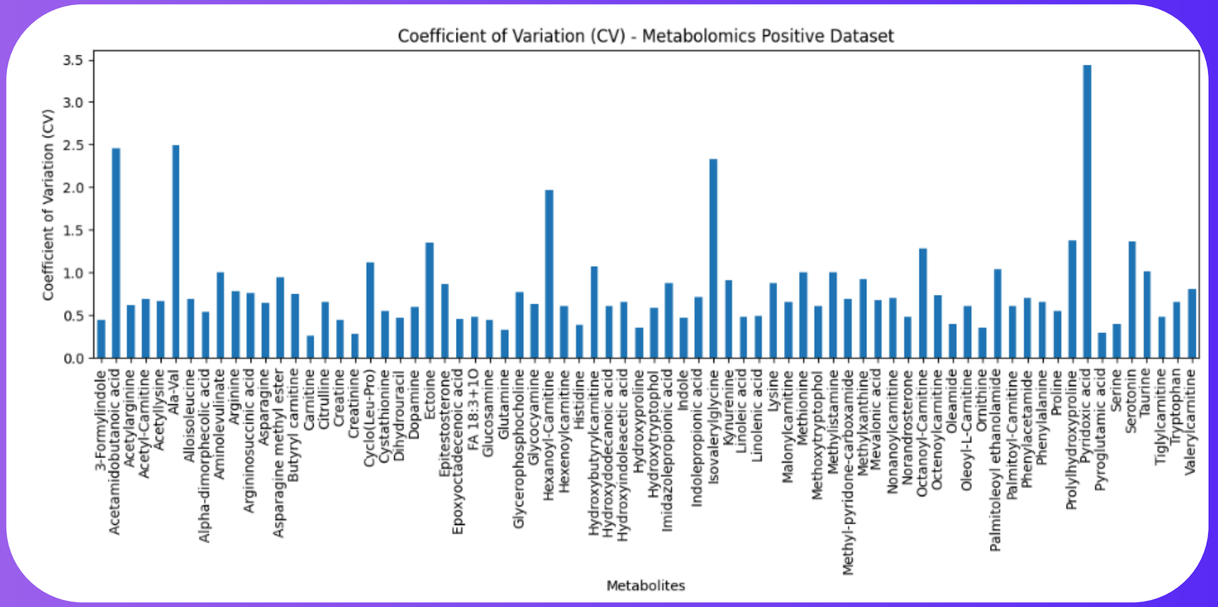
POST PROCESSING

- QC removal
- Class Divison (CHD and CTRL)
- Preprocessing
- Multiblock and autoscaling
- Sum PCA
- Dimensionality reduction
- Outlier Removal
- Division in train and test set

