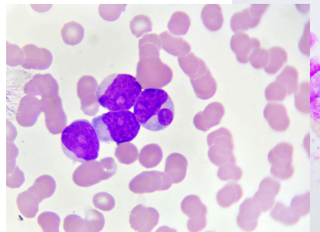
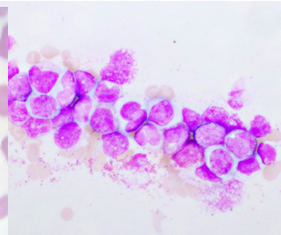
# DATA SCIENCE APPLICATION TO PREDICT THE SUBTYPE OF ACUTE LEUKEMIA (AML VS ALL) BY USING GEN EXPRESSION DATA

By Harish Muhammad

Science
AAAS

**Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring**
T. R. Golub *et al.*
*Science* **286**, 531 (1999);
DOI: 10.1126/science.286.5439.531

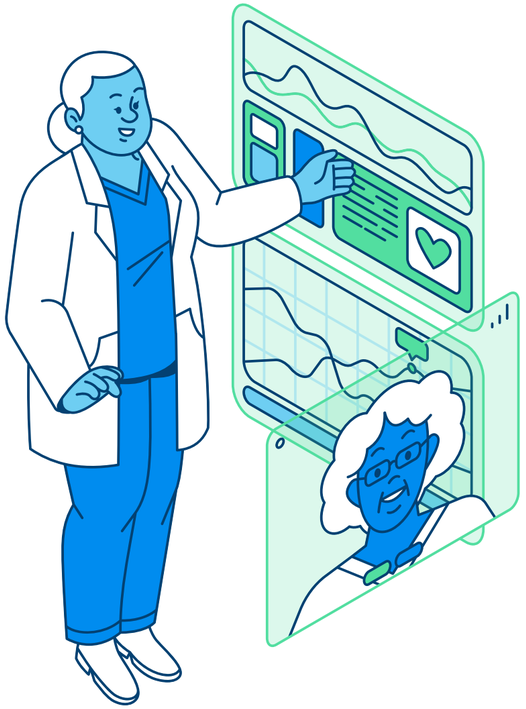



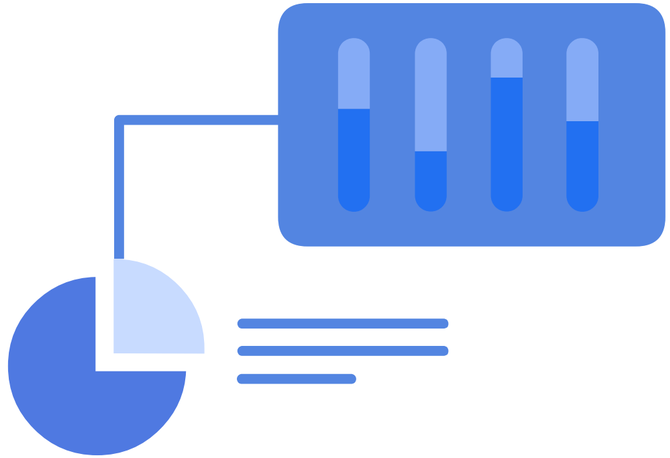


# Outlines

- Business problem and background
- Business objective (Menentukan objektif bisnis)
- Technical data science objective (Menentukan tujuan teknis data science)
- Examining data (Menelaah data)
- Validate data (Memvalidasi data)
- Determine object data (Menentukan objek data)
- Data construction (Mengkonstruksi data)
- Building a model scenario (Membangun skenario model)
- Model building (Membangun model)
- Model evaluation (Mengevaluasi hasil pemodelan)
- Reviewing the modeling process (Melakukan review pemodelan)

# BUSINESS PROBLEM & BACKGROUND

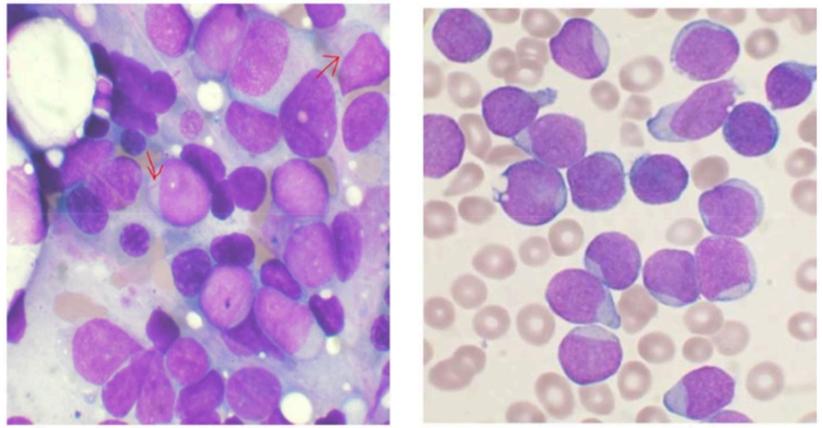




# Bussines problem Understanding

**Background & Business Problem**

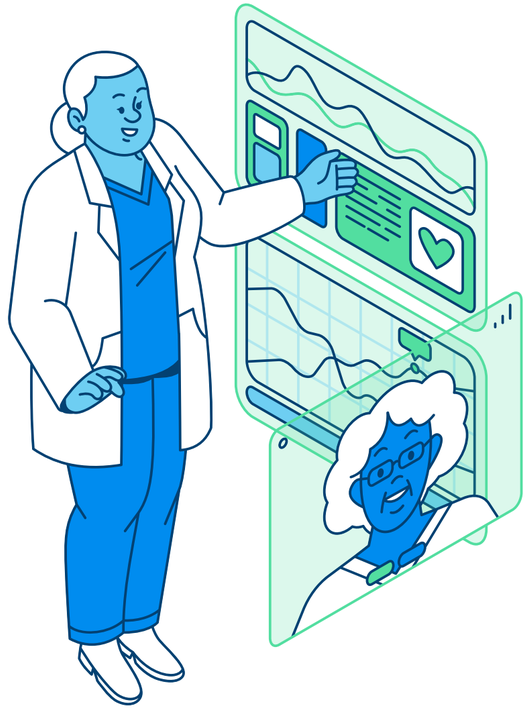
- **Medical context**: Acute leukemia (AML and ALL) symptoms often overlap, causing diagnostic challenges.
- **Clinical risk**: Incorrect classification can lead to ineffective treatment, exposure to toxic drugs (Chemotherapy), and lower survival rates.
- **Current limitation:** Conventional diagnoses (symptom-based and basic lab tests) are insufficient for accurate early-stage detection.
- **Need**:
  - Conventional diagnoses (symptom-based & lab tests) → often limited for accurate early-stage detection
  - Leverage gene expression data to support clinical decision-making → objective, faster, & potentially more affordable