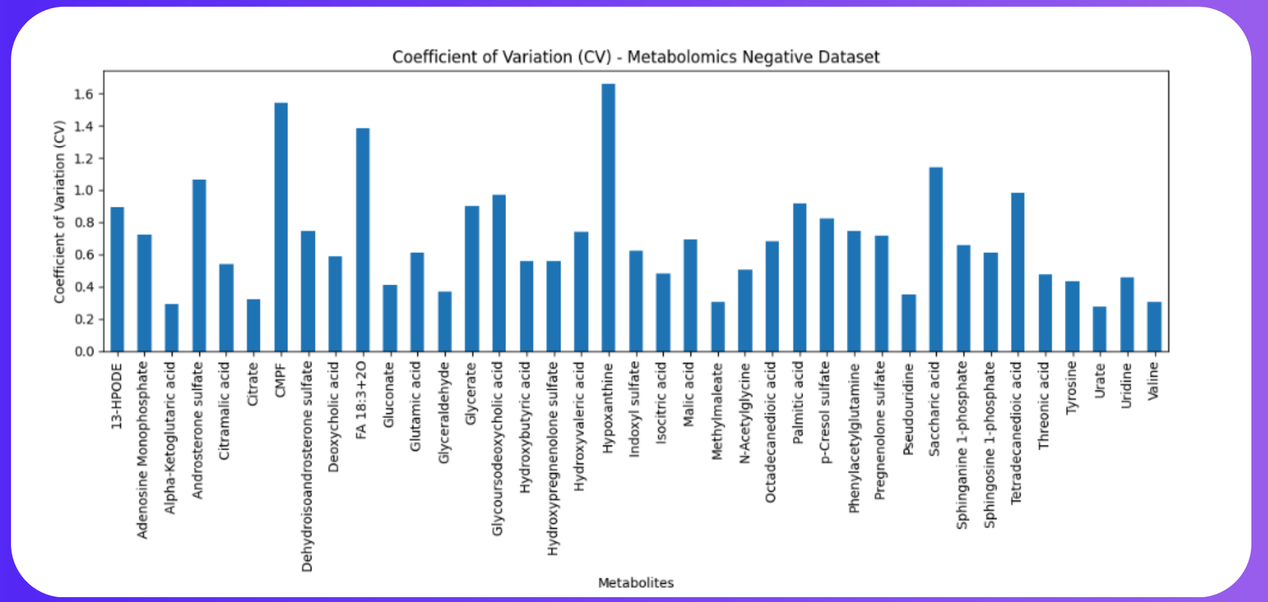
01

VARIABILITY ANALYSIS OF METABOLOMICS DATASETS

ESI+



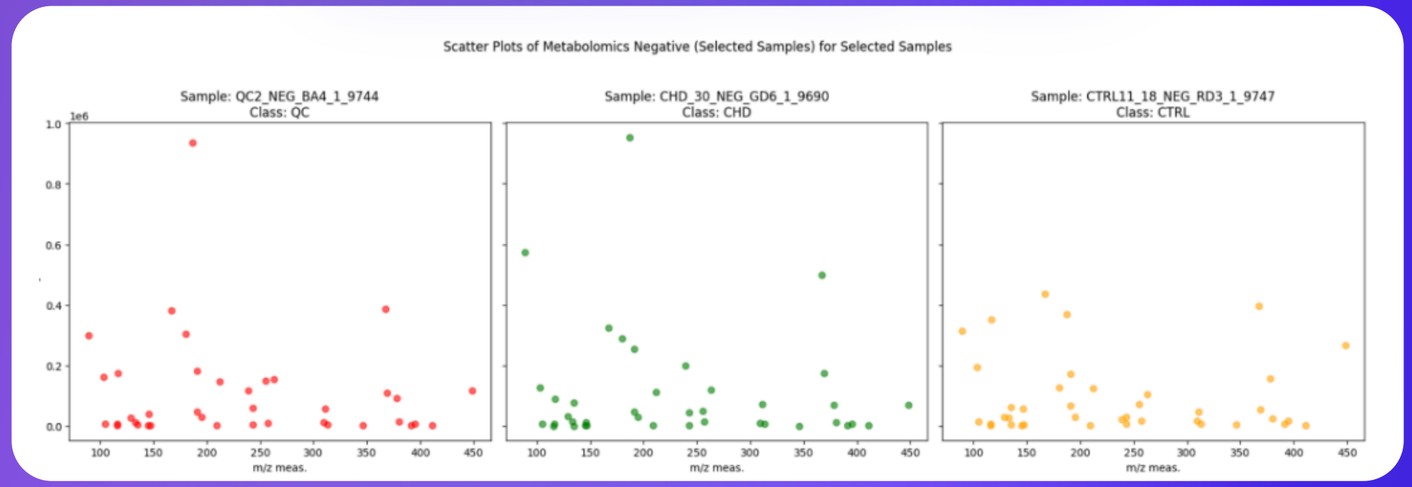
ESI-



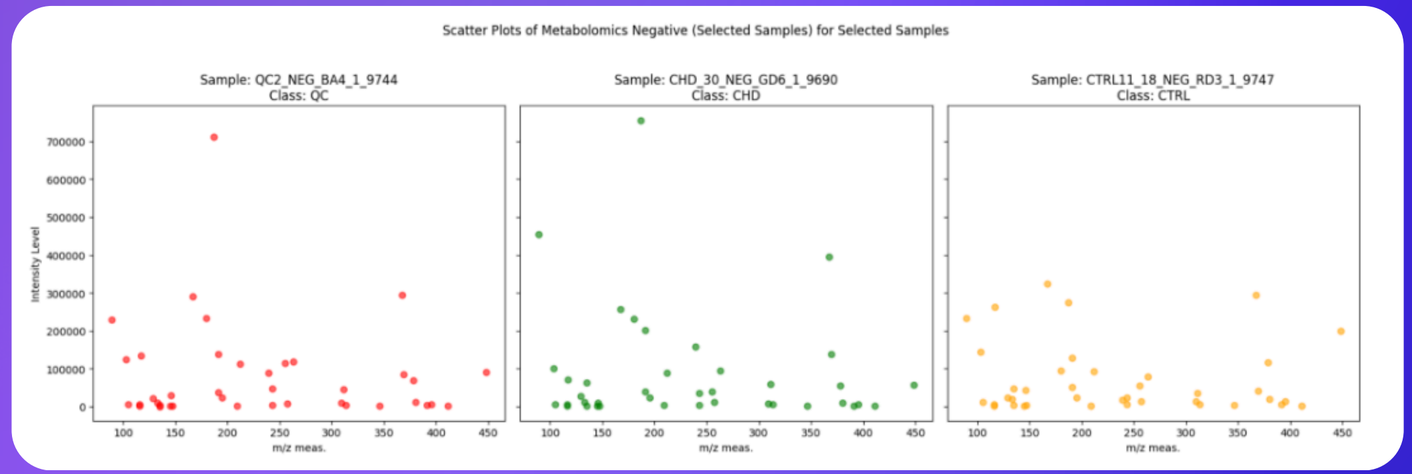
01

DATA VISUALIZATION PRE AND POST NORMALIZATION: ESI-

PRE NORM ESI-



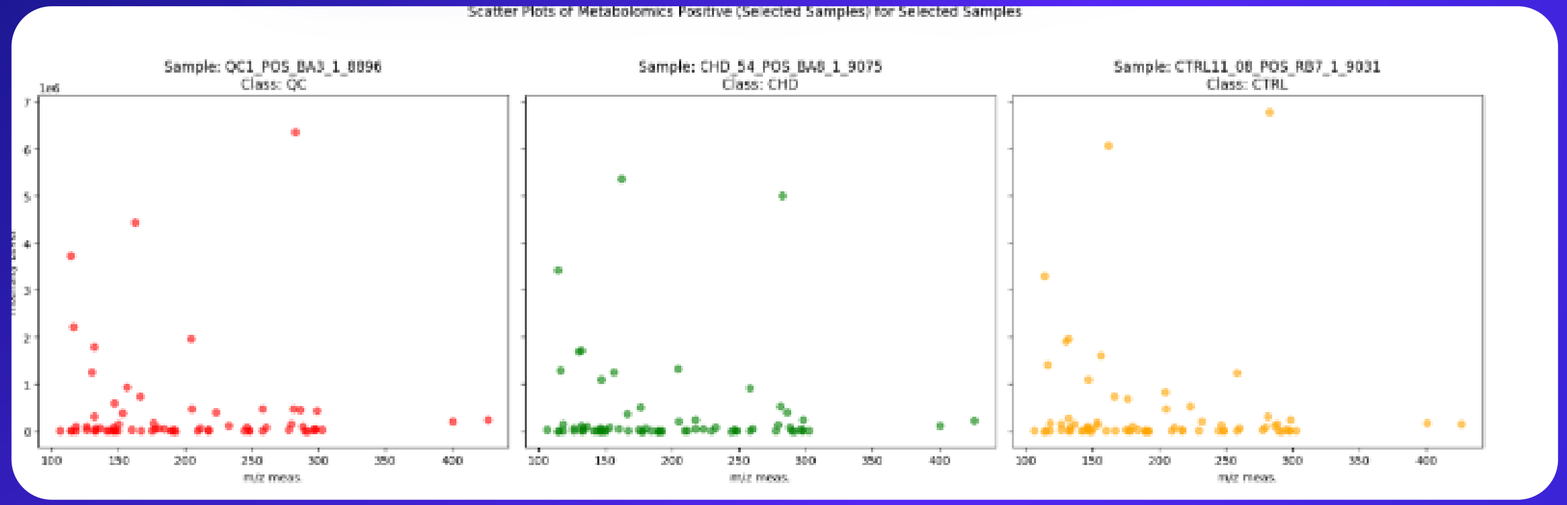
POST NORM ESI-



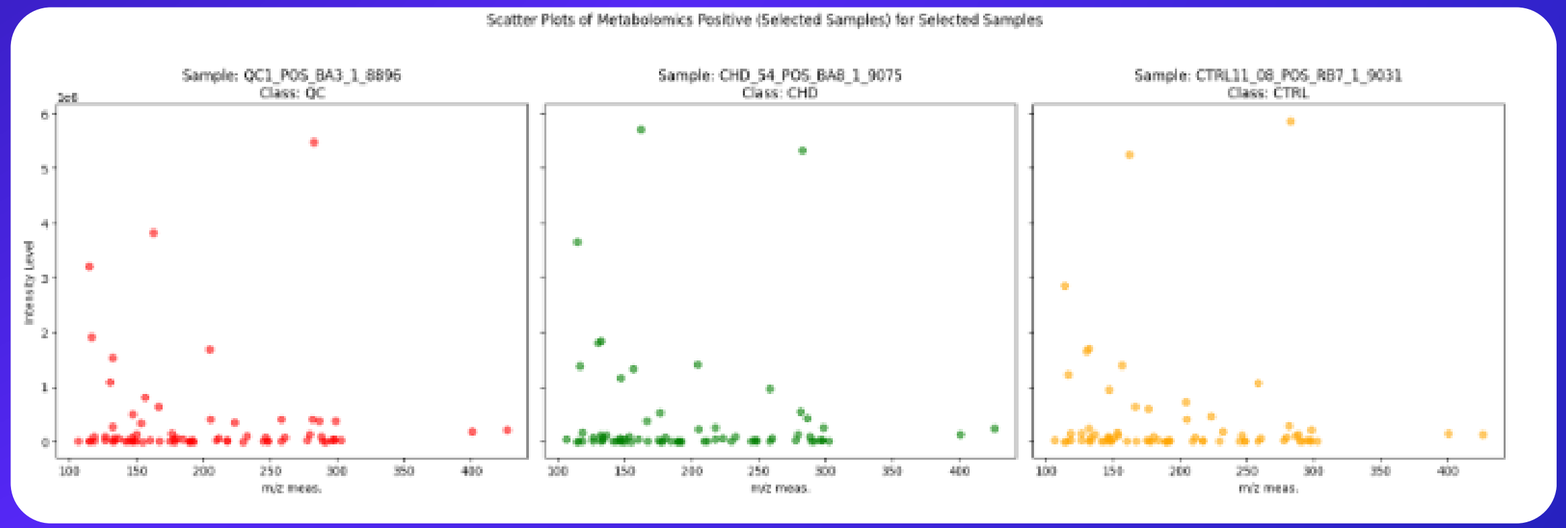
01

DATA VISUALIZATION PRE AND POST NORMALIZATION: ESI +

PRE NORM ESI+



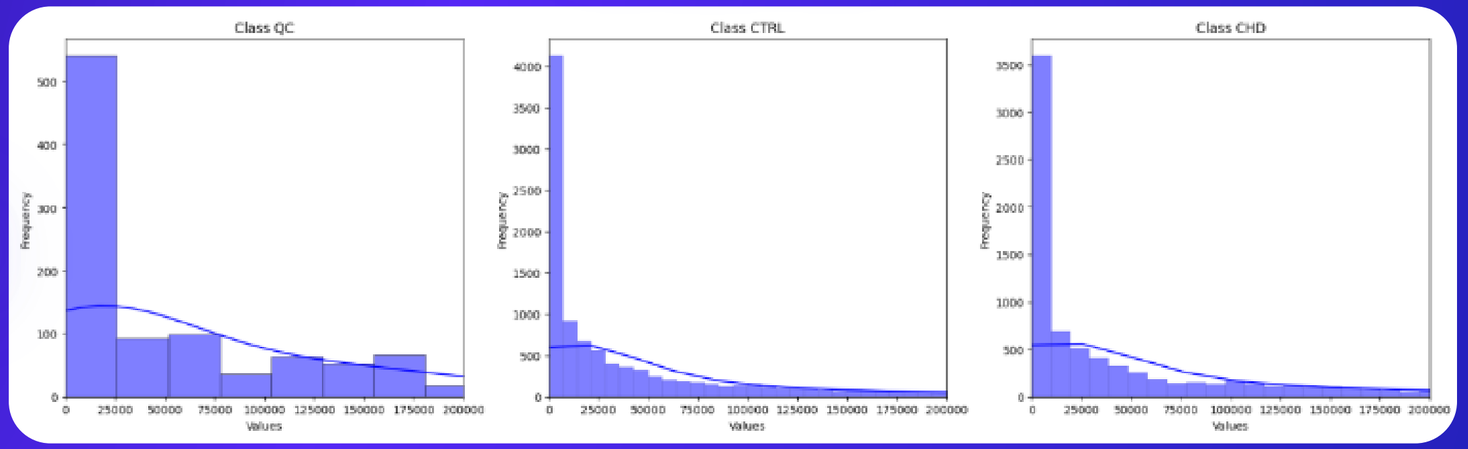
POST NORM ESI+



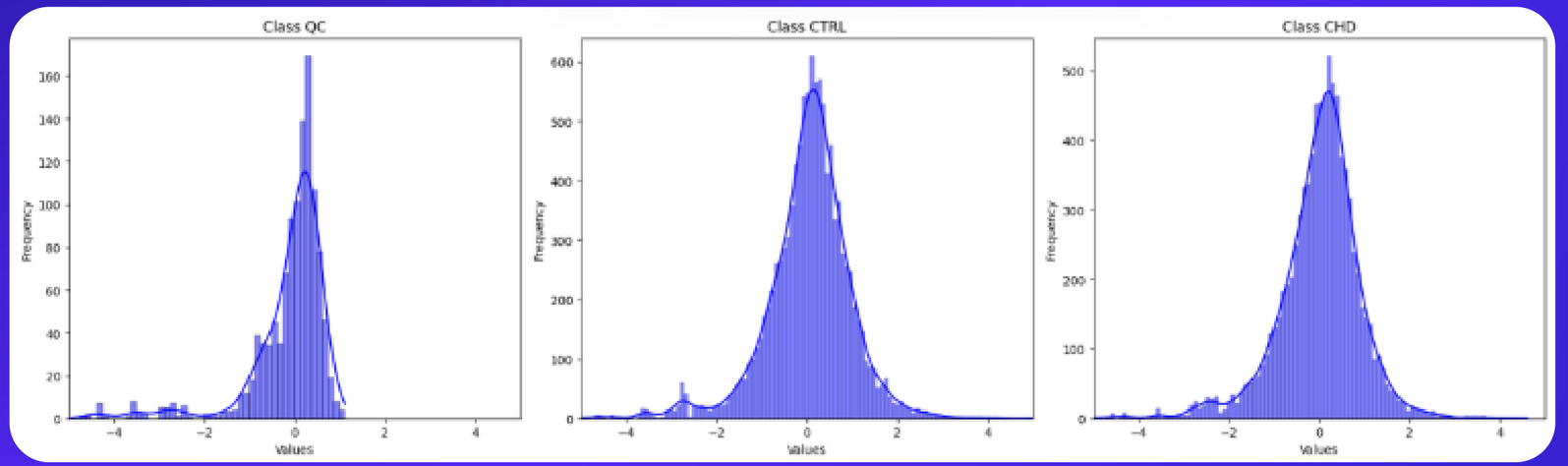
01

DATA VISUALIZATION: HISTOGRAMS ESI-

PRE NORM HIST

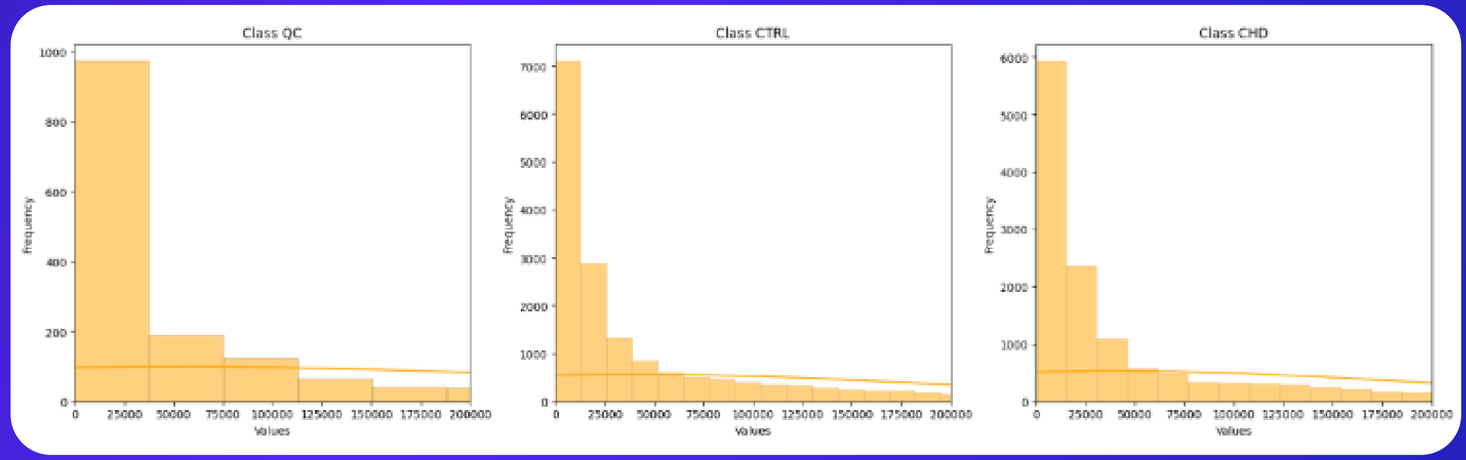


POST NORM,LOG,AUTOSCALING

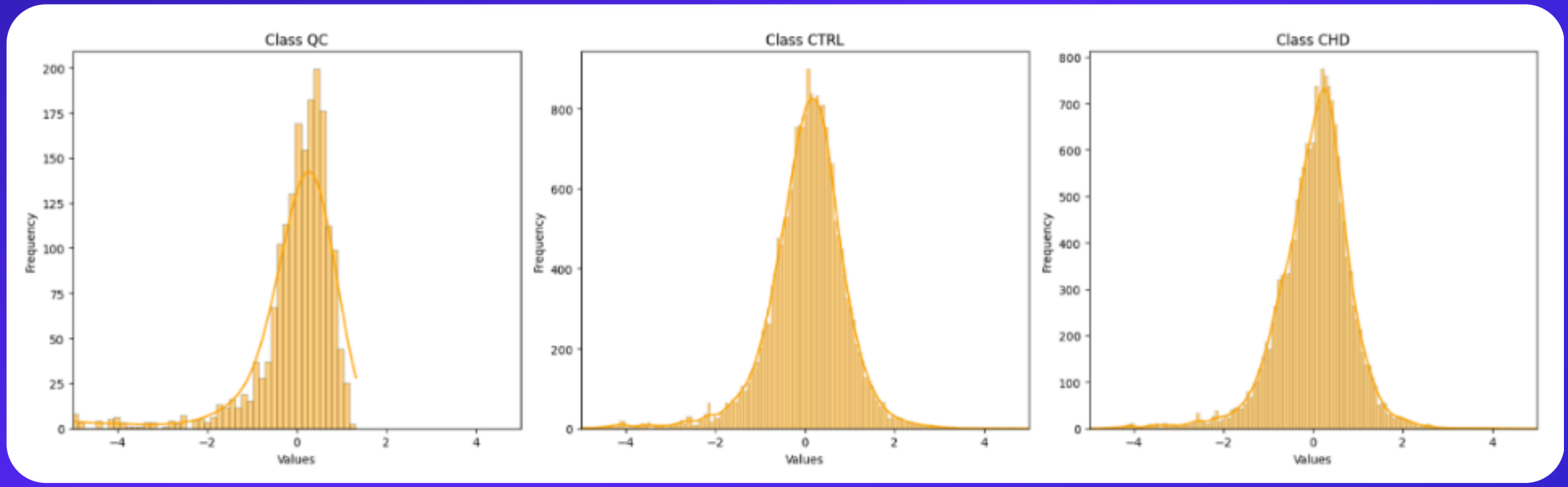


DATA VISUALIZATION: HISTOGRAMS ESI+

PRE NORM HIST

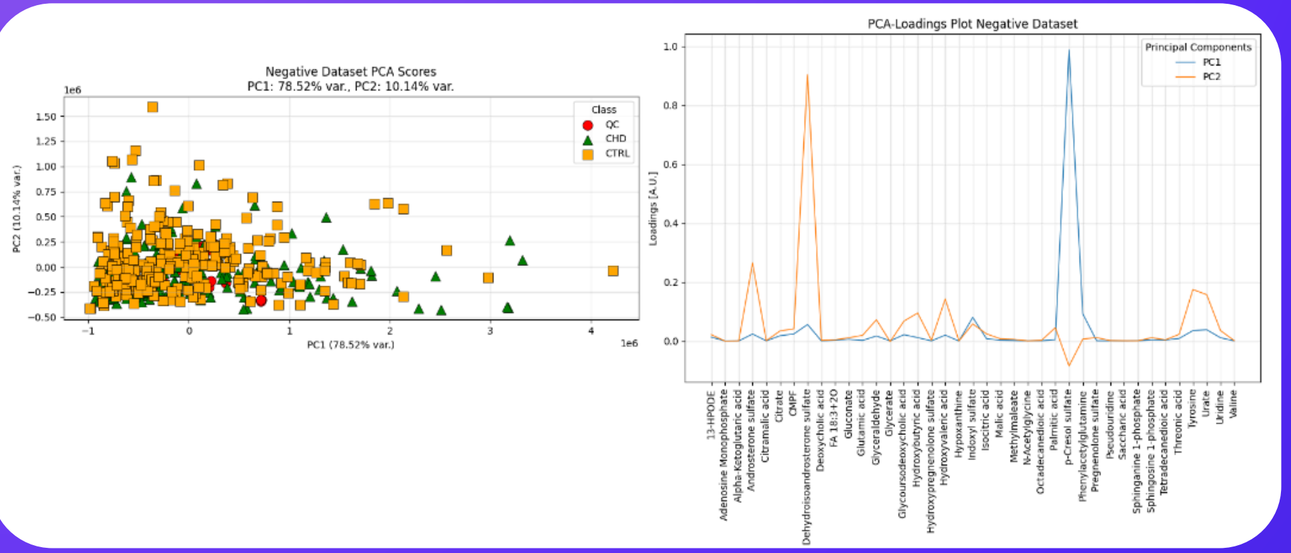


POST NORM,LOG,AUTOSCALING

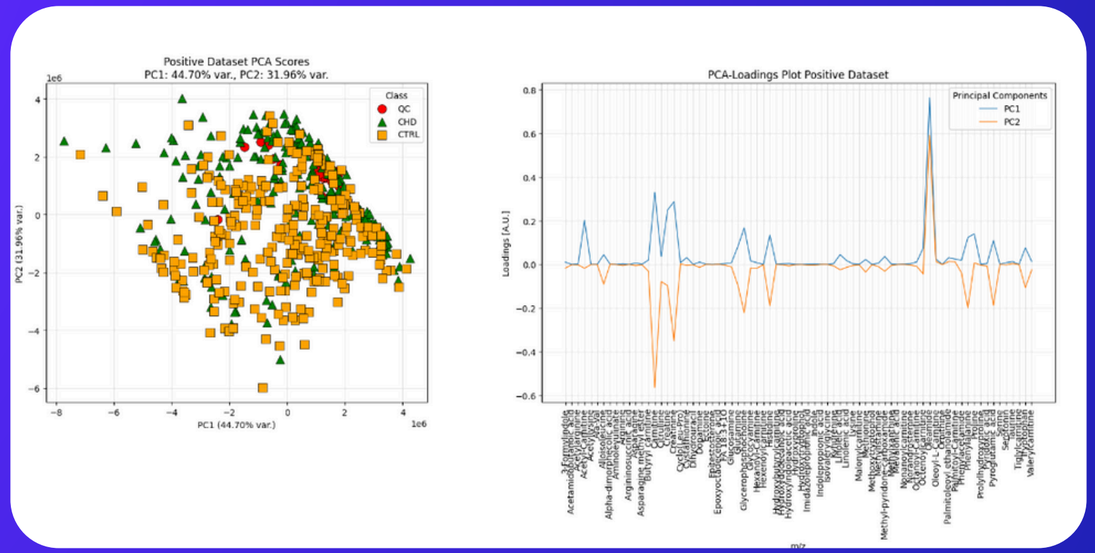


DATA VISUALIZATION: PCA VIEW PRE NORMALIZATION

ESI-

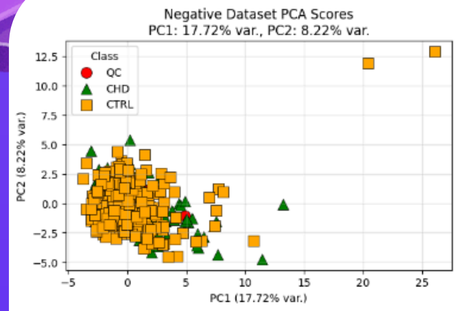


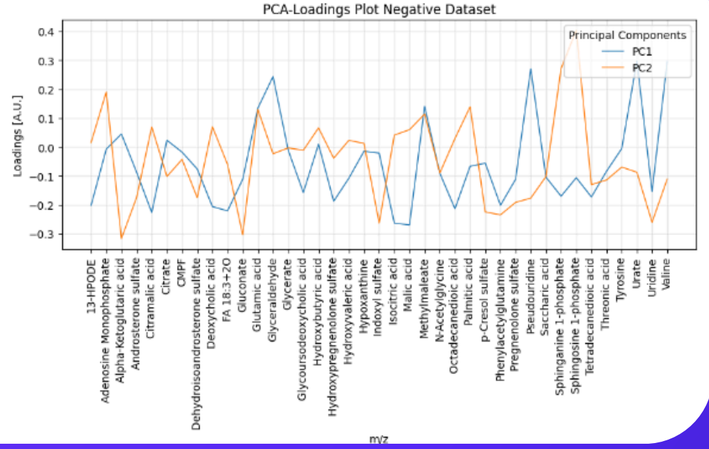
ESI+



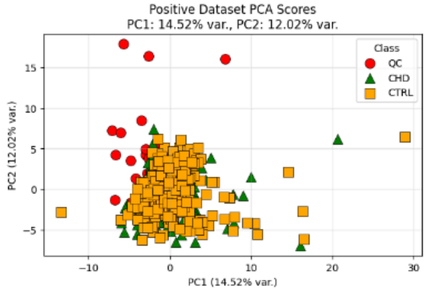
DATA VISUALIZATION: PCA VIEW POST NORMALIZATION

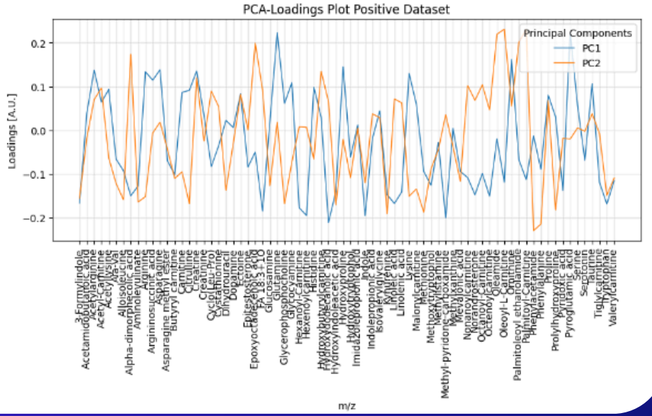
ESI-





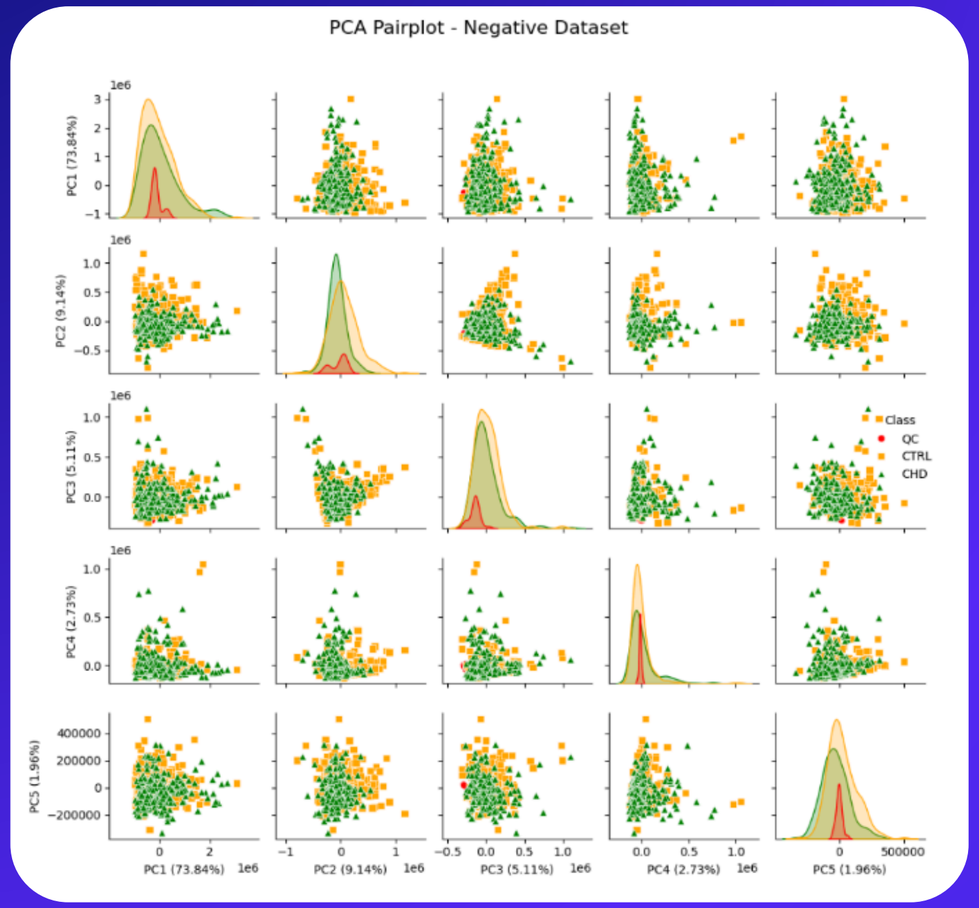
ESI+



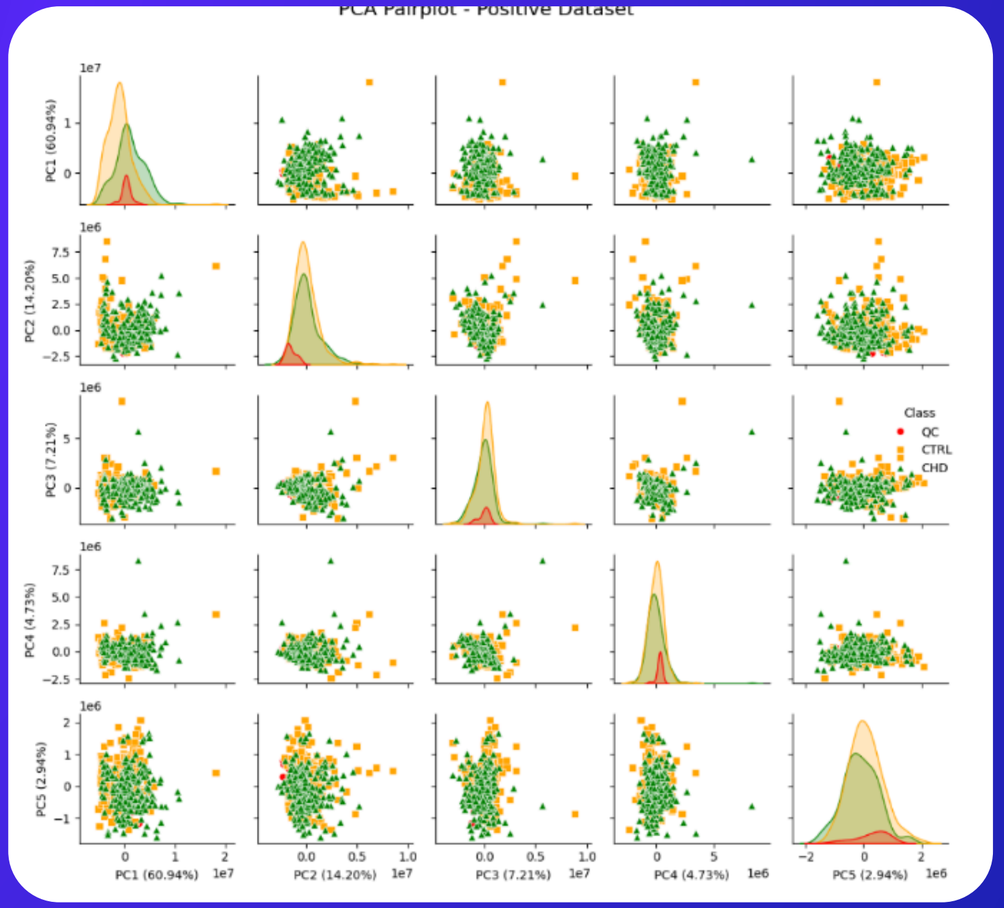


PAIRPLOTS

ESI-

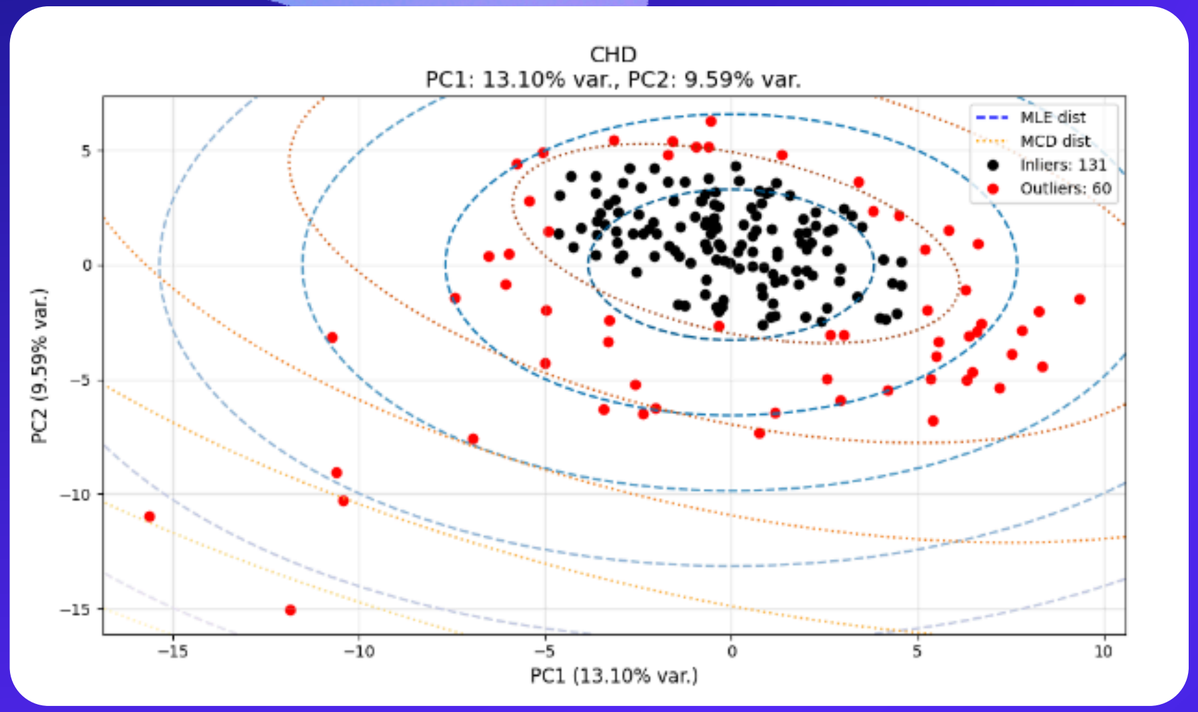


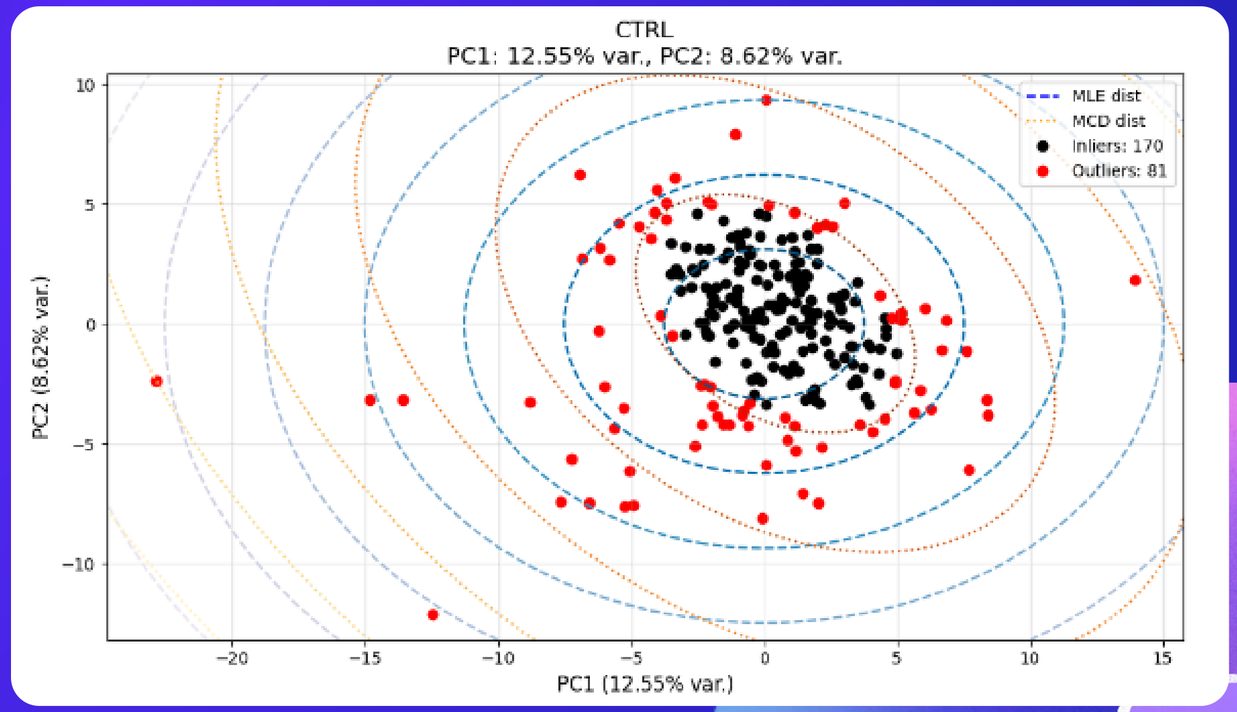
ESI+



POST PROCESSING OUTLIER REMOVAL

POST MULTIBLOCK AND PCA ON DIVIDED CLASSES

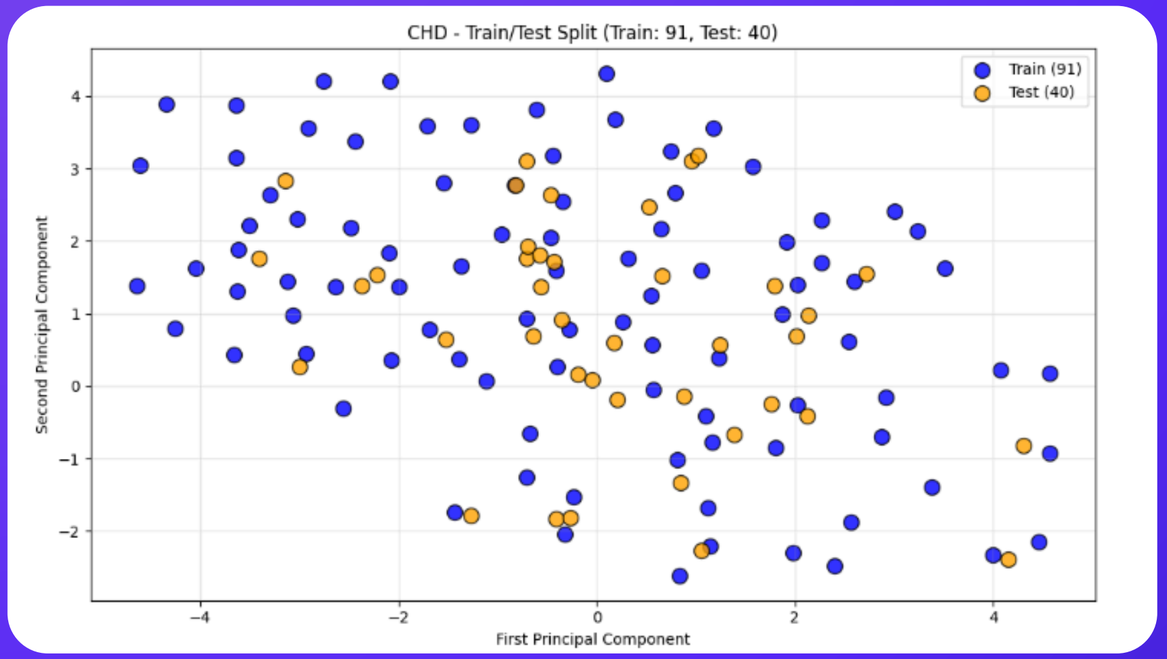




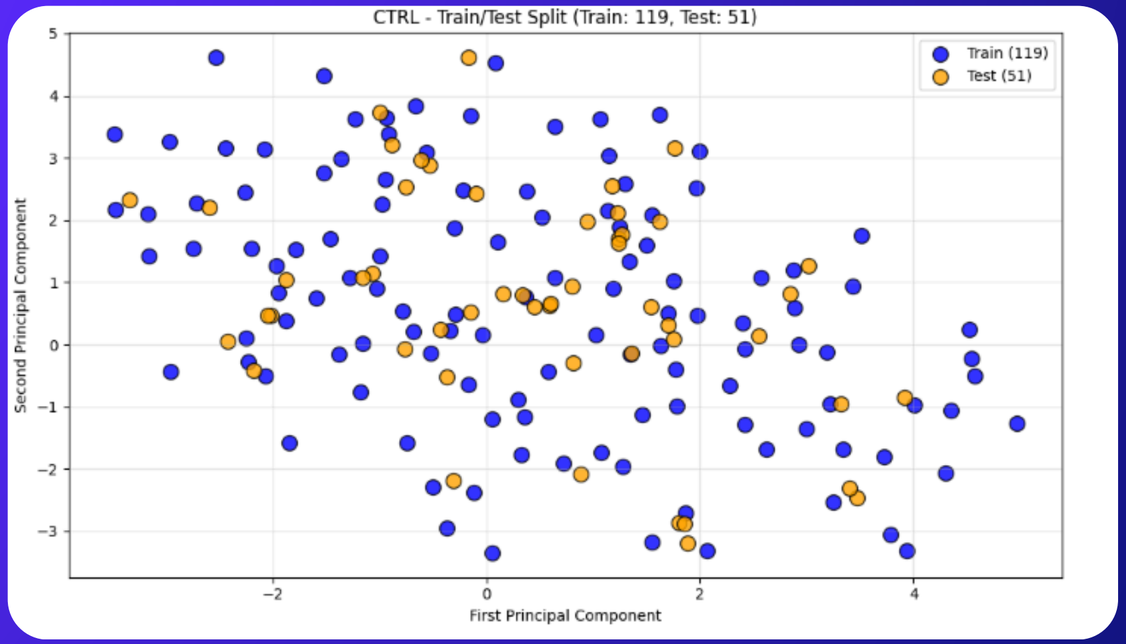
03

SPLIT IN TRAIN AND TEST SET

CHD



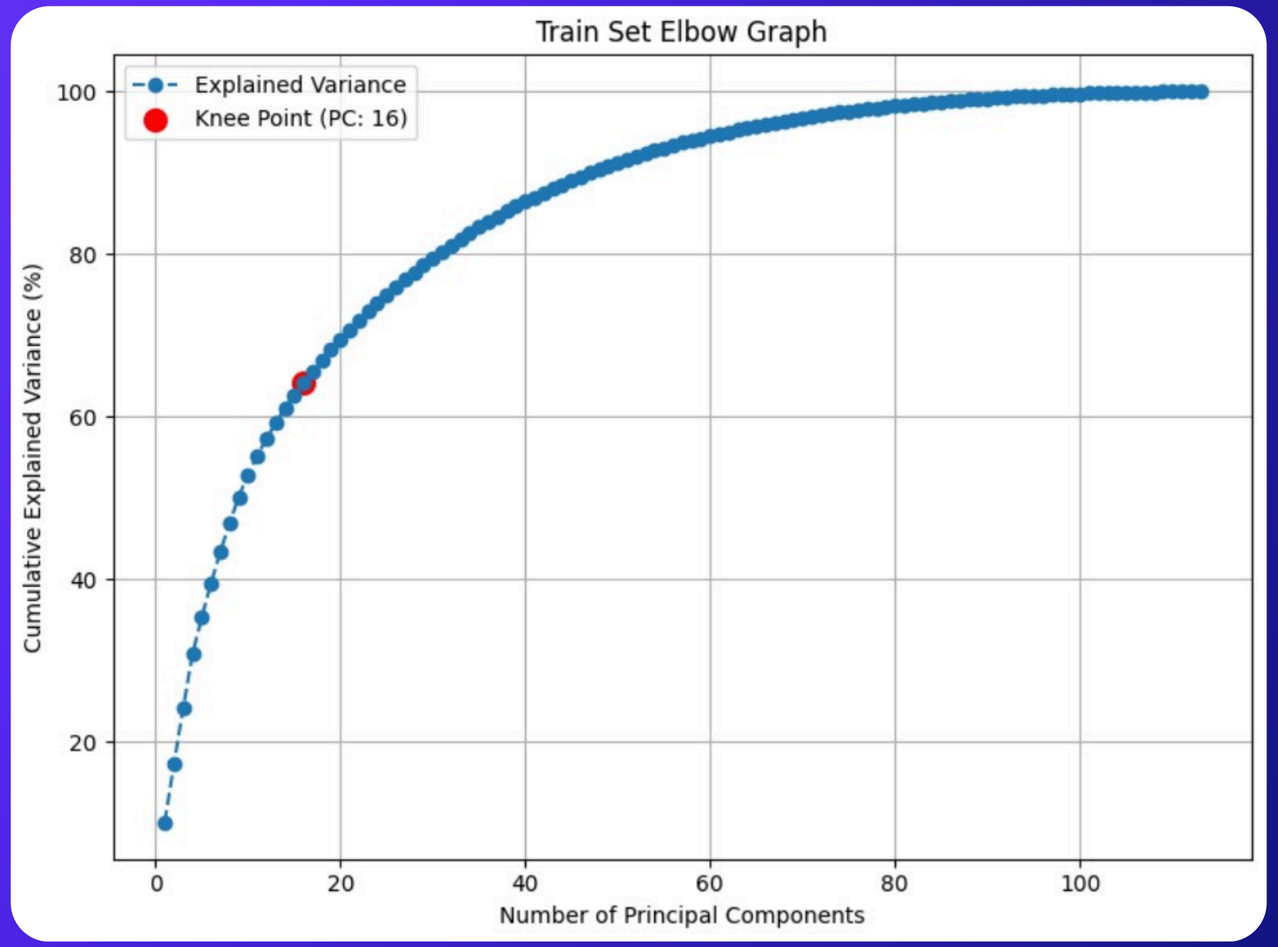
CTRL



04

POST CLEANING RAW DATA AND MULTIBLOCK ON TRAIN AND TEST SET:

DIMENSIONALITY REDUCTION

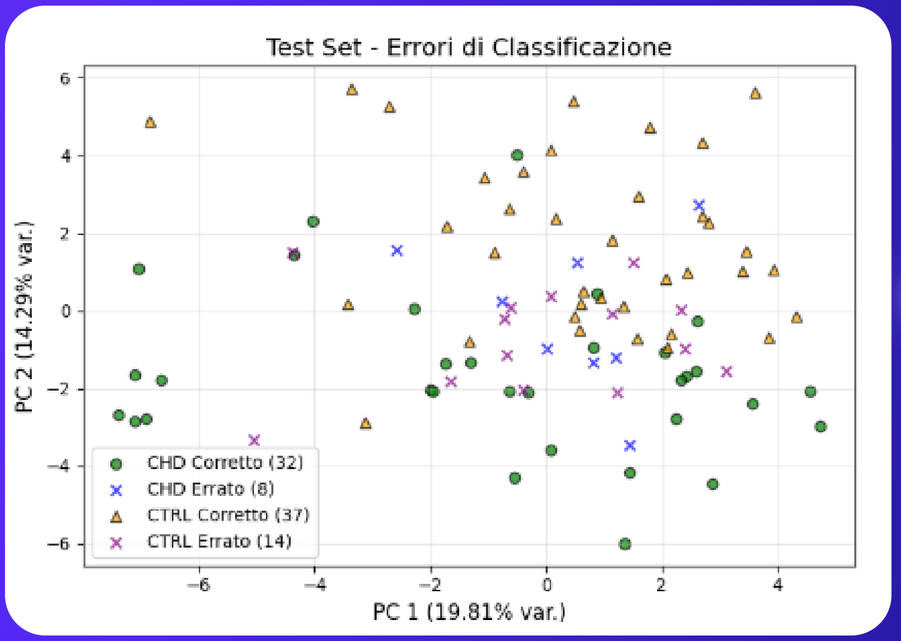


PQN: FIND MODEL PARAMETERS
FIT MODELS AND RESULTS

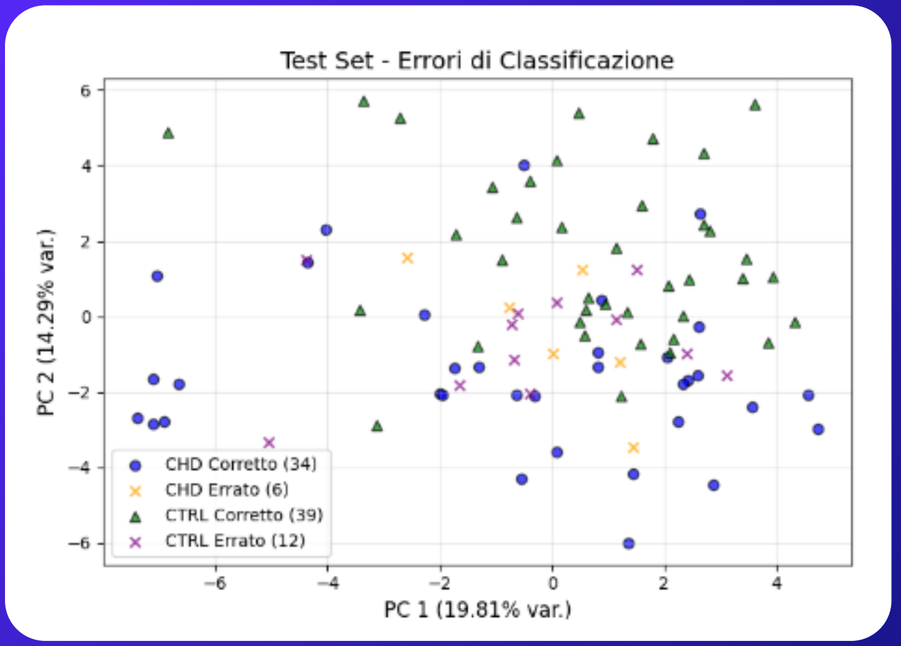
04

SVM

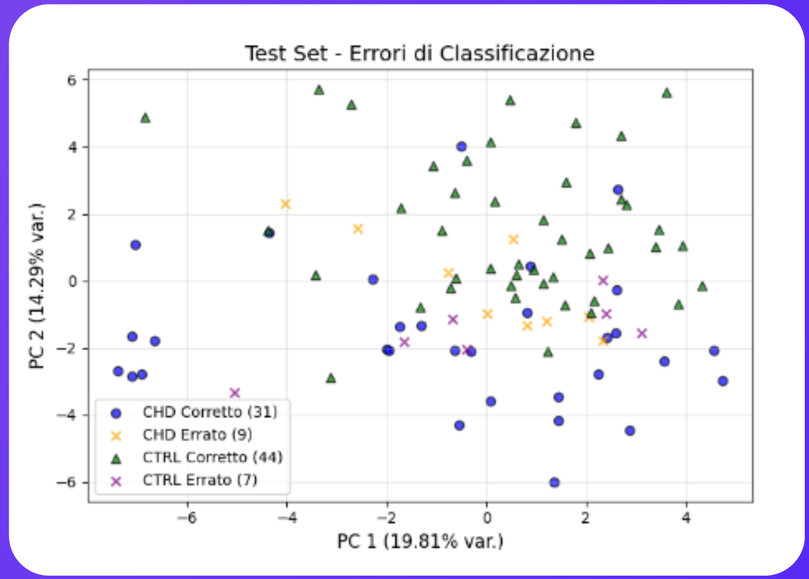
Metric	Grid Search (RBF)	Leave-One-Out (RBF)
Precision (Class CHD)	0.74	0.74
Precision (Class CTRL)	0.87	0.87
Recall (Class CHD)	0.85	0.85
Recall (Class CTRL)	0.76	0.76
F1-Score (Class CHD)	0.79	0.79
F1-Score (Class CTRL)	0.80	0.81
Accuracy	0.80	0.80
Macro Avg F1-Score	0.80	0.80
Weighted Avg F1-Score	0.81	0.80



GRID SEARCH CV



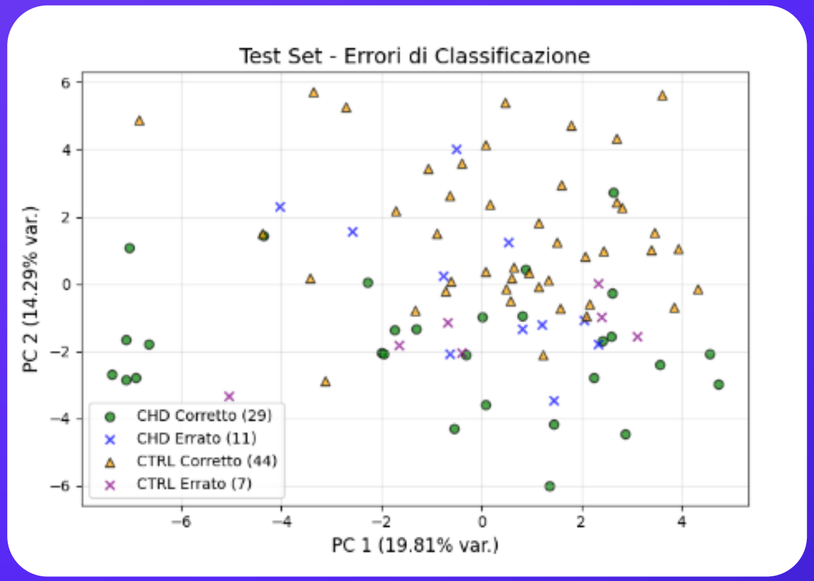
LOOCV



GRID SEARCH CV

RANDOM FOREST

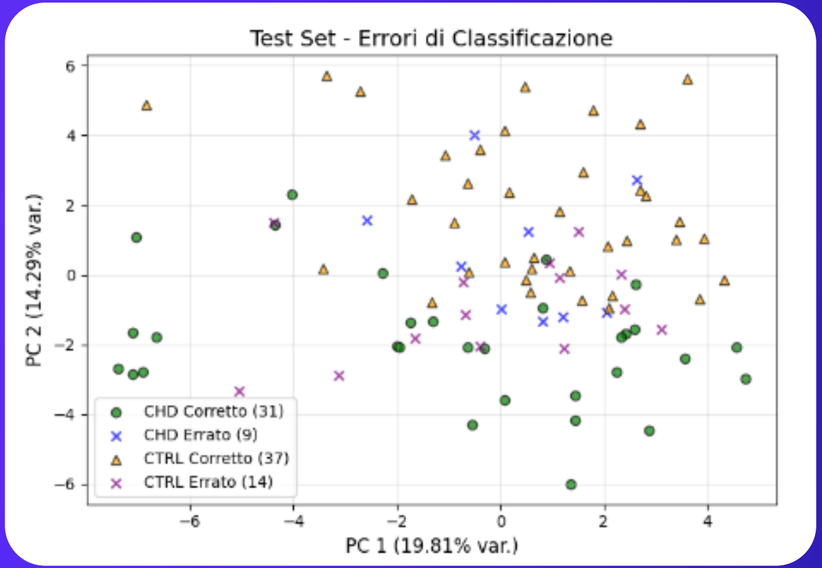
Metric	Grid Search	Leave-One-Out
Precision (Class 0)	0.82	0.82
Precision (Class 1)	0.83	0.83
Recall (Class 0)	0.78	0.78
Recall (Class 1)	0.86	0.86
F1-Score (Class 0)	0.79	0.79
F1-Score (Class 1)	0.85	0.85
Accuracy	0.82	0.82
Macro Avg F1-Score	0.82	0.82
Weighted Avg F1-Score	0.82	0.82