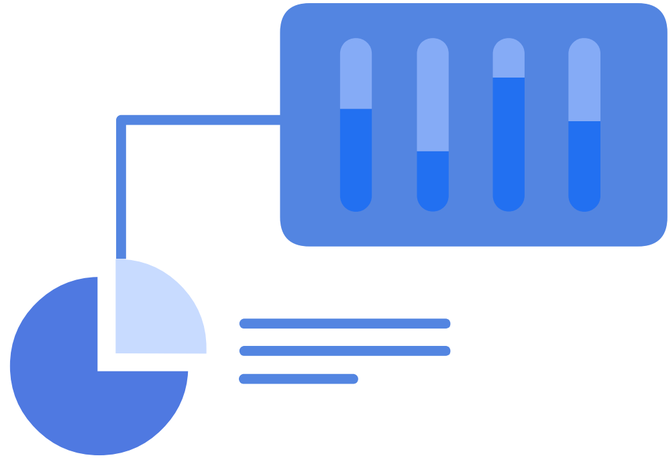


a: AML b: ALL

| Feature | **Acute Myeloid Leukemia** | **Acute Lymphoblastic Leukemia** |
|---|---|---|
| Origin | Myeloid cell precursors | Lymphoid cell precursors |
| Common in | Adults | Children |
| Cell morphology | heterogeneous myeloblasts | uniform lymphoblasts |
| Progression | Rapid | Rapid |
| Treatment sensitivity | require more aggressive therapy | Responds well to chemotherapy |

# DEFINE BUSINESS OBJECTIVE (MENENTUKAN OBJEKTIF BISNIS)





# Bussines Objective

🎯 **Business Objective**

- To develop a model that enables accurate early diagnosis of leukemia subtypes (AML vs ALL) to support clinical decision-making.

🔎 **Key Benefits**

- Improve clinical diagnostic workflow by integrating a predictive tool.
- Faster diagnosis process compared to the manual/pathologist method
- Reduce the need for costly and time-consuming lab tests.
- Empower medical personnel with a decision-support system and minimize errors.

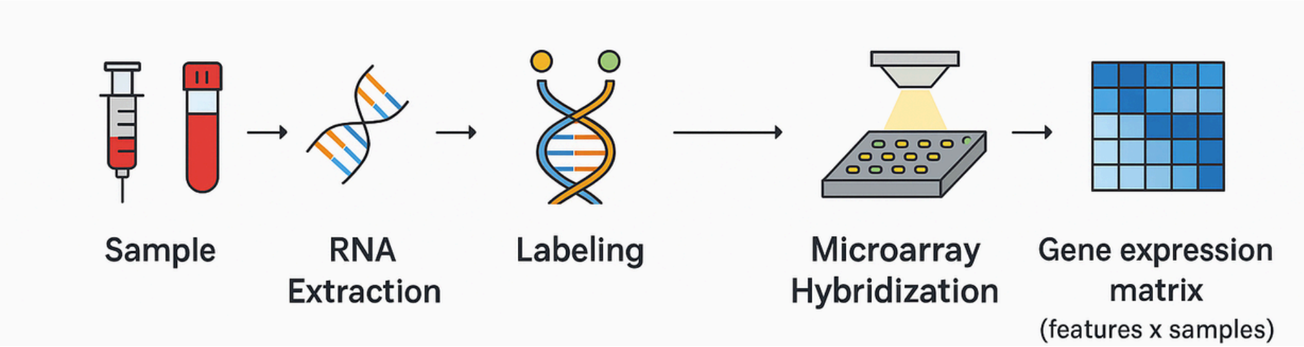
# Bussines Objective

✓ **Successful (business) metrics**

- Diagnosis accuracy ≥ 95% and stable performance
- Scalable & reproducible system across hospitals/labs
- Interpretability: Key components (PCA) contain leukemia-relevant genes

**Confirmation with Stakeholders**

- Metrics and objectives discussed with clinical researchers and ML experts
- Aligned with the needs of diagnostic support systems and ML implementation standards



# Terminology, resources, scope & limitation

**Terminology**

- AML/ALL: Types of acute leukemia
- Gene expression: RNA activity levels per gene
- PCA → Dimensionality reduction technique
- SMOTE → Oversampling for Imbalanced Data

**Resources**

- Dataset: Golub et.al
- Library & packages: Scikit-learn, imbalanced-learn, pandas, matplotlib
- Python notebook
- Pipeline-based implementation

**Assumptions:**

- Gene expression data accurately reflects cancer subtype (source: scientific article)

**Limitations**

- Small sample size (72 patients)
- High-dimensional (7129 gene features)
- External validation data not available

**Stakeholder Agreement**

- Documented assumptions and limits will be discussed with medical staff

# Risks and alternatives

**Identified Risks**

- The model is not accurate enough to be implemented
- Overfitting the model due to high dimensionality
- Model performance may not generalize to new samples
- Losing key information

**Mitigation Plans**

- The model will be used as a supporting tool for diagnosis, not the main decision maker
- PCA  will be used to reduce the dimensions
- Cross-validation will be applied, and stability will be evaluated
- Compare multiple ML models and tune the best-performing ones
- PCA maintains ≥ 95% variance

# Cost & Benefit Analysis

**Estimated cost components**

| Component | Cost |
|---|---|
| Cloud computing & notebooks | Free (Google Colab) |
| Work hours (analysis, coding) | ~80 hours |
| Work hours (Reporting, slides) | ~50 hours |
| Research & mentoring | Internal |

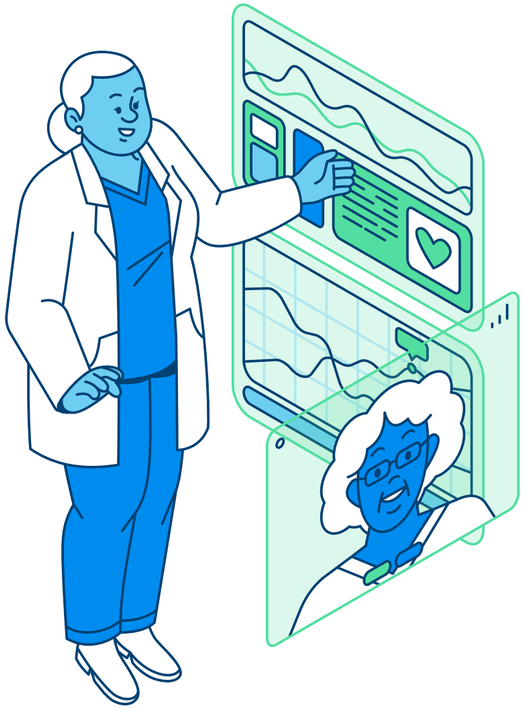
**Expected benefits**

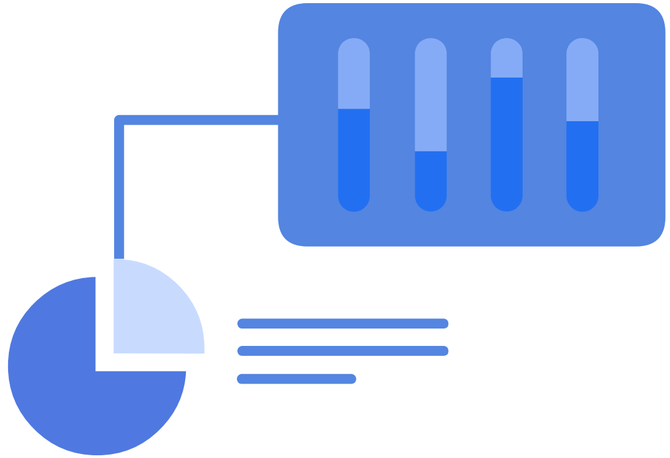
- Improved diagnostic support tool
- Faster, more accurate leukemia classification

**Stakeholder Agreement**

- Time, tools, and potential reuse approved as proportional to project value

# TECHNICAL DATA SCIENCE OBJECTIVE (MENENTUKAN TUJUAN TEKNIS DATA SCIENCE)





# Technical Data Science Objective

🎯 **Technical Data Science Objective**

- To build and evaluate a supervised machine learning model that classifies samples into AML or ALL based on high-dimensional gene expression data.

🔎 **Key Points:**

- Build a binary classification model using ML
- Normalize and clean high-dimensional gene expression data
- Handle any missing values or inconsistencies
- Apply preprocessing: outlier handling, scaling, resampling, and dimensionality reduction
- Use a pipeline for robust preprocessing
- Compare multiple algorithm performances
- Optimize best-performing models (XGBoost, LGBM, RF)
- Evaluate performance using metrics: accuracy and ROC-AUC

# Technical Data Science Objective

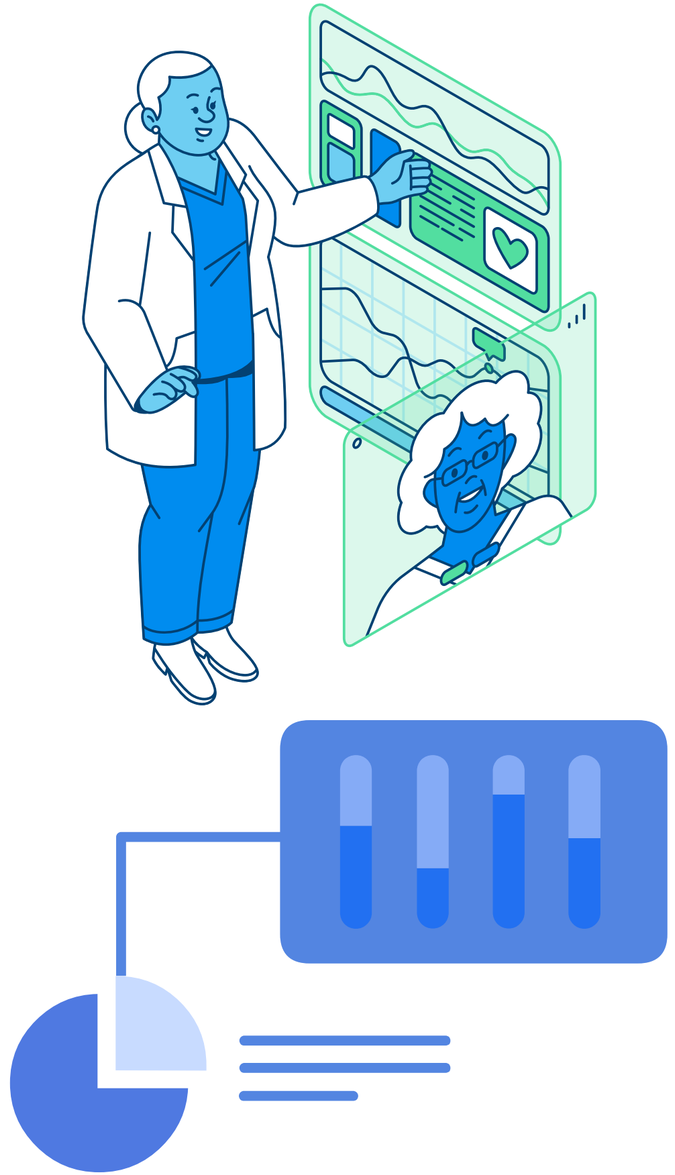
## 🎯 Technical Data Science Success Criteria

| Criteria | Description |
|---|---|
| Cross-val & test Accuracy ≥ 95% | Ensures the model generalizes well and has high accuracy |
| ROC AUC (Test) ≥ 0.95 | Indicates excellent model discriminative ability |
| PCA maintains > 95% variance | Ensure dimensionality reduction doesn’t lose key information |
| Generalization | No significant difference between training and testing |

**Stakeholder approval**

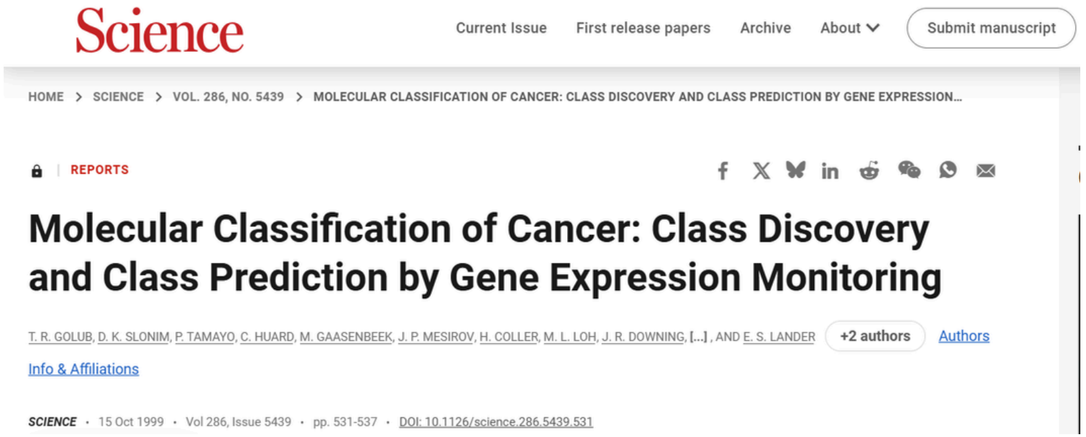
- Metrics will be further discussed with medical diagnostic and machine learning experts for approval

# EXAMINING DATA (MENELAAH DATA)



# Dataset overview

## Source: Golub et.al (Scientific paper)



Science

Current Issue | First release papers | Archive | About | Submit manuscript

HOME > SCIENCE > VOL. 286, NO. 5439 > MOLECULAR CLASSIFICATION OF CANCER: CLASS DISCOVERY AND CLASS PREDICTION BY GENE EXPRESSION...