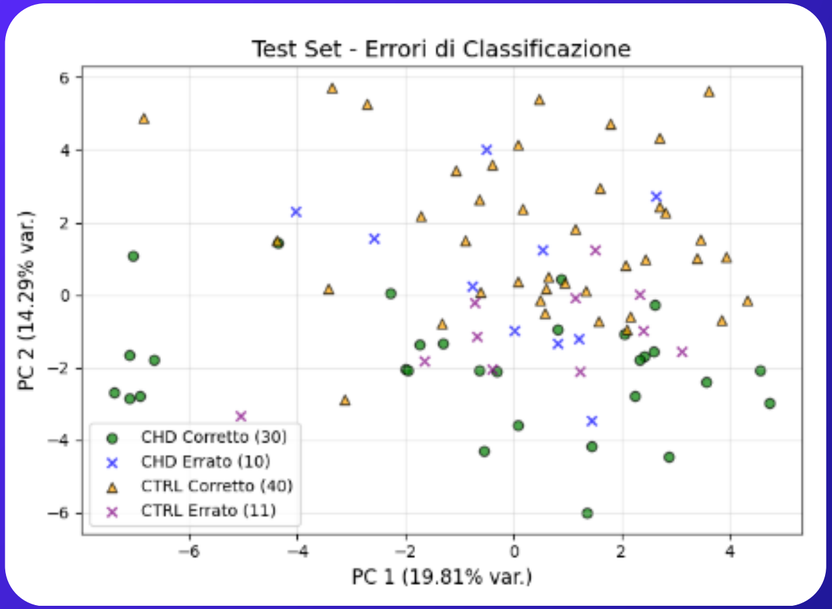
LOOCV

LOGISTIC REGRESSION

Metric	Grid Search (LR)	Leave-One-Out (LR)
Precision (Class CHD)	0.69	0.73
Precision (Class CTRL)	0.80	0.80
Recall (Class CHD)	0.78	0.75
Recall (Class CTRL)	0.73	0.78
F1-Score (Class CHD)	0.73	0.74
F1-Score (Class CTRL)	0.76	0.79
Accuracy	0.75	0.77
Macro Avg F1-Score	0.75	0.77
Weighted Avg F1-Score	0.75	0.77



GRID SEARCH CV

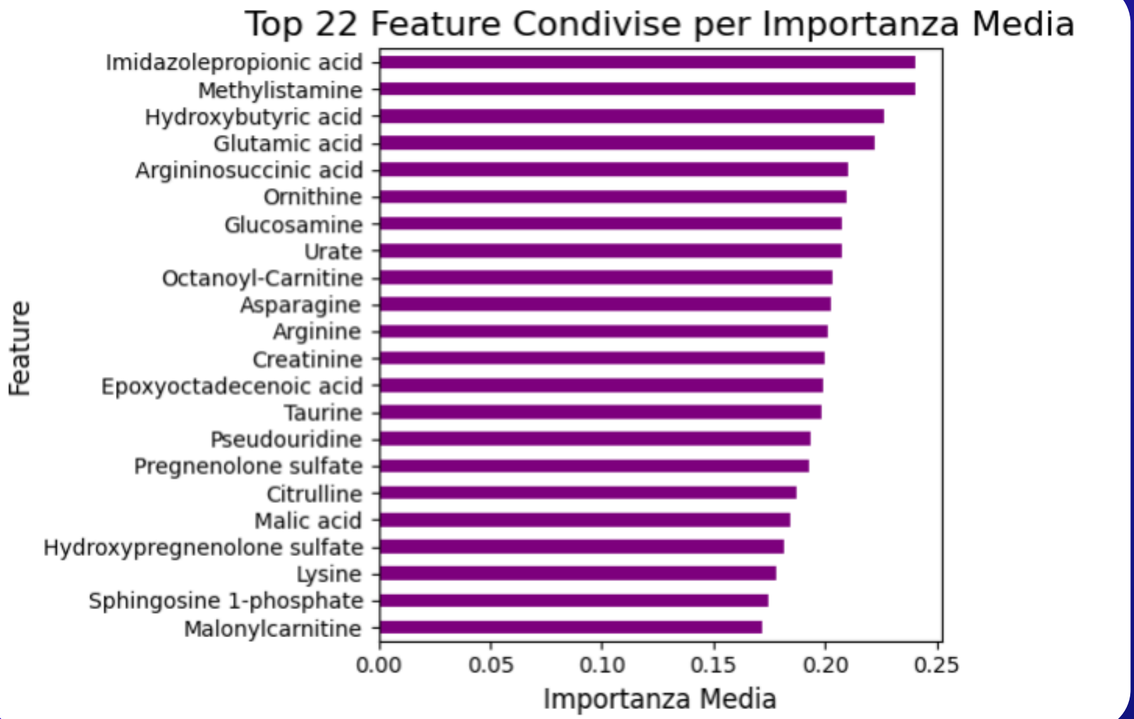


LOOCV

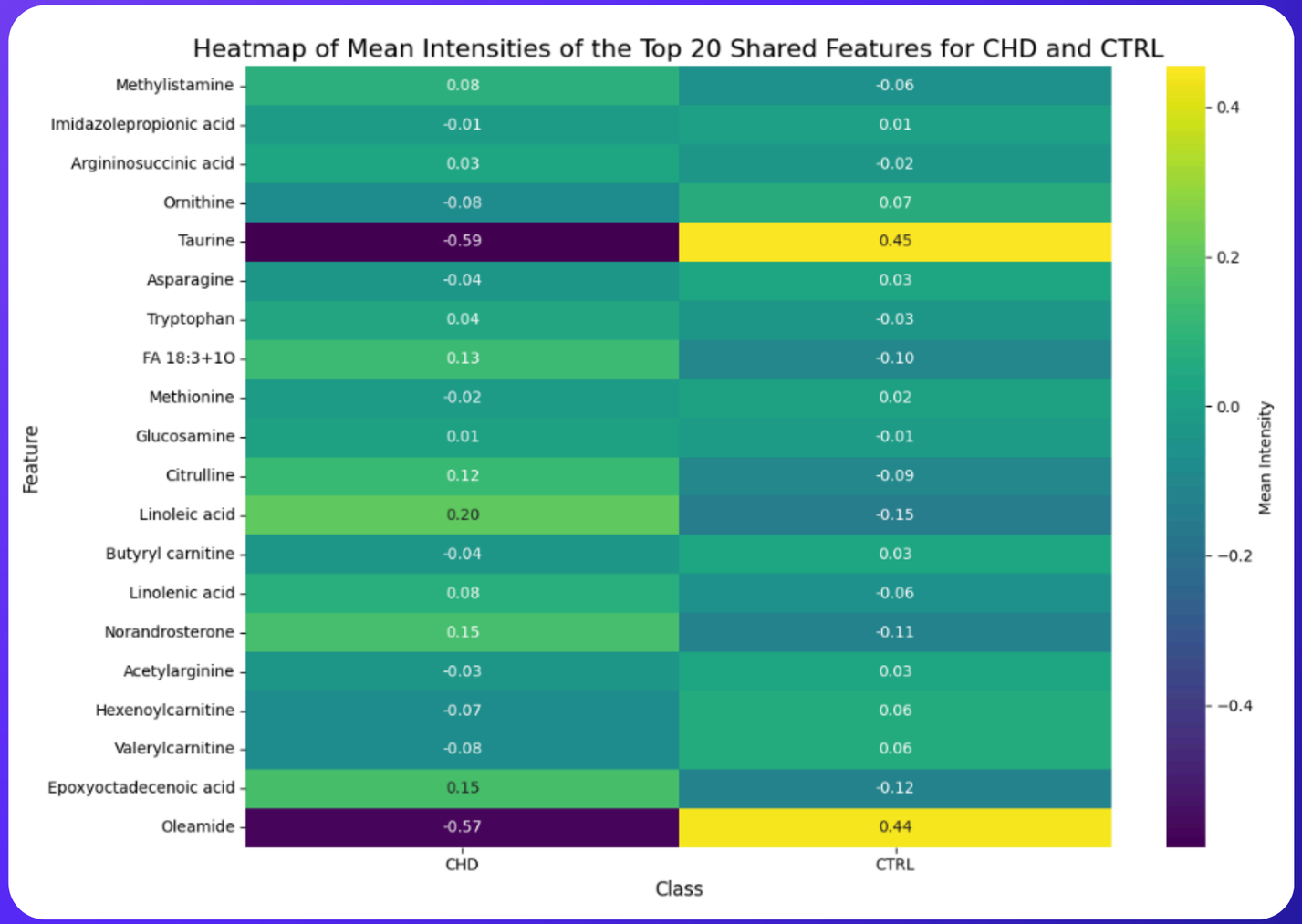
FIND MOST IMPORTANT FEATURES: TECHNIQUES

05

- LOGISTIC REGRESSION: SHAP
- SVM: SHAP
- RANDOM FOREST: FEATURE IMPORTANCES

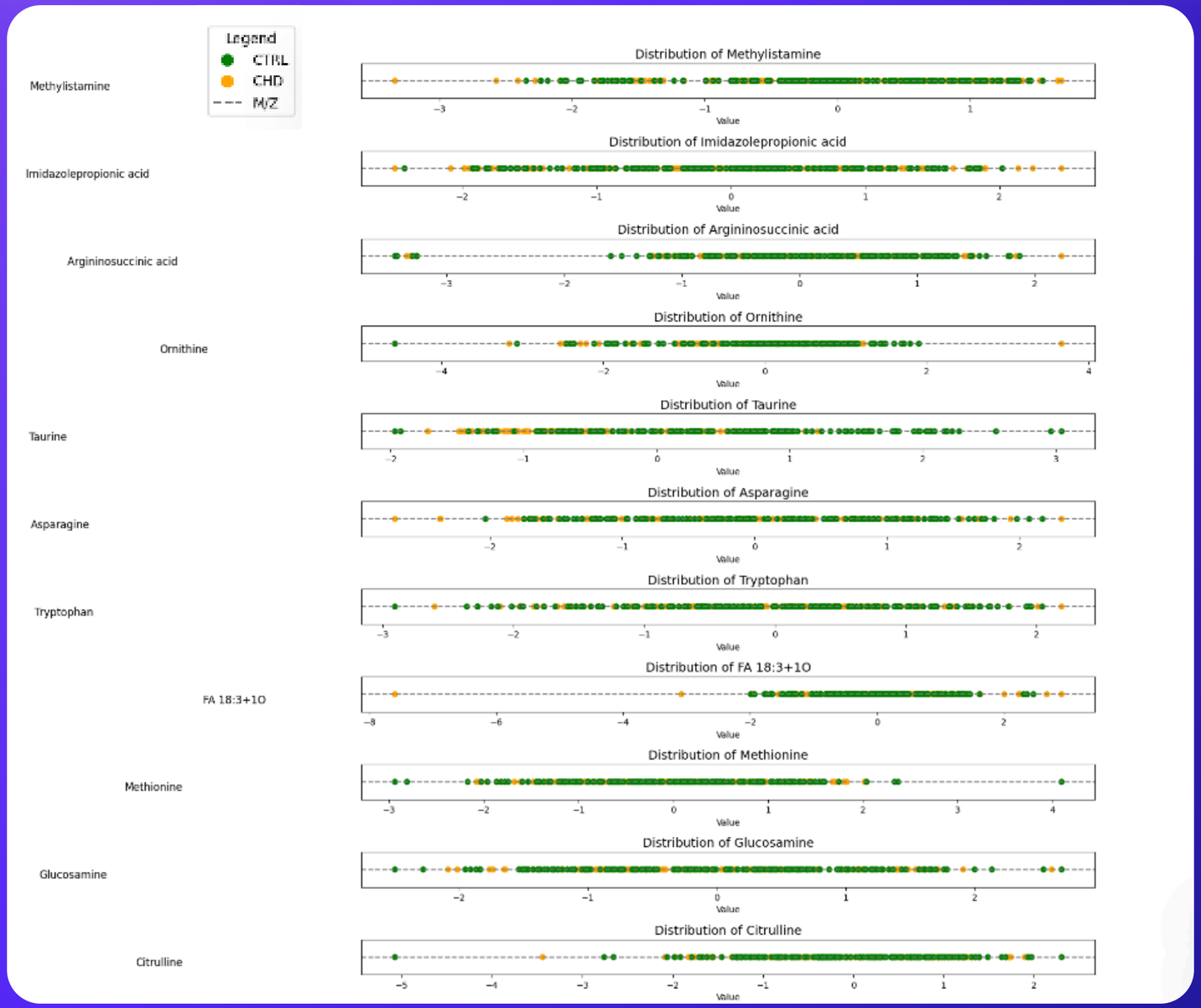


UNIVARIATE ANALYSIS: HEATMAP



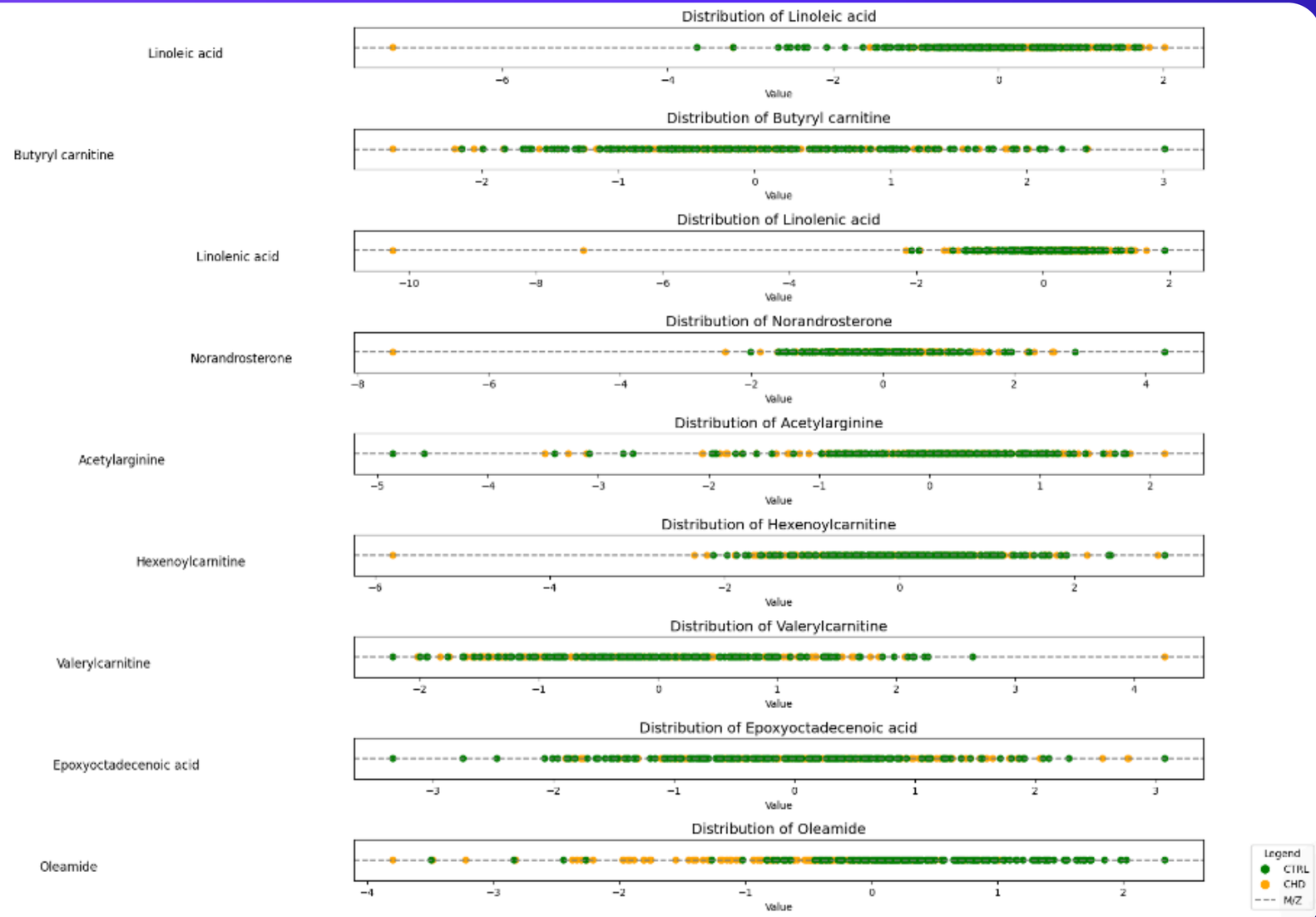
UNIVARIATE ANALYSIS: DISTRIBUTION PLOT

05



UNIVARIATE ANALYSIS: DISTRIBUTION PLOT

05



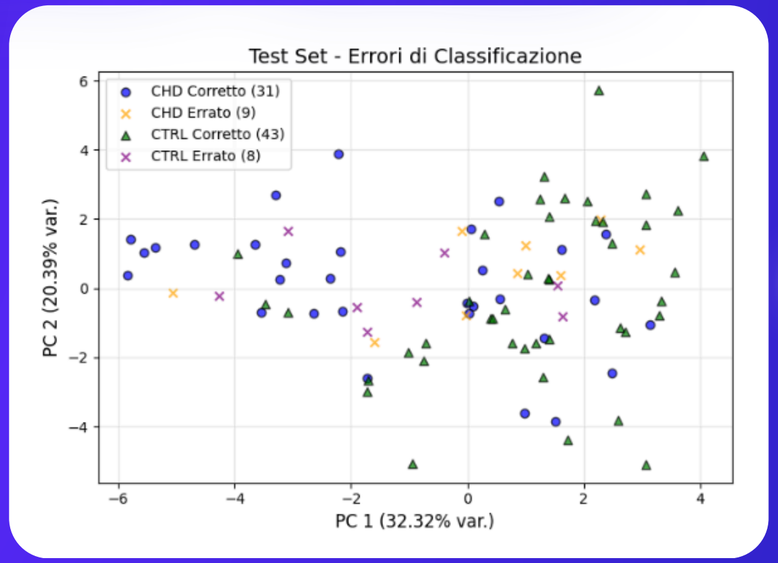
FIT ON MOST IMPORTANT FEATURES

SVM OLD

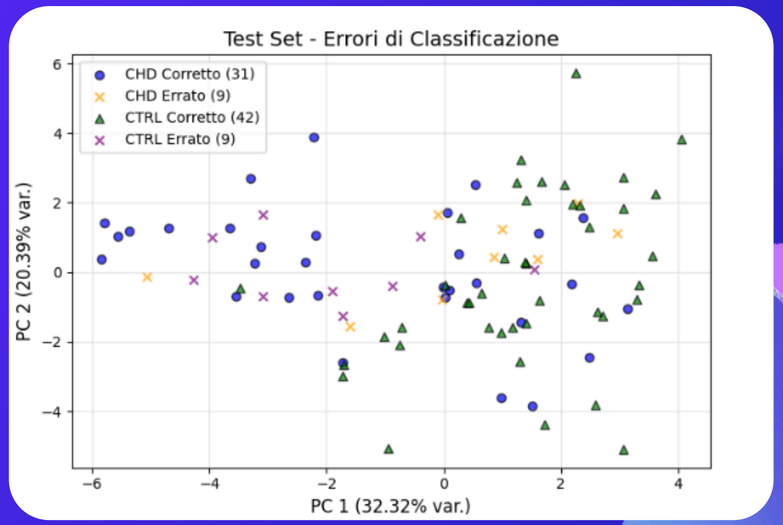
Metric	Grid Search (RBF)	Leave-One-Out (RBF)
Precision (Class CHD)	0.74	0.74
Precision (Class CTRL)	0.87	0.87
Recall (Class CHD)	0.85	0.85
Recall (Class CTRL)	0.76	0.76
F1-Score (Class CHD)	0.79	0.79
F1-Score (Class CTRL)	0.80	0.81
Accuracy	0.80	0.80
Macro Avg F1-Score	0.80	0.80
Weighted Avg F1-Score	0.81	0.80

SVM NEW

Metric	Grid Search (RBF)	Leave-One-Out (RBF)
Precision (Class CHD)	0.79	0.78
Precision (Class CTRL)	0.83	0.82
Recall (Class CHD)	0.78	0.78
Recall (Class CTRL)	0.84	0.82
F1-Score (Class CHD)	0.78	0.78
F1-Score (Class CTRL)	0.83	0.82
Accuracy	0.81	0.80
Macro Avg F1-Score	0.81	0.80
Weighted Avg F1-Score	0.81	0.80



GRID SEARCH CV



LOOCV

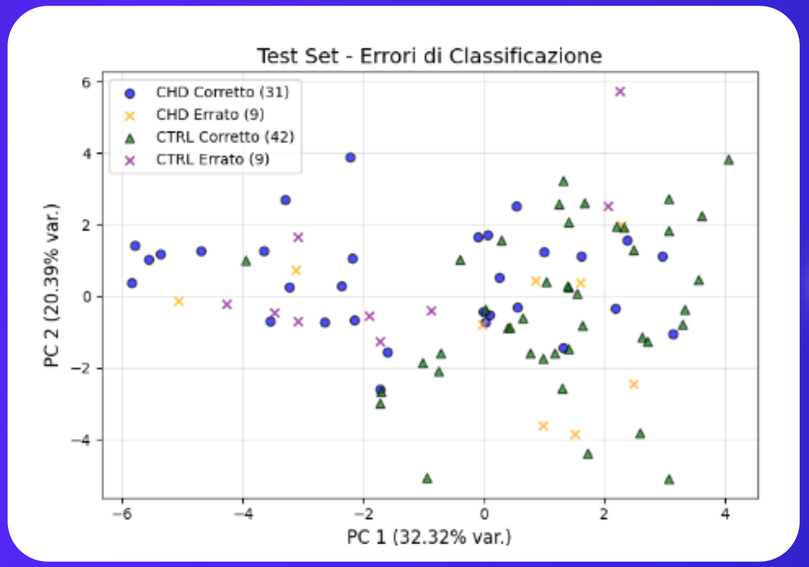
06

RANDOM FOREST OLD

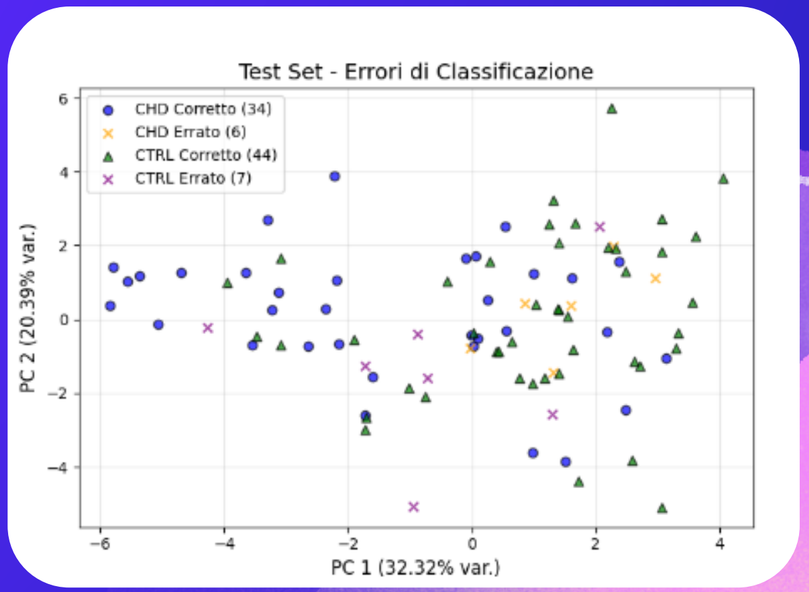
Metric	Grid Search	Leave-One-Out
Precision (Class 0)	0.82	0.82
Precision (Class 1)	0.83	0.83
Recall (Class 0)	0.78	0.78
Recall (Class 1)	0.86	0.86
F1-Score (Class 0)	0.79	0.79
F1-Score (Class 1)	0.85	0.85
Accuracy	0.82	0.82
Macro Avg F1-Score	0.82	0.82
Weighted Avg F1-Score	0.82	0.82

RANDOM FOREST NEW

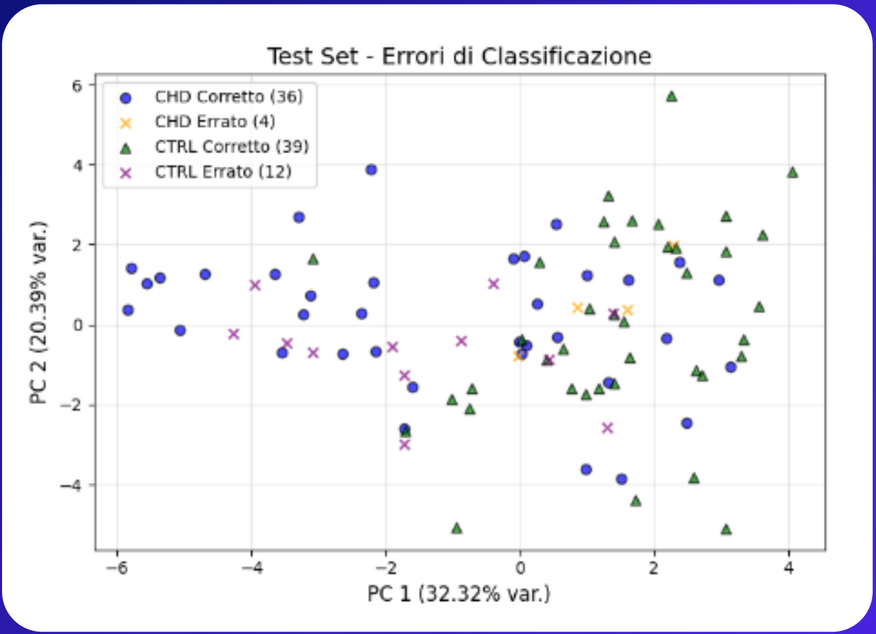
Metric	Grid Search	Leave-One-Out
Precision (Class CHD)	0.78	0.81
Precision (Class CTRL)	0.82	0.80
Recall (Class CHD)	0.78	0.72
Recall (Class CTRL)	0.82	0.86
F1-Score (Class CHD)	0.78	0.76
F1-Score (Class CTRL)	0.83	0.83
Accuracy	0.81	0.80
Macro Avg F1-Score	0.81	0.80
Weighted Avg F1-Score	0.81	0.80



GRID SEARCH CV



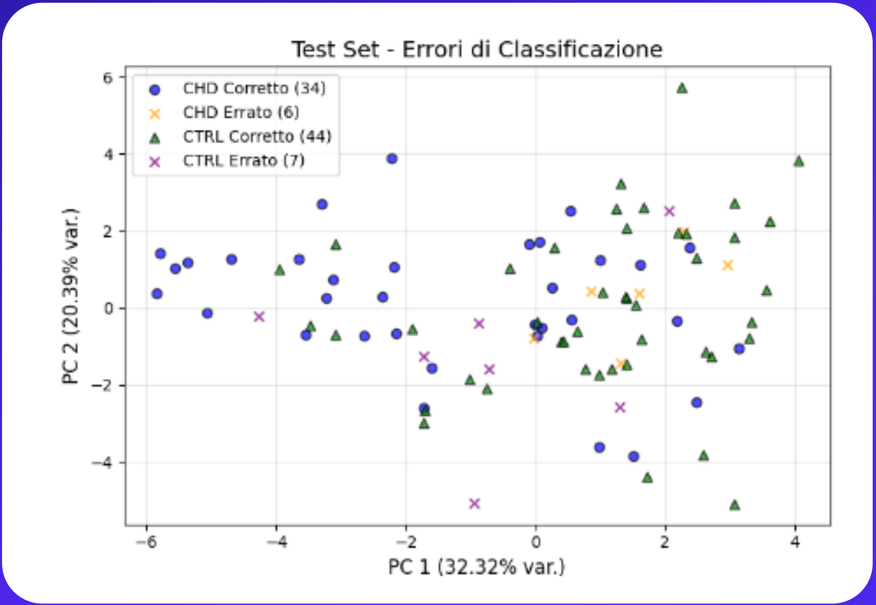
LOOCV



GRID SEARCH CV

LOGISTIC REGRESSION OLD

Metric	Grid Search (LR)	Leave-One-Out (LR)
Precision (Class CHD)	0.69	0.73
Precision (Class CTRL)	0.80	0.80
Recall (Class CHD)	0.78	0.75
Recall (Class CTRL)	0.73	0.78
F1-Score (Class CHD)	0.73	0.74
F1-Score (Class CTRL)	0.76	0.79
Accuracy	0.75	0.77
Macro Avg F1-Score	0.75	0.77
Weighted Avg F1-Score	0.75	0.77



LOOCV

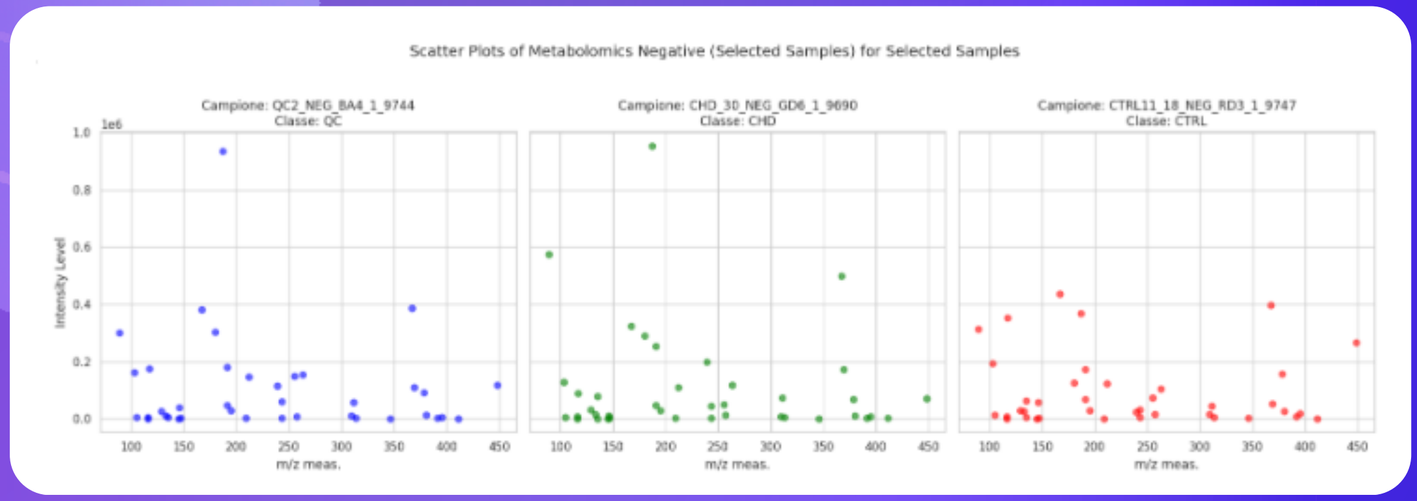
LOGISTIC REGRESSION NEW

Metric	Grid Search	Leave-One-Out
Precision (Class CHD)	0.75	0.83
Precision (Class CTRL)	0.91	0.88
Recall (Class CHD)	0.90	0.85
Recall (Class CTRL)	0.76	0.86
F1-Score (Class CHD)	0.82	0.84
F1-Score (Class CTRL)	0.83	0.87
Accuracy	0.82	0.86
Macro Avg F1-Score	0.83	0.86
Weighted Avg F1-Score	0.84	0.86