REPORTS

**Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring**

T. R. GOLUB, D. K. SLONIM, P. TAMAYO, C. HUARD, M. GAASENBEEK, J. P. MESIROV, H. COLLER, M. L. LOH, J. R. DOWNING, [...], AND E. S. LANDER +2 authors Authors

Info & Affiliations

*SCIENCE* • 15 Oct 1999 • Vol 286, Issue 5439 • pp. 531-537 • DOI: 10.1126/science.286.5439.531

## Data dimension (raw)

```
# Reviewing Data dimension
# Number of columns and rows
print(f' Number of columns and rows of df_actual: {df_actual.shape}')
print(f' Number of columns and rows of df_train: {df_train.shape}')
print(f' Number of columns and rows of df_test: {df_test.shape}')
```

```
 Number of columns and rows of df_actual: (72, 2)
 Number of columns and rows of df_train: (7129, 78)
 Number of columns and rows of df_test: (7129, 70)
```

# Dataset overview

**df_actual**

| | patient | cancer |
|---|---|---|
| 0 | 1 | ALL |
| 1 | 2 | ALL |
| 2 | 3 | ALL |
| 3 | 4 | ALL |
| 4 | 5 | ALL |
| ... | ... | ... |
| 67 | 68 | ALL |
| 68 | 69 | ALL |
| 69 | 70 | ALL |
| 70 | 71 | ALL |
| 71 | 72 | ALL |

72 rows × 2 columns

**df_actual**

- Each row represents a patient
- Total 72 patients + label
- Target/label: Binary, AML or ALL

**df_ train & df_test**

- Rows → Transposed into columns
- Overall, 7129 features → high dimensionality
- All features: continuous numeric values

**df_train**

| | Gene Description | Gene Accession Number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | ... | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | AFFX-BioB-5_at (endogenous control) | AFFX-BioB-5_at | -214 | -139 | -76 | -135 | -106 | -138 | -72 | -413 | ... | 15 | -318 | -32 | -124 | -135 | -20 | 7 | -213 | -25 | -72 |
| 1 | AFFX-BioB-M_at (endogenous control) | AFFX-BioB-M_at | -153 | -73 | -49 | -114 | -125 | -85 | -144 | -260 | ... | -114 | -192 | -49 | -79 | -186 | -207 | -100 | -252 | -20 | -139 |
| 2 | AFFX-BioB-3_at (endogenous control) | AFFX-BioB-3_at | -58 | -1 | -307 | 265 | -76 | 215 | 238 | 7 | ... | 2 | -95 | 49 | -37 | -70 | -50 | -57 | 136 | 124 | -1 |
| 3 | AFFX-BioC-5_at (endogenous control) | AFFX-BioC-5_at | 88 | 283 | 309 | 12 | 168 | 71 | 55 | -2 | ... | 193 | 312 | 230 | 330 | 337 | 101 | 132 | 318 | 325 | 392 |
| 4 | AFFX-BioC-3_at (endogenous control) | AFFX-BioC-3_at | -295 | -264 | -376 | -419 | -230 | -272 | -399 | -541 | ... | -51 | -139 | -367 | -188 | -407 | -369 | -377 | -209 | -396 | -324 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 7124 | PTGER3 Prostaglandin E receptor 3 (subtype EP3... | X83863_at | 793 | 782 | 1138 | 627 | 250 | 645 | 1140 | 1799 | ... | 279 | 737 | 588 | 1170 | 2315 | 834 | 752 | 1293 | 1733 | 1567 |
| 7125 | HMG2 High-mobility group (nonhistone chromosom... | Z17240_at | 329 | 295 | 777 | 170 | 314 | 341 | 482 | 446 | ... | 51 | 227 | 361 | 284 | 250 | 557 | 295 | 342 | 304 | 627 |
| 7126 | RB1 Retinoblastoma 1 (including osteosarcoma) | L49218_f_at | 36 | 11 | 41 | -50 | 14 | 26 | 10 | 59 | ... | 6 | -9 | -26 | 39 | -12 | -12 | 28 | 26 | 12 | 21 |
| 7127 | GB DEF = Glycophorin Sta (type A) exons 3 and ... | M71243_f_at | 191 | 76 | 228 | 126 | 56 | 193 | 369 | 781 | ... | 2484 | 371 | 133 | 298 | 790 | 335 | 1558 | 246 | 3193 | 2520 |
| 7128 | GB DEF = mRNA (clone 1A7) | Z78285_f_at | -37 | -14 | -41 | -91 | -25 | -53 | -42 | 20 | ... | -2 | -31 | -32 | -3 | -10 | -65 | -67 | 23 | -33 | 0 |

7129 rows × 40 columns

**df_test**

| | Gene Description | Gene Accession Number | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | ... | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | AFFX-BioB-5_at (endogenous control) | AFFX-BioB-5_at | -342 | -87 | -62 | 22 | 86 | -146 | -187 | -56 | ... | -161 | -48 | -62 | -58 | -76 | -154 | -79 | -55 | -59 | -131 |
| 1 | AFFX-BioB-M_at (endogenous control) | AFFX-BioB-M_at | -200 | -248 | -23 | -153 | -36 | -74 | -187 | -43 | ... | -215 | -531 | -198 | -217 | -98 | -136 | -118 | -44 | -114 | -126 |
| 2 | AFFX-BioB-3_at (endogenous control) | AFFX-BioB-3_at | 41 | 262 | -7 | 17 | -141 | 170 | 312 | 43 | ... | -46 | -124 | -5 | 63 | -153 | 49 | -30 | 12 | 23 | -50 |
| 3 | AFFX-BioC-5_at (endogenous control) | AFFX-BioC-5_at | 328 | 295 | 142 | 276 | 252 | 174 | 142 | 177 | ... | 146 | 431 | 141 | 95 | 237 | 180 | 68 | 129 | 146 | 211 |
| 4 | AFFX-BioC-3_at (endogenous control) | AFFX-BioC-3_at | -224 | -226 | -233 | -211 | -201 | -32 | 114 | -116 | ... | -172 | -496 | -256 | -191 | -215 | -257 | -110 | -108 | -171 | -206 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 7124 | PTGER3 Prostaglandin E receptor 3 (subtype EP3... | X83863_at | 1074 | 67 | 245 | 893 | 1235 | 354 | 304 | 625 | ... | 809 | 466 | 707 | 423 | 441 | 524 | 742 | 320 | 348 | 874 |
| 7125 | HMG2 High-mobility group (nonhistone chromosom... | Z17240_at | 475 | 263 | 164 | 297 | 9 | -42 | -1 | 173 | ... | 445 | 349 | 354 | 41 | 99 | 249 | 234 | 174 | 208 | 393 |
| 7126 | RB1 Retinoblastoma 1 (including osteosarcoma) | L49218_f_at | 48 | -33 | 84 | 6 | 7 | -100 | -207 | 63 | ... | -2 | 0 | -22 | 0 | -8 | 40 | 72 | -4 | 0 | 34 |
| 7127 | GB DEF = Glycophorin Sta (type A) exons 3 and ... | M71243_f_at | 168 | -33 | 100 | 1971 | 1545 | 45 | 112 | 63 | ... | 210 | 284 | 260 | 1777 | 80 | -68 | 109 | 176 | 74 | 237 |
| 7128 | GB DEF = mRNA (clone 1A7) | Z78285_f_at | -70 | -21 | -18 | -42 | -81 | -108 | -190 | -62 | ... | 16 | -73 | 5 | -49 | -12 | -1 | -30 | 40 | -12 | -2 |