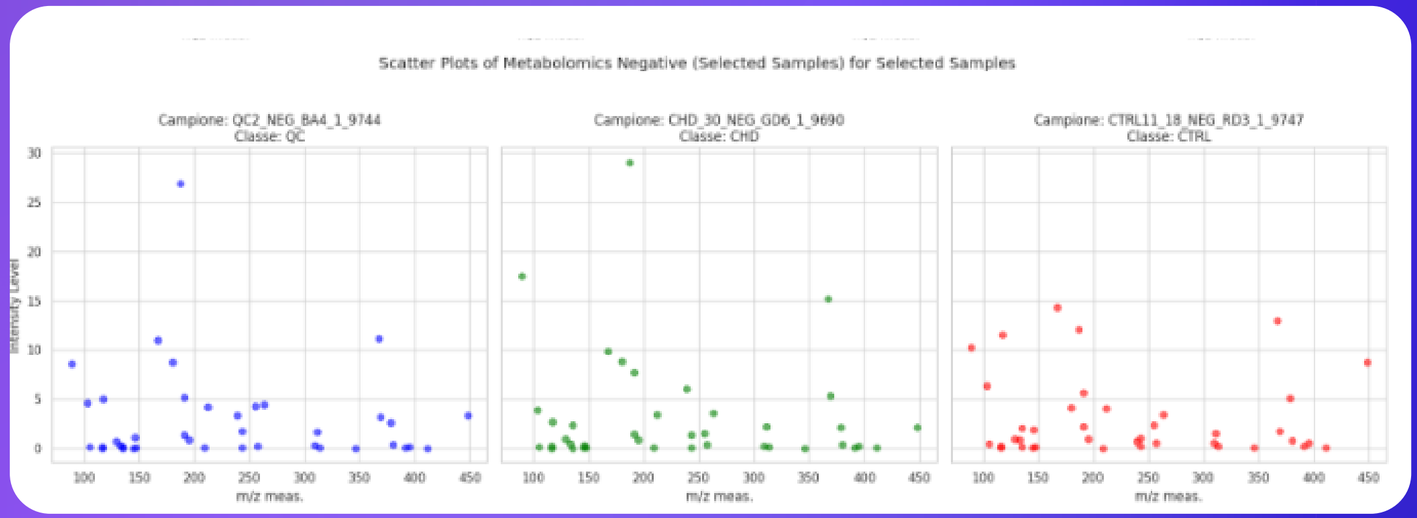
07

OTHER NORMALIZATION METHODS:TIC DATA VISUALIZATION

PRE NORM ESI-

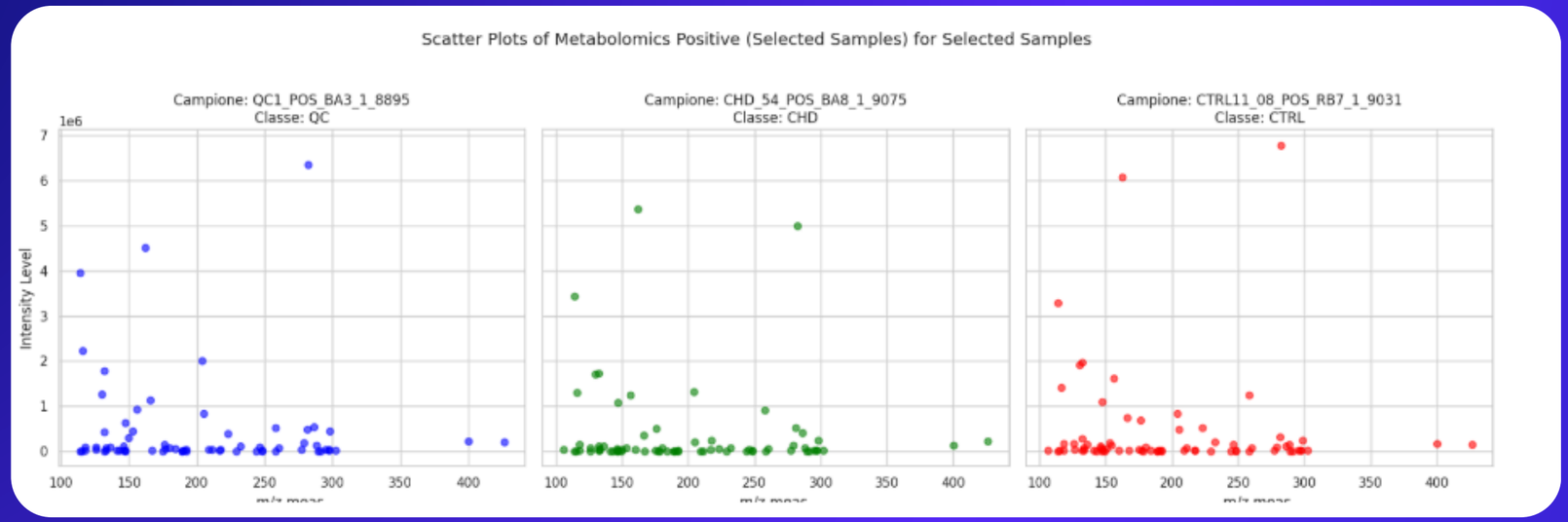


POST NORM ESI-

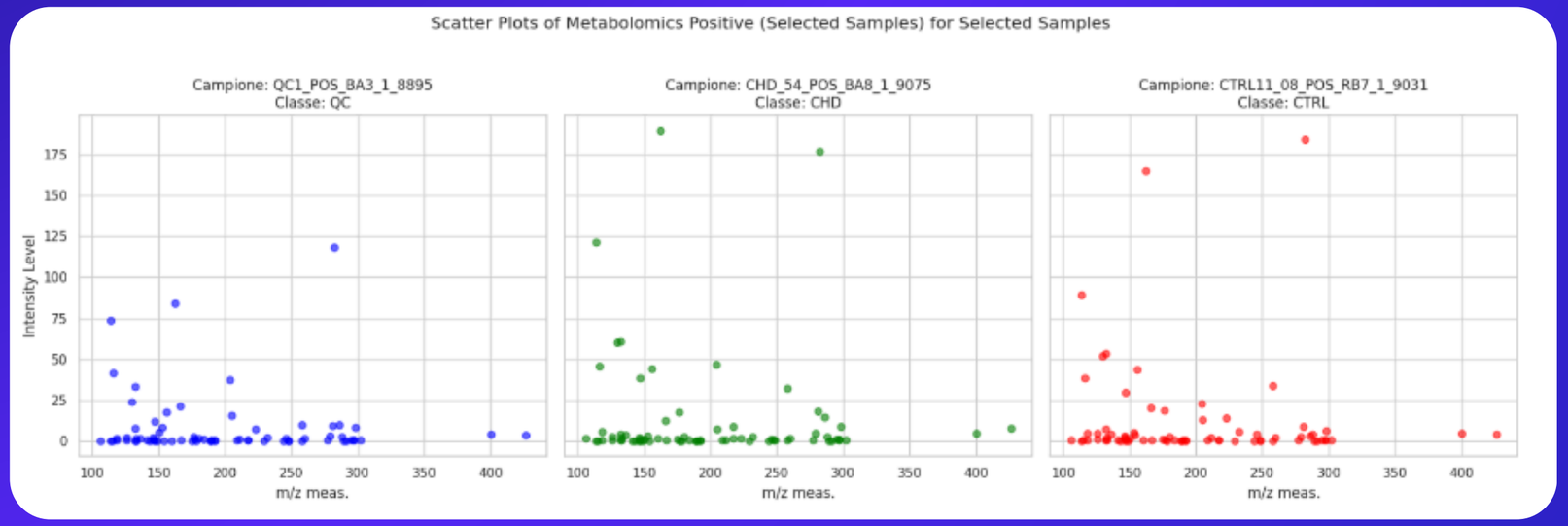


07

PRE NORM ESI+

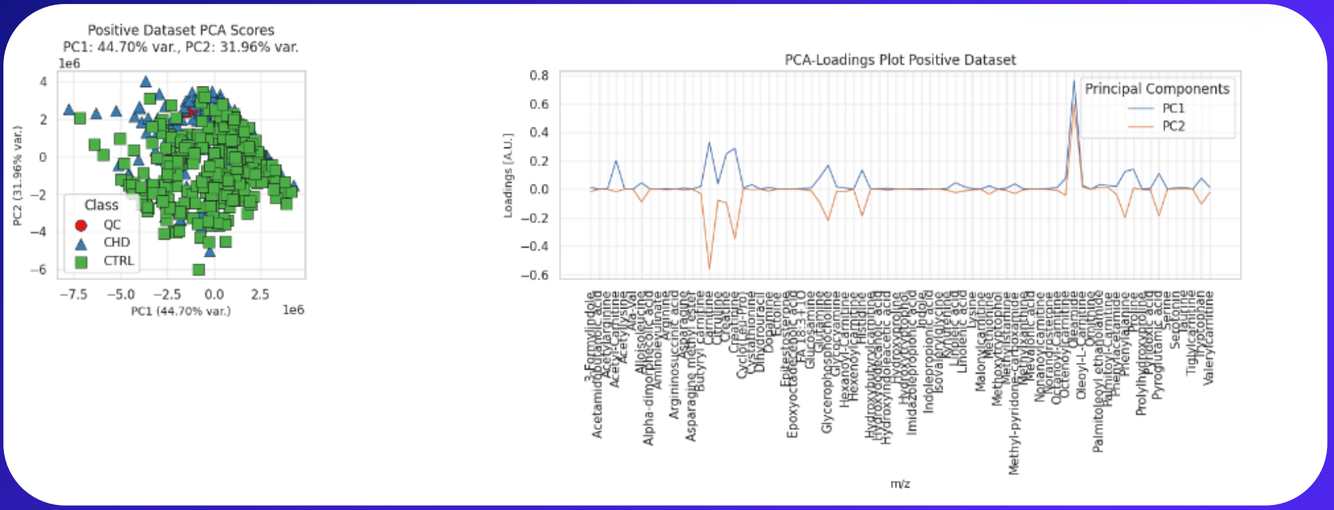


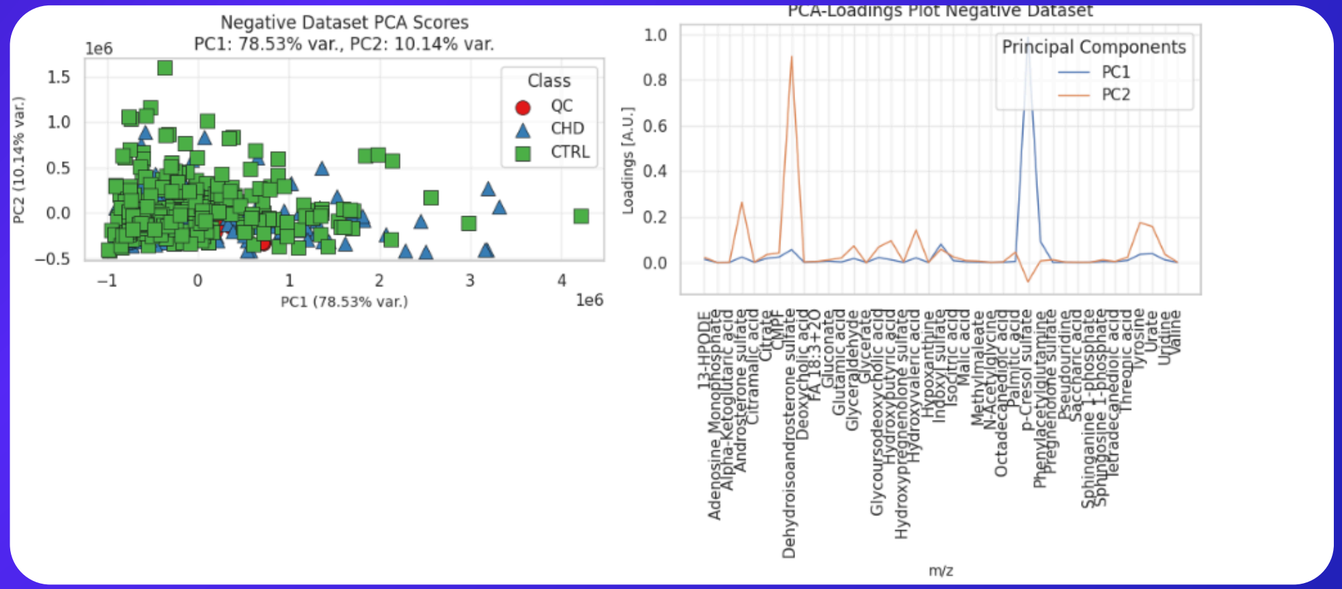
POST NORM ESI+



PCA VIEW

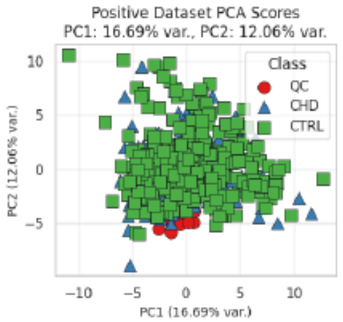
07

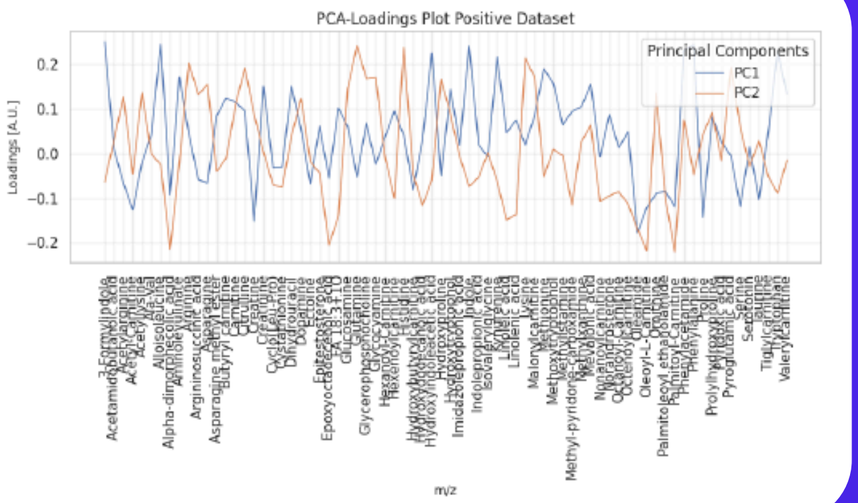

Positive Dataset PCA Scores
PC1: 44.70% var., PC2: 31.96% var.
1e6
PC2 (31.96% var.)
PC1 (44.70% var.)
1e6
Class
QC
CHD
CTRL
PCA-Loadings Plot Positive Dataset
Loadings [A.U.]
Principal Components
PC1
PC2
m/z

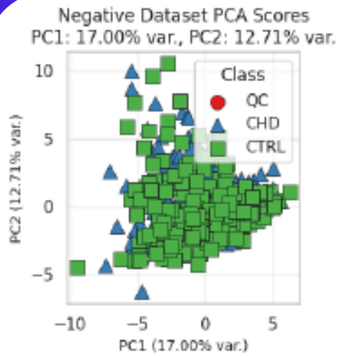

Negative Dataset PCA Scores
PC1: 78.53% var., PC2: 10.14% var.
1e6
PC2 (10.14% var.)
PC1 (78.53% var.)
1e6
Class
QC
CHD
CTRL
PCA-Loadings Plot Negative Dataset
Loadings [A.U.]
Principal Components
PC1
PC2
m/z

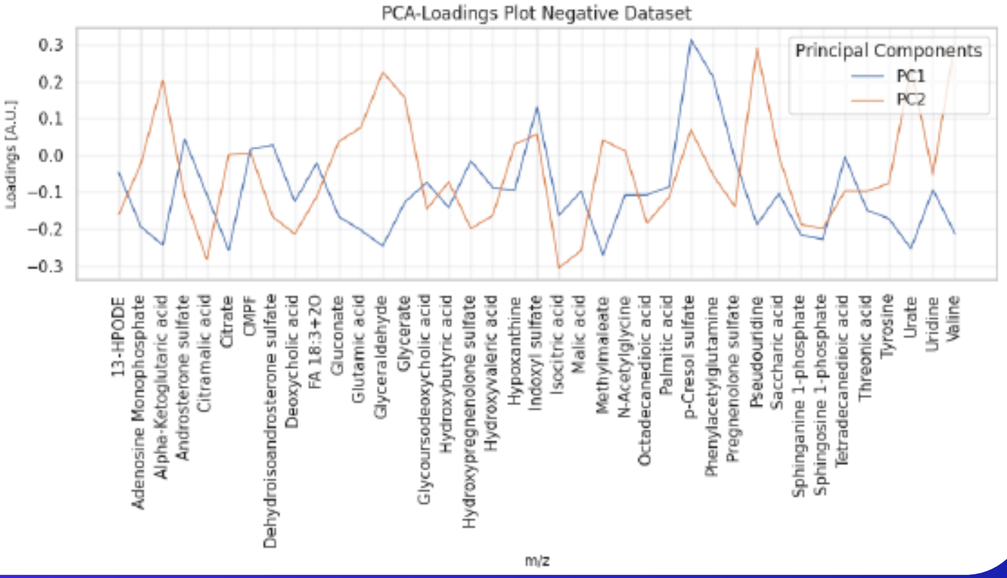
PCA VIEW

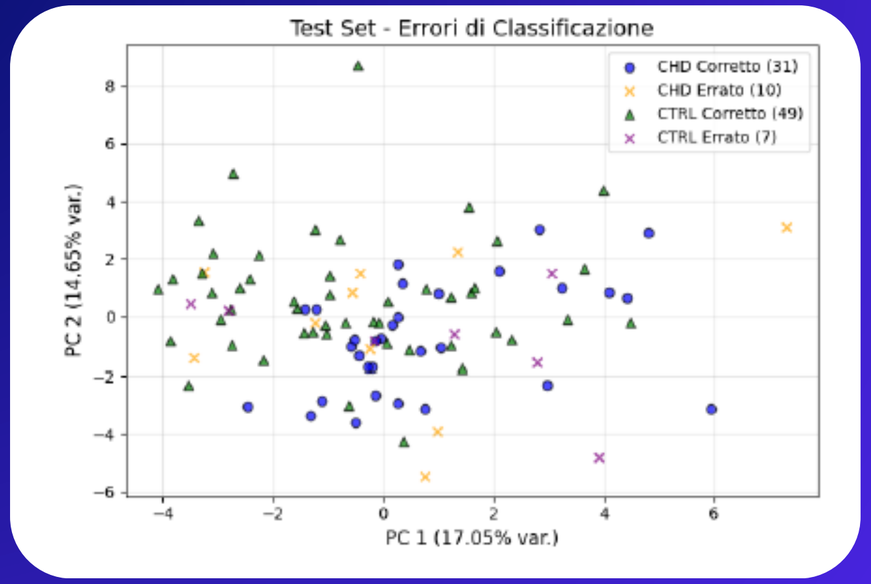
07


Positive Dataset PCA Scores
PC1: 16.69% var., PC2: 12.06% var.
Class
QC
CHD
CTRL
PC2 (12.06% var.)
PC1 (16.69% var.)


PCA-Loadings Plot Positive Dataset
Principal Components
PC1
PC2
Loadings [A.U.]
m/z


Negative Dataset PCA Scores
PC1: 17.00% var., PC2: 12.71% var.
Class
QC
CHD
CTRL
PC2 (12.71% var.)
PC1 (17.00% var.)

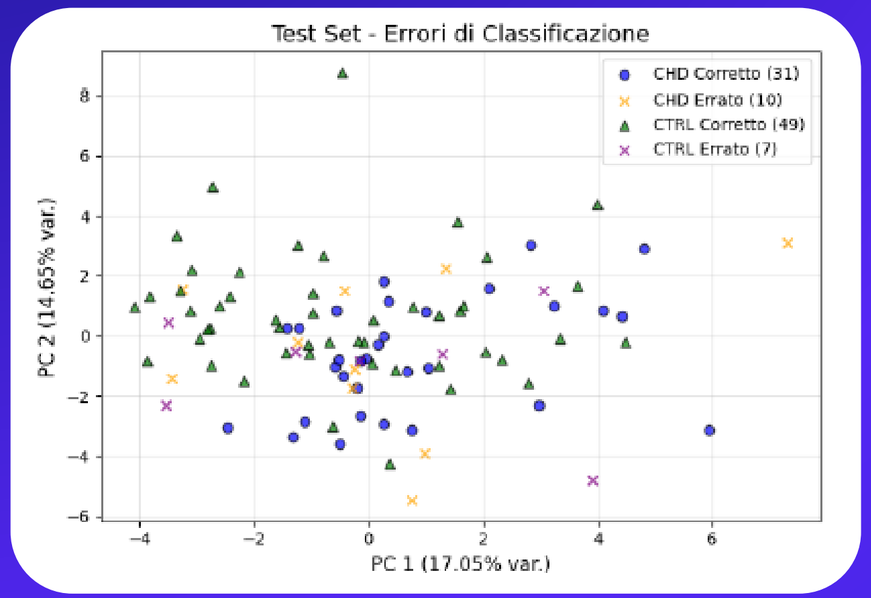

PCA-Loadings Plot Negative Dataset
Principal Components
PC1
PC2
Loadings [A.U.]
13-HPODE
Adenosine Monophosphate
Alpha-Ketoglutaric acid
Androsterone sulfate
Citramalic acid
Citrate
CMPF
Dehydroisoandrosterone sulfate
Deoxycholic acid
FA 18:3+2O
Gluconate
Glutamic acid
Glyceraldehyde
Glycerate
Glycoursodeoxycholic acid
Hydroxybutyric acid
Hydroxypregnenolone sulfate
Hydroxyvaleric acid
Hypoxanthine
Indoxyl sulfate
Isocitric acid
Malic acid
Methylmaleate
N-Acetylglycine
Octadecanedioic acid
Palmitic acid
p-Cresol sulfate
Phenylacetylglutamine
Pregnenolone sulfate
Pseudouridine
Saccharic acid
Sphinganine 1-phosphate
Sphingosine 1-phosphate
Tetradecanedioic acid
Threonic acid
Tyrosine
Urate
Uridine
Valine
m/z



GRID SEARCH CV

SVM OLD

Metric	Grid Search	Leave-One-Out
Precision (CHD)	0.86	0.89
Precision (CTRL)	0.0.85	0.85
Recall (CHD)	0.78	0.78
Recall (CTRL)	0.91	0.93
F1-Score (CHD)	0.82	0.83
F1-Score (CTRL)	0.88	0.89
Accuracy	0.86	0.87
Macro Avg F1-Score	0.85	0.86
Weighted Avg F1-Score	0.85	0.86

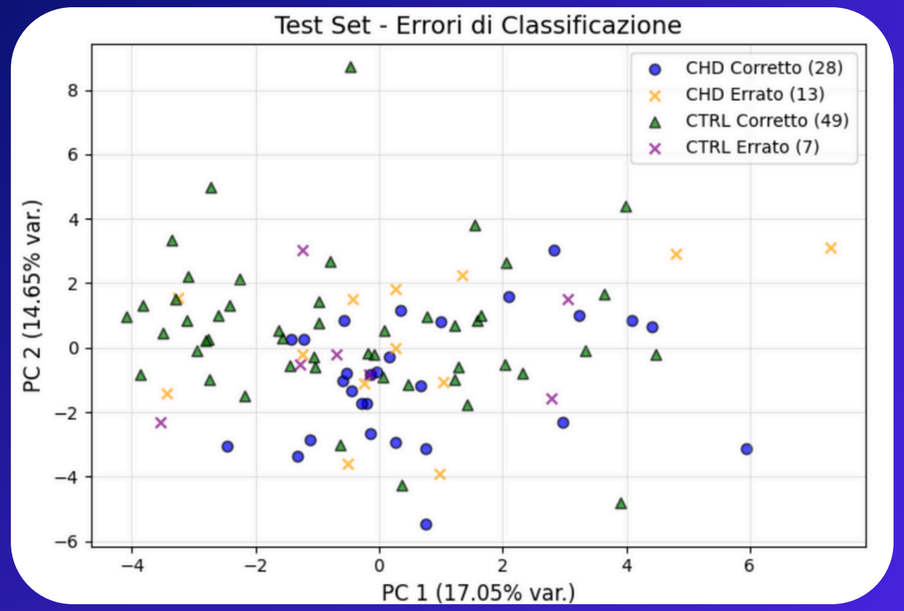


LOOCV

SVM NEW

Metric	Grid Search (RBF)	Leave-One-Out (RBF)
Precision (Class CHD)	0.82	0.82
Precision (Class CTRL)	0.83	0.83
Recall (Class CHD)	0.76	0.76
Recall (Class CTRL)	0.88	0.88
F1-Score (Class CHD)	0.78	0.78
F1-Score (Class CTRL)	0.85	0.85
Accuracy	0.82	0.82
Macro Avg F1-Score	0.82	0.82
Weighted Avg F1-Score	0.82	0.82

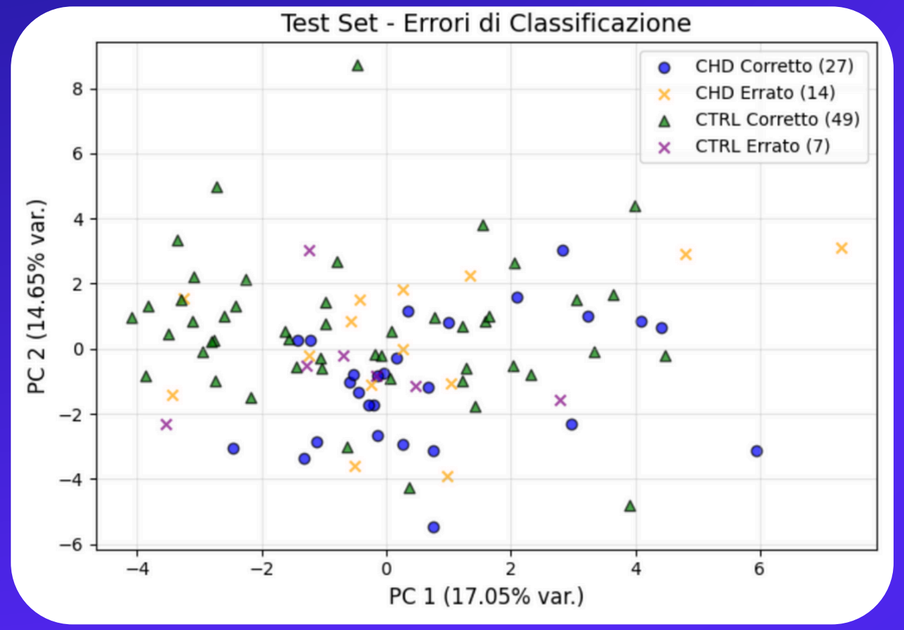
07



GRID SEARCH CV

RANDOM FOREST OLD

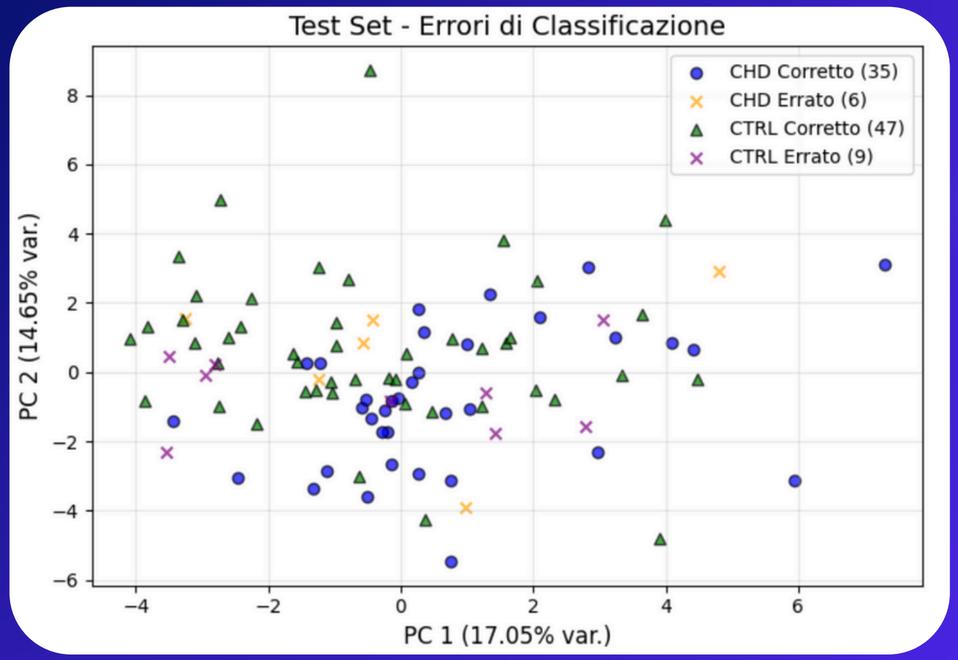
Metric	Grid Search	Leave-One-Out
Precision (CHD)	0.77	0.77
Precision (CTRL)	0.81	0.81
Recall (CHD)	0.73	0.73
Recall (CTRL)	0.84	0.84
F1-Score (CHD)	0.75	0.75
F1-Score (CTRL)	0.82	0.82
Accuracy	0.79	0.79
Macro Avg F1-Score	0.79	0.79
Weighted Avg F1-Score	0.79	0.79



LOOCV

RANDOM FOREST NEW

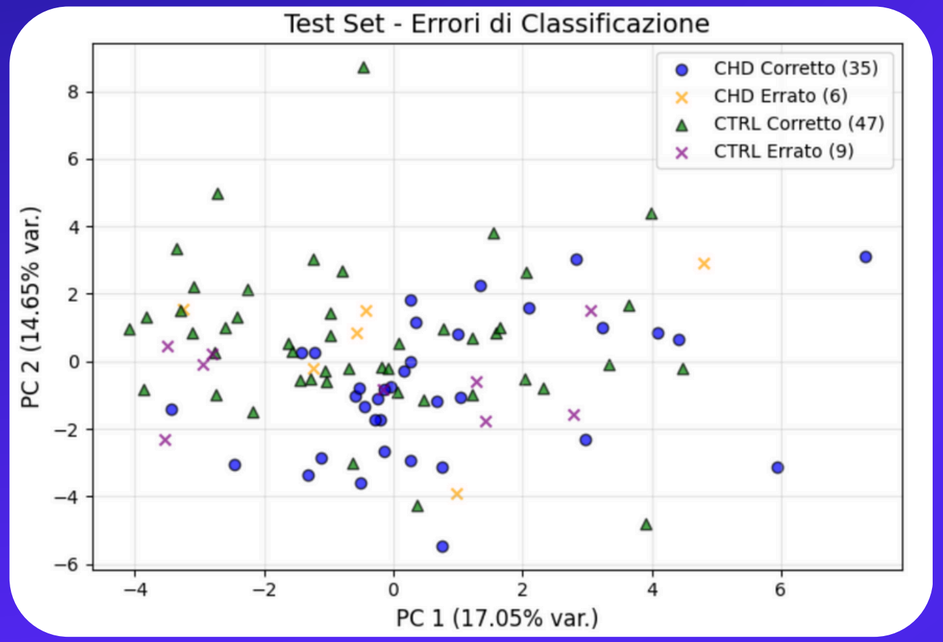
Metric	Grid Search	Leave-One-Out
Precision (Class 0)	0.80	0.79
Precision (Class 1)	0.79	0.78
Recall (Class 0)	0.68	0.66
Recall (Class 1)	0.88	0.88
F1-Score (Class 0)	0.74	0.72
F1-Score (Class 1)	0.83	0.82
Accuracy	0.79	0.78
Macro Avg F1-Score	0.78	0.77
Weighted Avg F1-Score	0.79	0.78



GRID SEARCH CV

LOGISTIC REGRESSION OLD

Metric	Grid Search	Leave-One-Out
Precision (CHD)	0.78	0.78
Precision (CTRL)	0.0.80	0.80
Recall (CHD)	0.71	0.71
Recall (CTRL)	0.86	0.86
F1-Score (CHD)	0.74	0.74
F1-Score (CTRL)	0.83	0.83
Accuracy	0.79	0.79
Macro Avg F1-Score	0.79	0.79
Weighted Avg F1-Score	0.79	0.79



LOOCV

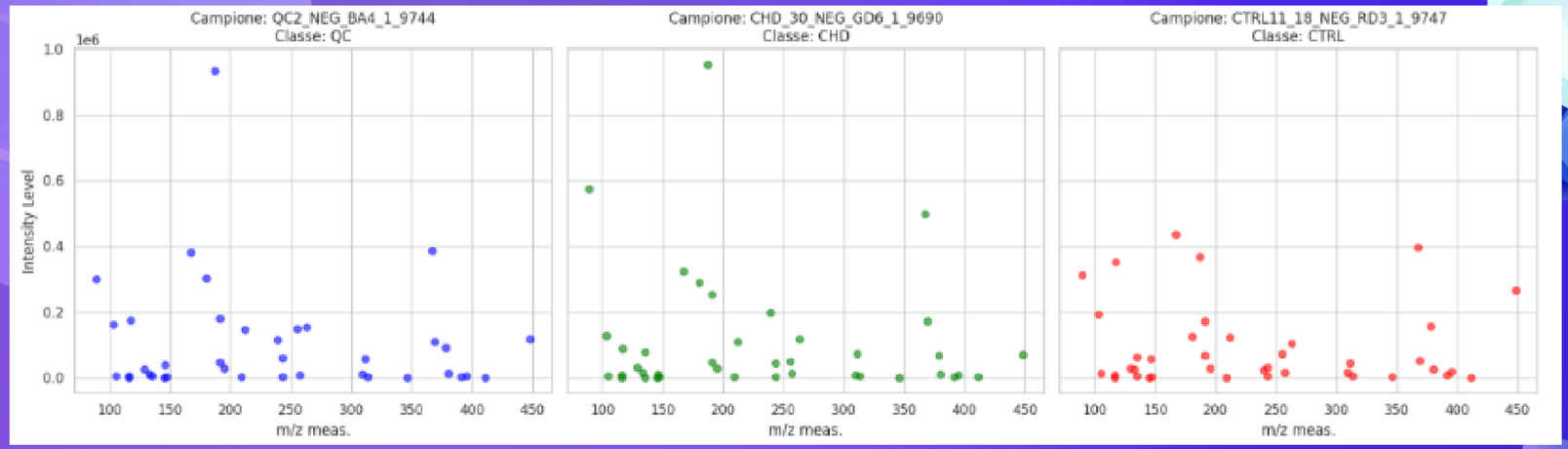
LOGISTIC REGRESSION NEW

Metric	Grid Search	Leave-One-Out
Precision (Class 0)	0.80	0.80
Precision (Class 1)	0.89	0.89
Recall (Class 0)	0.85	0.85
Recall (Class 1)	0.84	0.84
F1-Score (Class 0)	0.82	0.82
F1-Score (Class 1)	0.86	0.86
Accuracy	0.85	0.85
Macro Avg F1-Score	0.84	0.84
Weighted Avg F1-Score	0.85	0.85

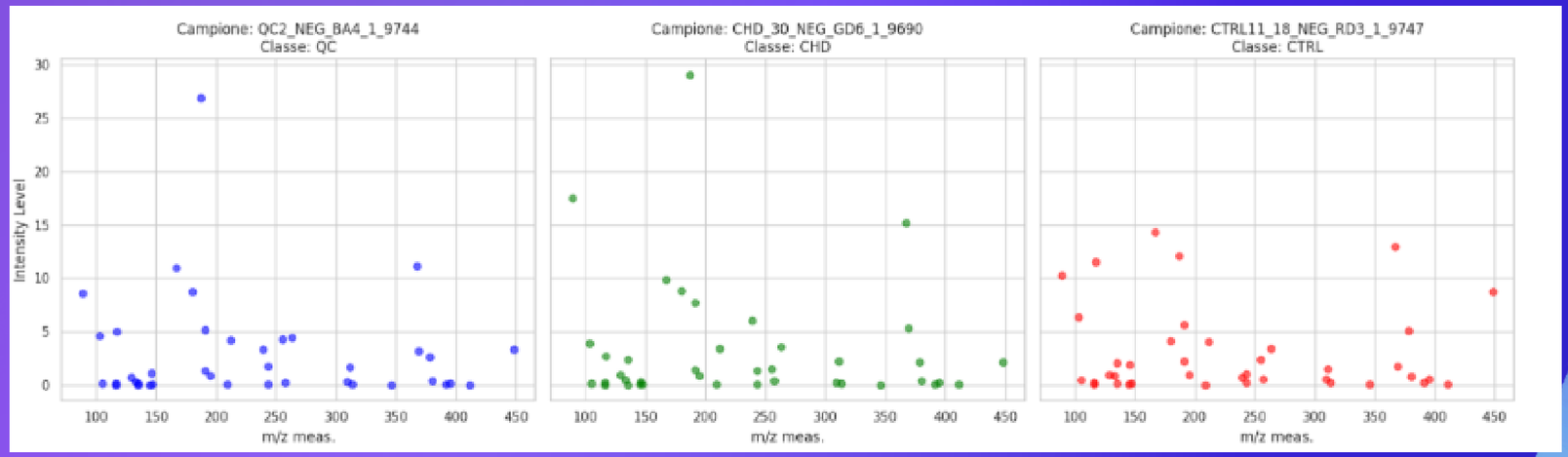
07

OTHER NORMALIZATION METHODS:MEDIAN DATA VISUALIZATION

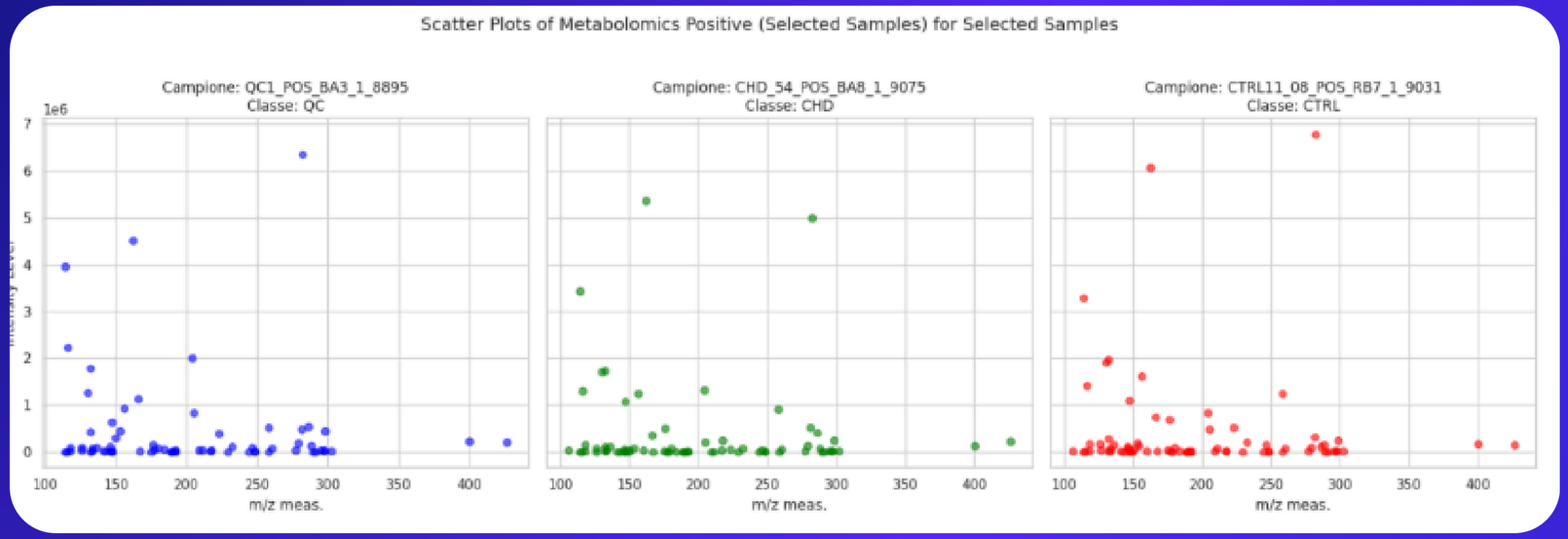
PRE NORM ESI-



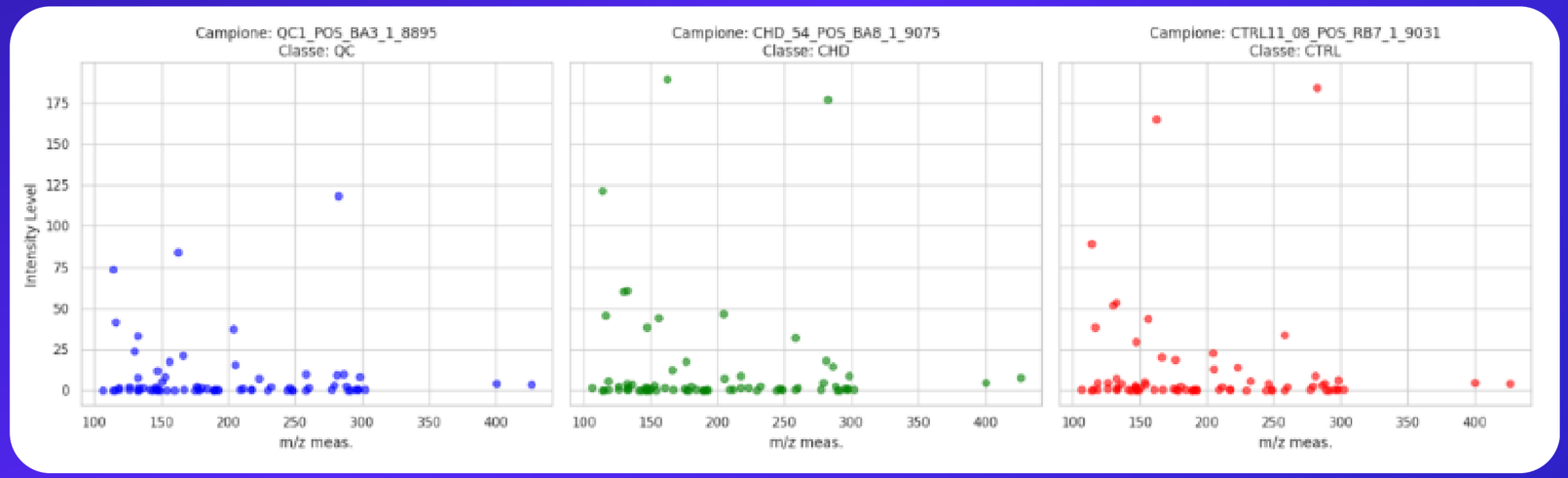
POST NORM ESI-



PRE NORM ESI-

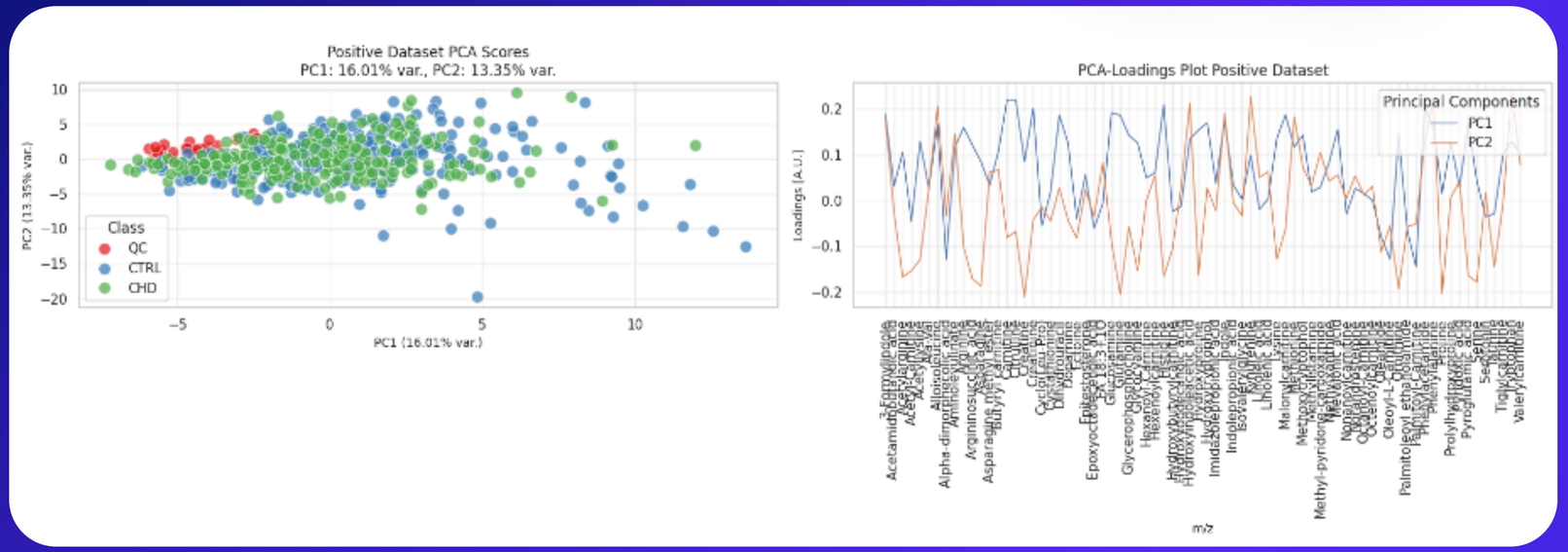


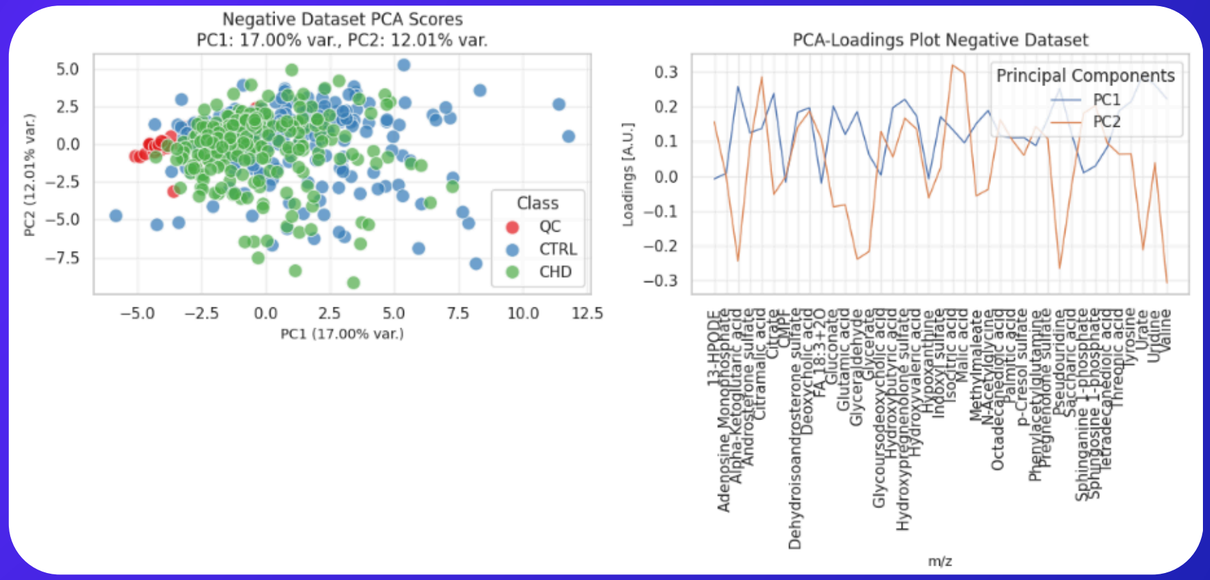
POST NORM ESI+

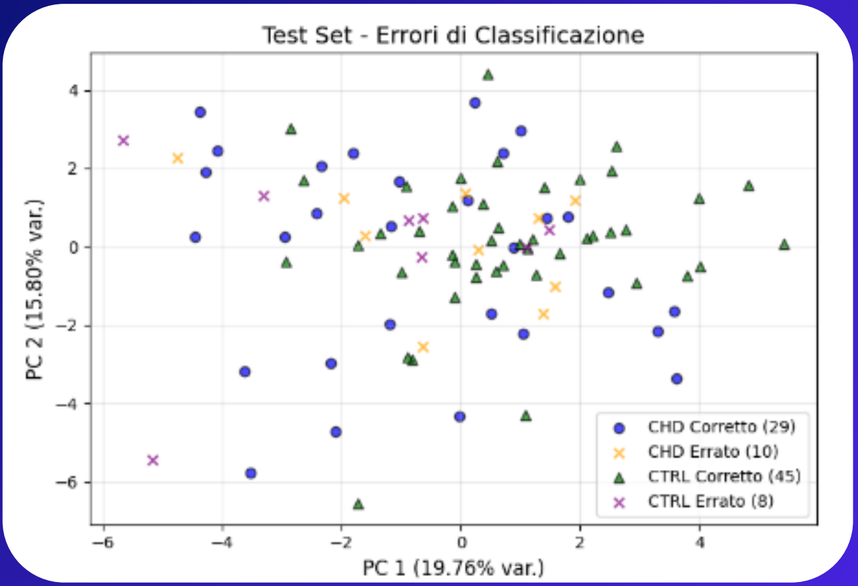


PCA VIEW

07


Positive Dataset PCA Scores
PC1: 16.01% var., PC2: 13.35% var.
PC2 (13.35% var.)
PC1 (16.01% var.)
Class
QC
CTRL
CHD
PCA-Loadings Plot Positive Dataset
Principal Components
PC1
PC2
Loadings [A.U.]
m/z