7129 rows × 36 columns

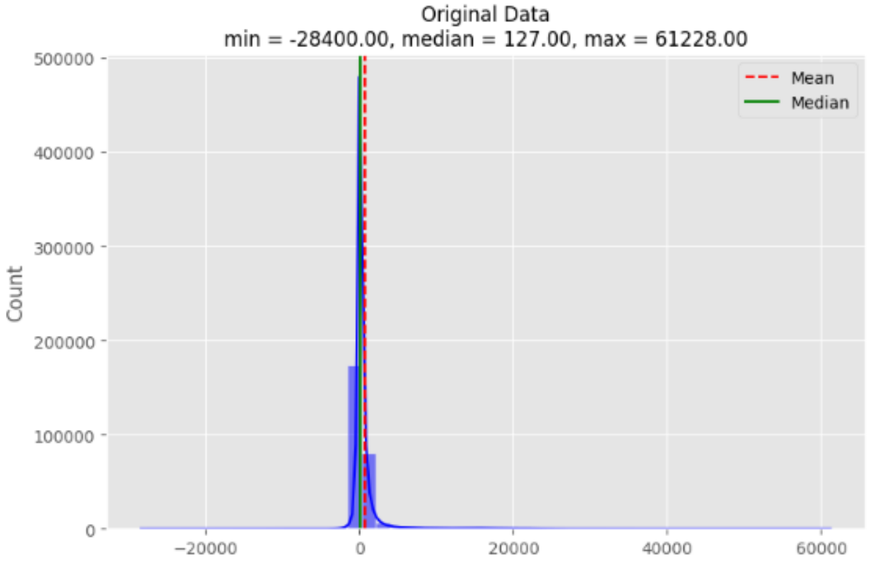
# Statistical descriptive

## Statistical descriptive

| Gene Accession Number | AFFX-BioB-5_at | AFFX-BioB-M_at | AFFX-BioB-3_at | AFFX-BioC-5_at | AFFX-BioC-3_at | AFFX-BioDn-5_at | AFFX-BioDn-3_at | AFFX-CreX-5_at | AFFX-CreX-3_at | AFFX-BioB-5_st | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 38.000000 | 38.000000 | 38.000000 | 38.000000 | 38.000000 | 38.000000 | 38.000000 | 38.000000 | 38.000000 | 38.000000 | ... |
| mean | -120.868421 | -150.526316 | -17.157895 | 181.394737 | -276.552632 | -439.210526 | -43.578947 | -201.184211 | 99.052632 | 112.131579 | ... |
| std | 109.555656 | 75.734507 | 117.686144 | 117.468004 | 111.004431 | 135.458412 | 219.482393 | 90.838989 | 83.178397 | 211.815597 | ... |
| min | -476.000000 | -327.000000 | -307.000000 | -36.000000 | -541.000000 | -790.000000 | -479.000000 | -463.000000 | -82.000000 | -215.000000 | ... |
| 25% | -138.750000 | -205.000000 | -83.250000 | 81.250000 | -374.250000 | -547.000000 | -169.000000 | -239.250000 | 36.000000 | -47.000000 | ... |
| 50% | -106.500000 | -141.500000 | -43.500000 | 200.000000 | -263.000000 | -426.500000 | -33.500000 | -185.500000 | 99.500000 | 70.500000 | ... |
| 75% | -68.250000 | -94.750000 | 47.250000 | 279.250000 | -188.750000 | -344.750000 | 79.000000 | -144.750000 | 152.250000 | 242.750000 | ... |
| max | 17.000000 | -20.000000 | 265.000000 | 392.000000 | -51.000000 | -155.000000 | 419.000000 | -24.000000 | 283.000000 | 561.000000 | ... |

8 rows × 7129 columns

## Data distribution

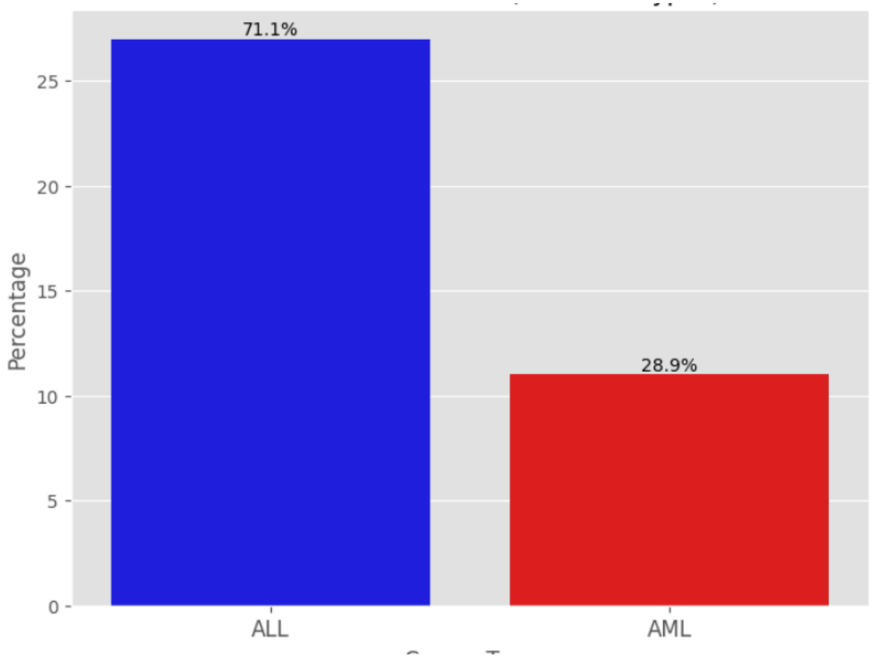



## Insights

- Gene expression (training data) range:  -28400.00 to 71369.00
- Gene expression (testing data) range:  -26775.00 to 71369.00
- Skewed distribution in many genes
- High variance → gene expressed differently among patients

# Data distribution

## Class distribution




Mild Imbalance class (7:3)

## Data distribution

The total number of features with normal distribution and asymetric distribution are 3879 features and 3250 features

| | Feature | D'Agostino-Pearson Statistic | P-value | Distribution | Skewness | Skewness Type |
|---|---|---|---|---|---|---|
| **0** | AFFX-BioB-5_at | 17.649720 | 1.470321e-04 | Not Normally Distributed | -1.517897 | Left Skew |
| **1** | AFFX-BioB-M_at | 2.221980 | 3.292329e-01 | Normally Distributed | -0.559021 | Left Skew |
| **2** | AFFX-BioB-3_at | 2.549140 | 2.795512e-01 | Normally Distributed | 0.453981 | Right Skew |
| **3** | AFFX-BioC-5_at | 6.111995 | 4.707575e-02 | Not Normally Distributed | -0.182510 | Left Skew |
| **4** | AFFX-BioC-3_at | 1.271057 | 5.296554e-01 | Normally Distributed | -0.222709 | Left Skew |
| **...** | ... | ... | ... | ... | ... | ... |
| **7124** | X83863_at | 8.838071 | 1.204585e-02 | Not Normally Distributed | 1.070252 | Right Skew |
| **7125** | Z17240_at | 20.939533 | 2.838168e-05 | Not Normally Distributed | 1.551995 | Right Skew |
| **7126** | L49218_f_at | 5.772584 | 5.578268e-02 | Normally Distributed | 0.425516 | Right Skew |
| **7127** | M71243_f_at | 36.476604 | 1.200069e-08 | Not Normally Distributed | 2.616494 | Right Skew |
| **7128** | Z78285_f_at | 0.354611 | 8.375239e-01 | Normally Distributed | -0.214672 | Left Skew |

7129 rows × 6 columns

- 3879 features → Normally distributed
- 3250 features→ Not normally distributed

# Outliers

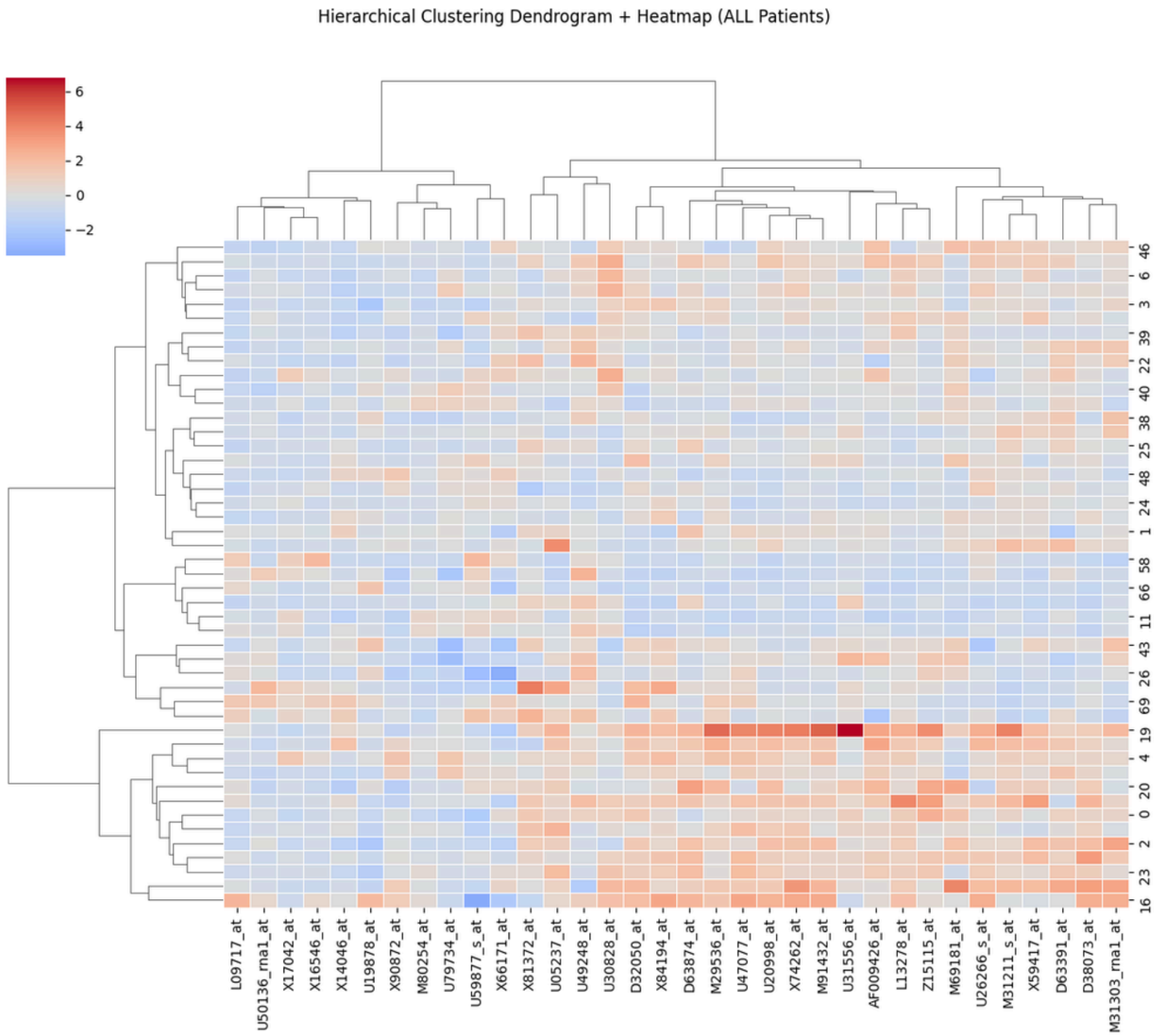
The total number of features with high percentage of outliers (more than 5% for each columns) were 2959 features