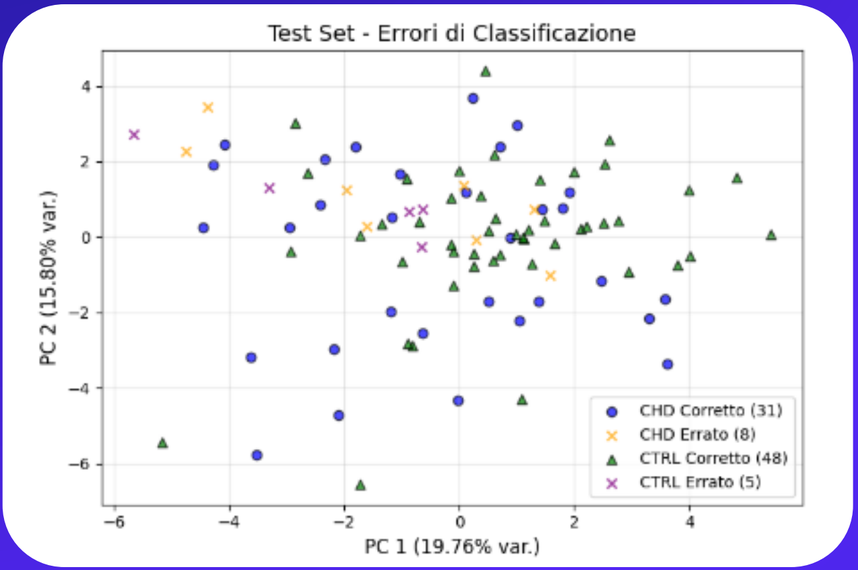

Negative Dataset PCA Scores
PC1: 17.00% var., PC2: 12.01% var.
PC2 (12.01% var.)
PC1 (17.00% var.)
Class
QC
CTRL
CHD
PCA-Loadings Plot Negative Dataset
Principal Components
PC1
PC2
Loadings [A.U.]
m/z



GRID SEARCH CV

SVM OLD

Metric	Grid Search	Leave-One-Out
Precision (CHD)	0.83	0.86
Precision (CTRL)	0.84	0.84
Recall (CHD)	0.77	0.77
Recall (CTRL)	0.89	0.91
F1-Score (CHD)	0.80	0.81
F1-Score (CTRL)	0.86	0.87
Accuracy	0.84	0.85
Macro Avg F1-Score	0.83	0.84
Weighted Avg F1-Score	0.84	0.85

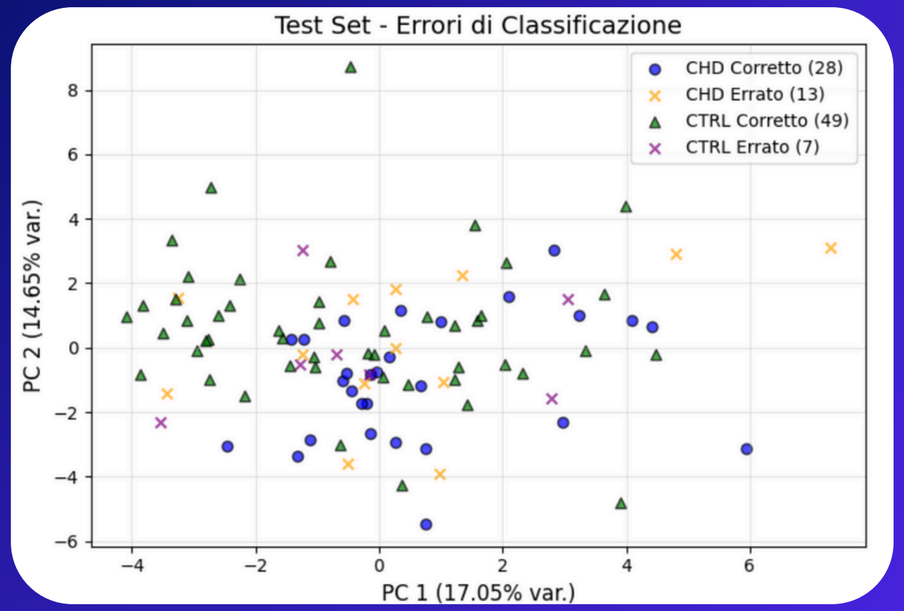


LOOCV

SVM NEW

Metric	Grid Search	Leave-One-Out
Precision (CHD)	0.78	0.86
Precision (CTRL)	0.82	0.86
Recall (CHD)	0.74	0.79
Recall (CTRL)	0.85	0.91
F1-Score (CHD)	0.76	0.83
F1-Score (CTRL)	0.83	0.88
Accuracy	0.80	0.86
Macro Avg F1-Score	0.80	0.85
Weighted Avg F1-Score	0.80	0.86

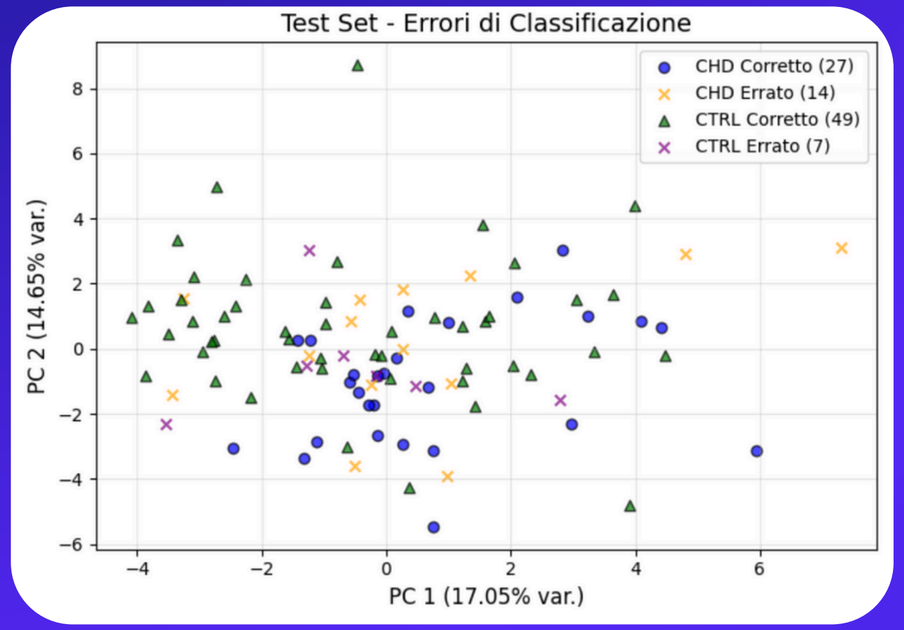
07



GRID SEARCH CV

RANDOM FOREST OLD

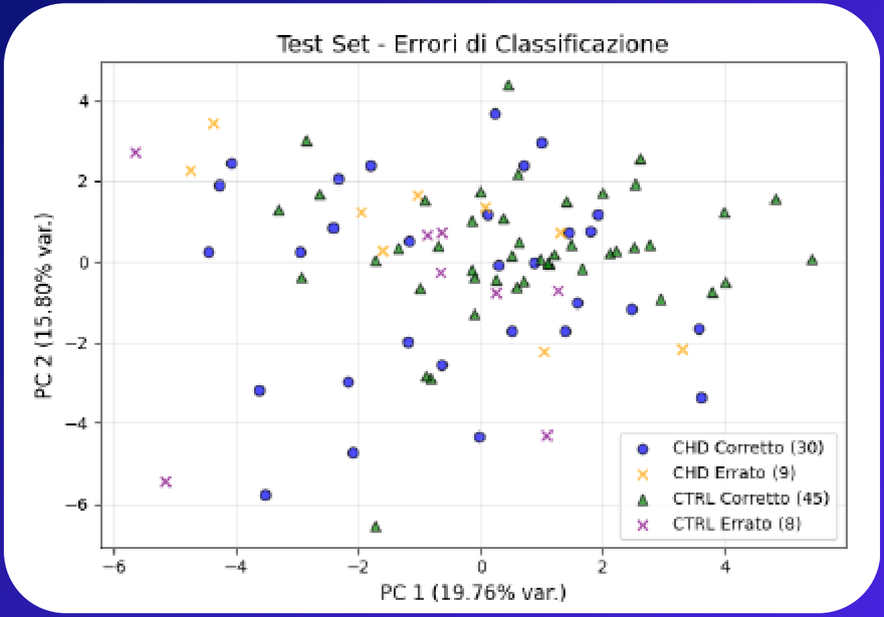
Metric	Grid Search	Leave-One-Out
Precision (CHD)	0.84	0.88
Precision (CTRL)	0.79	0.82
Recall (CHD)	0.67	0.72
Recall (CTRL)	0.91	0.92
F1-Score (CHD)	0.74	0.79
F1-Score (CTRL)	0.84	0.87
Accuracy	0.80	0.84
Macro Avg F1-Score	0.79	0.83
Weighted Avg F1-Score	0.80	0.83



LOOCV

RANDOM FOREST NEW

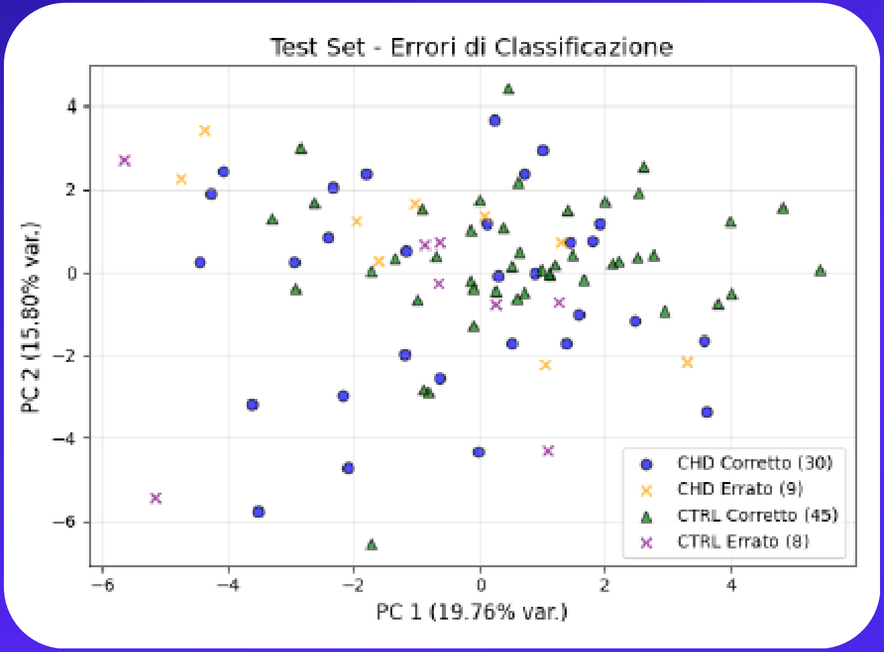
Metric	Grid Search	Leave-One-Out
Precision (CHD)	0.78	0.86
Precision (CTRL)	0.82	0.86
Recall (CHD)	0.74	0.79
Recall (CTRL)	0.85	0.91
F1-Score (CHD)	0.76	0.83
F1-Score (CTRL)	0.83	0.88
Accuracy	0.80	0.86
Macro Avg F1-Score	0.80	0.85
Weighted Avg F1-Score	0.80	0.86



GRID SEARCH CV

LOGISTIC REGRESSION OLD

Metric	Grid Search	Leave-One-Out
Precision (CHD)	0.84	0.86
Precision (CTRL)	0.80	0.86
Recall (CHD)	0.69	0.79
Recall (CTRL)	0.91	0.91
F1-Score (CHD)	0.76	0.83
F1-Score (CTRL)	0.85	0.88
Accuracy	0.82	0.86
Macro Avg F1-Score	0.81	0.85
Weighted Avg F1-Score	0.81	0.86

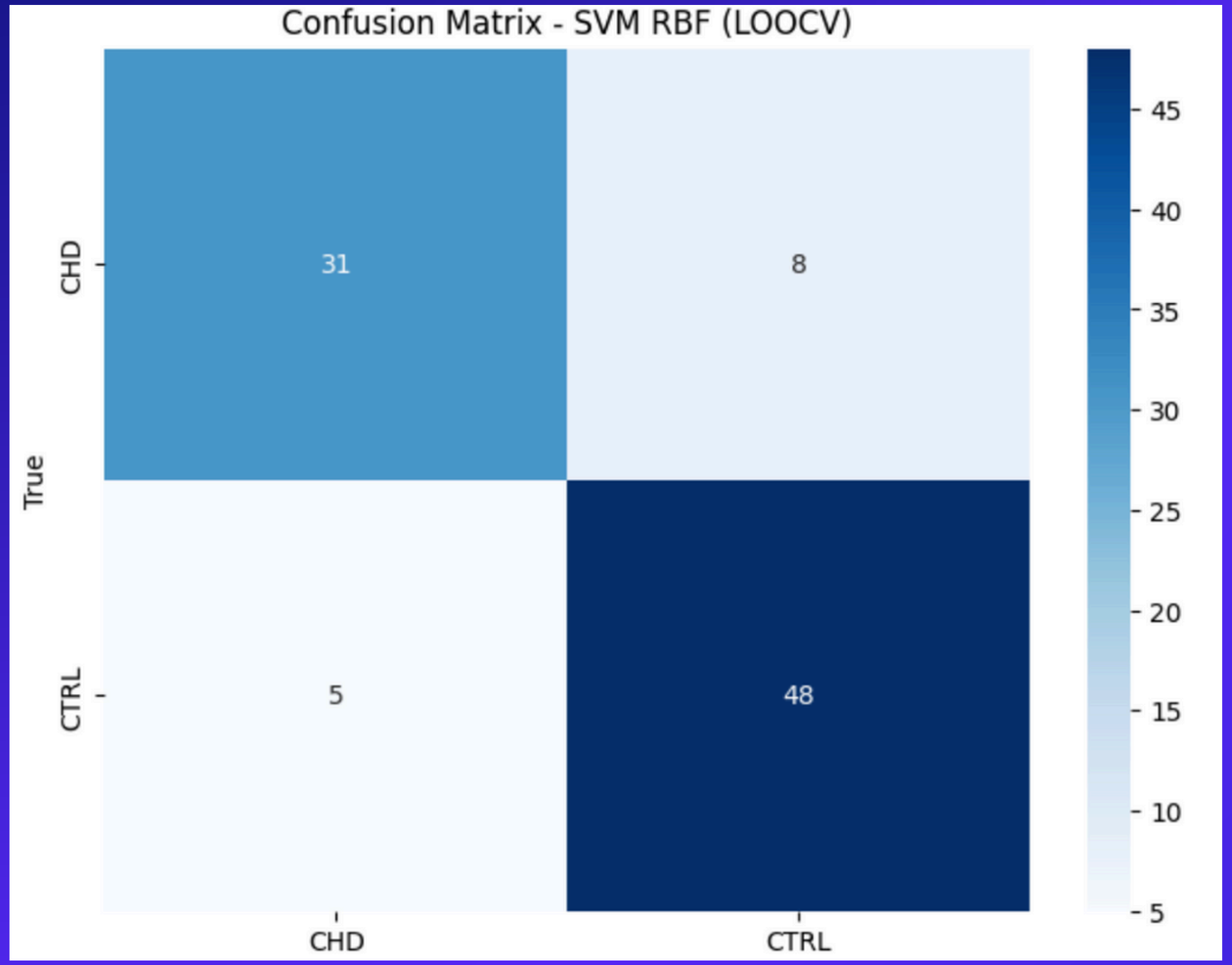


LOOCV

LOGISTIC REGRESSION NEW

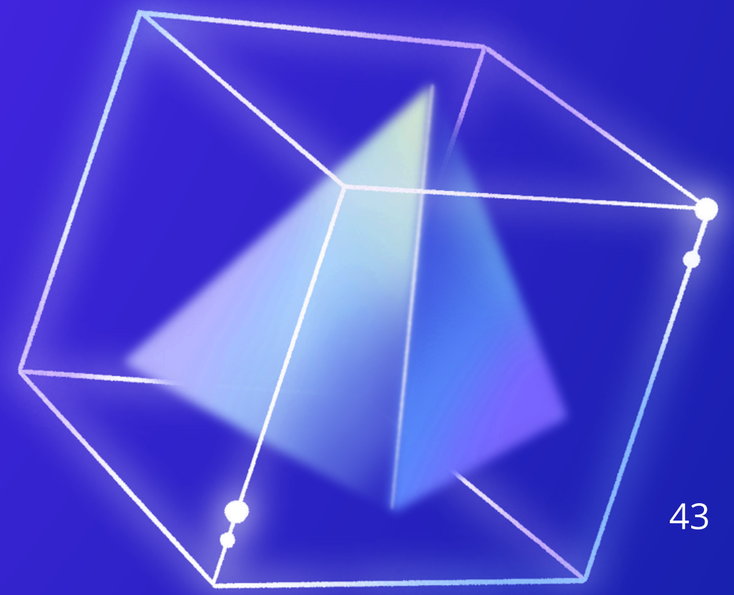
Metric	Grid Search	Leave-One-Out
Precision (CHD)	0.79	0.79
Precision (CTRL)	0.83	0.83
Recall (CHD)	0.77	0.77
Recall (CTRL)	0.85	0.85
F1-Score (CHD)	0.78	0.78
F1-Score (CTRL)	0.84	0.84
Accuracy	0.82	0.82
Macro Avg F1-Score	0.81	0.81
Weighted Avg F1-Score	0.81	0.81

SVM WITH RBF AND LOOCV



Classification Report:

	precision	recall	f1-score	support
CHD	0.86	0.79	0.83	39
CTRL	0.86	0.91	0.88	53
accuracy			0.86	92
macro avg	0.86	0.85	0.85	92
weighted avg	0.86	0.86	0.86	92



FIND THE MOST COMMON IMPORTANT FEATURES BETWEEN PQN, TIC AND MEDIAN NORMALIZATION

Number	Feature
1	Methylistamine
2	Pseudouridine
3	Asparagine
4	Imidazolepropionic acid
5	Ornithine

07

THANK YOU!