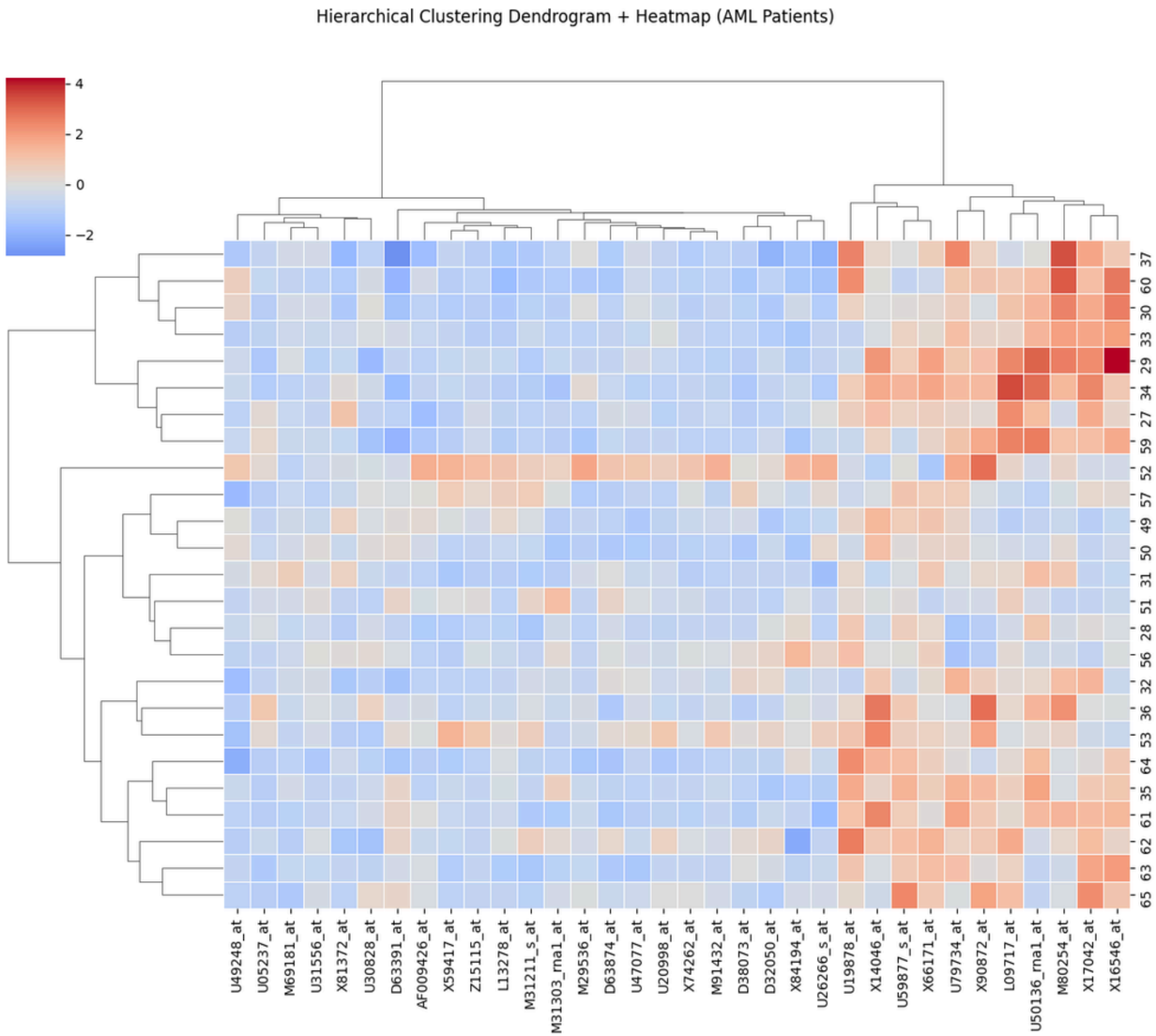
| | Columns | Number of Outliers | (%) of Outliers | Lower bound | Upper bound |
|---|---|---|---|---|---|
| **0** | AFFX-BioB-5_at | 5 | 13.16 | -244.500 | 37.500 |
| **1** | AFFX-BioB-M_at | 0 | 0.00 | -370.375 | 70.625 |
| **2** | AFFX-BioB-3_at | 2 | 5.26 | -279.000 | 243.000 |
| **3** | AFFX-BioC-5_at | 0 | 0.00 | -215.750 | 576.250 |
| **4** | AFFX-BioC-3_at | 0 | 0.00 | -652.500 | 89.500 |
| **...** | ... | ... | ... | ... | ... |
| **7124** | X83863_at | 1 | 2.63 | -179.875 | 1887.125 |
| **7125** | Z17240_at | 4 | 10.53 | -2.375 | 624.625 |
| **7126** | L49218_f_at | 5 | 13.16 | -25.375 | 63.625 |
| **7127** | M71243_f_at | 4 | 10.53 | -390.875 | 1014.125 |
| **7128** | Z78285_f_at | 3 | 7.89 | -89.625 | 35.375 |

7129 rows × 5 columns

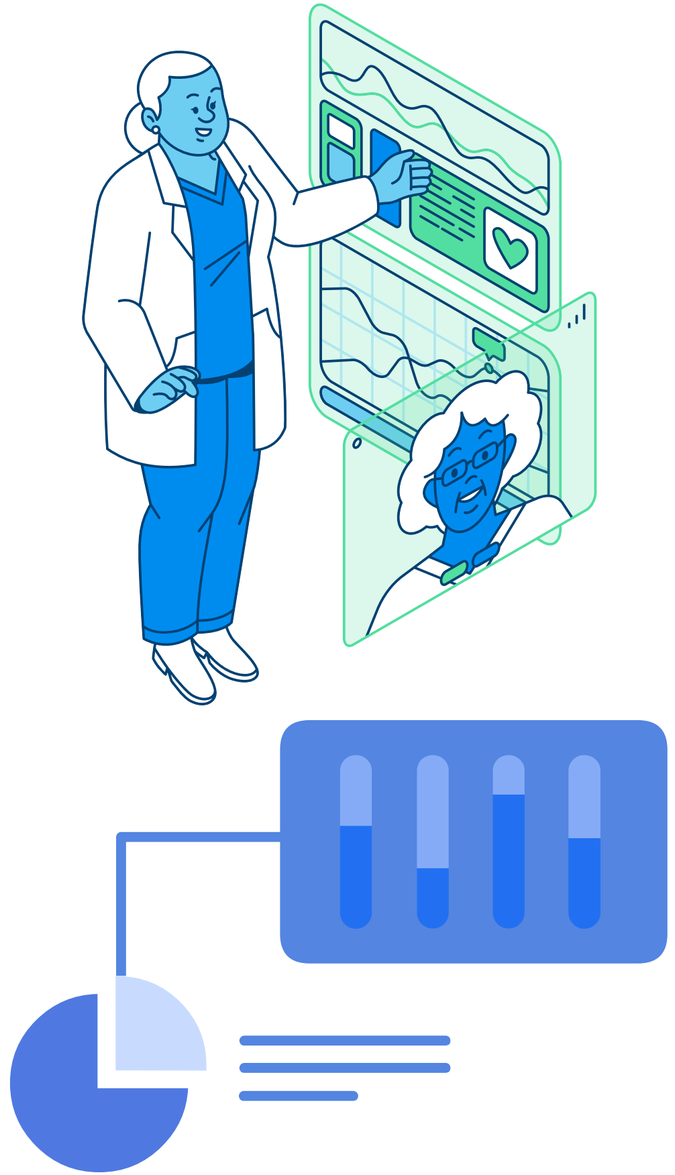
**2959 features contain more than 5% outliers per column**

# Data Relationship



# VALIDATE DATA (MEMVALIDASI DATA)

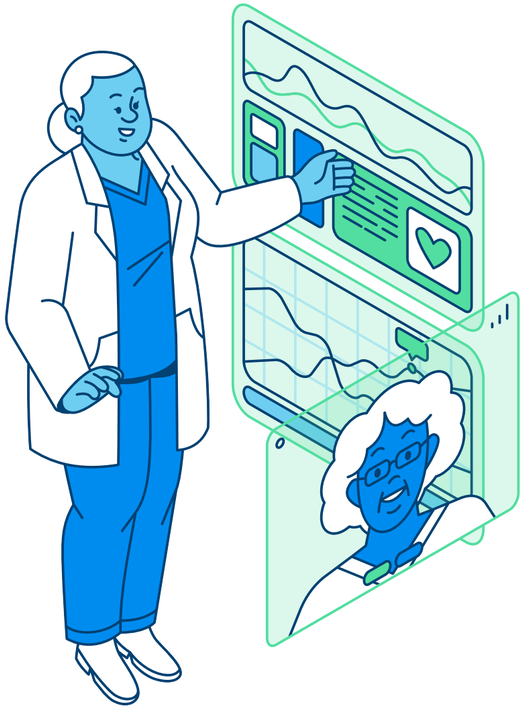


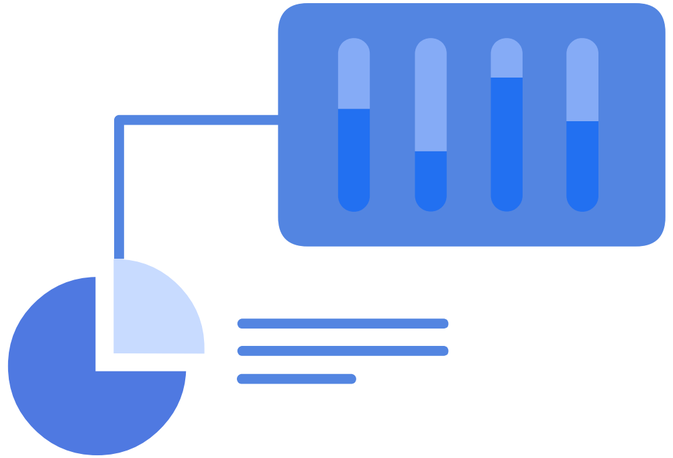
# Validate data

## Validation step performed

| Validation Task | Result | Recommendations |
|---|---|---|
| Missing Values Check | No missing values detected in gene expression matrix & target | No action needed |
| Data Type Consistency | All features are numerical (float), all targets (labels) are categorical | Maintain current data types. Encoding target label. |
| feature sufficiency | Adequate number of features (>7000 features) | Perform dimensionality reduction technique |
| Duplicate Rows/Sampl | No duplicate patient samples identified | No action needed |
| Outlier Detection | Some features contain extreme values (expected in biology) | Do not remove data, apply winsorization |
| Value Range Check | Gene expression levels are within expected microarray ranges | No action needed |
| Label Verification | AML/ALL labels match reference file | No action needed |

# DETERMINE OBJECT DATA (MENENTUKAN OBJEK DATA)





# Determine object data

## Object data

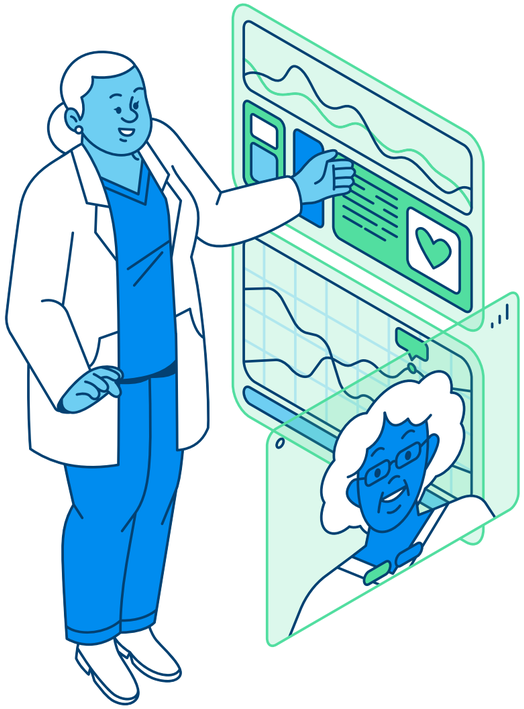
| Component | Description |
|---|---|
| Selected Feature | Gene expression (≈7000 genes per sample) (numeric continous) (Wide format) |
| Deleted features | Gene description, Gene accession number, call.1, call.2, call... |
| Record (row) Analysis | Each row represents a patient |
| Target Variable | Diagnosis class: ALL (Acute Lymphoblastic Leukemia) or AML (Myeloid) |

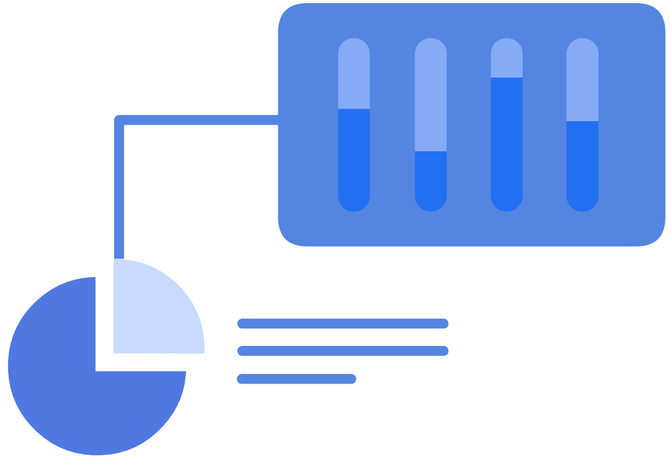
Gene expression (Features)

Individual patients (Observation/records)

| | AFFX-BioB-5_at | AFFX-BioB-M_at | AFFX-BioB-3_at | AFFX-BioC-5_at | AFFX-BioC-3_at | AFFX-BioDn-5_at | AFFX-BioDn-3_at | AFFX-CreX-5_at | AFFX-CreX-3_at | AFFX-BioB-5_st | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | -214 | -153 | -58 | 88 | -295 | -558 | 199 | -176 | 252 | 206 | ... |
| **1** | -139 | -73 | -1 | 283 | -264 | -400 | -330 | -168 | 101 | 74 | ... |
| **2** | -76 | -49 | -307 | 309 | -376 | -650 | 33 | -367 | 206 | -215 | ... |

# DATA CLEANING (MEMBERSIHKAN DATA)





# Data Cleaning strategy (part 1)

## Handling Missing values

```
# Checking and calculating missing values
print(f'Total missing values in target/label dataset: {df_actual.isna().sum().sum()}')
print(f'Total missing values in training dataset: {df_train.isna().sum().sum()}')
print(f'Total missing values in testing dataset: {df_test.isna().sum().sum()}')
```

```
Total missing values in target/label dataset: 0
Total missing values in training dataset: 0
Total missing values in testing dataset: 0
```

## Handling duplicates

```
# Checking and calculating duplicates
print(f'Total duplicates in target/label dataset: {df_actual.duplicated().sum()}')
print(f'Total duplicates in training dataset: {df_train.duplicated().sum()}')
print(f'Total duplicates in testing dataset: {df_test.duplicated().sum()}')
```

```
Total duplicates in target/label dataset: 0
Total duplicates in training dataset: 0
Total duplicates in testing dataset: 0
```

Missing Values df_actual

| | |
|---|---|
| patient | 0 |
| cancer | 0 |

Missing Values df_train

| | |
|---|---|
| Gene Description | 0 |
| Gene Accession Number | 0 |
| 1 | 0 |
| call | 0 |
| 2 | 0 |
| ... | ... |
| call.35 | 0 |
| 32 | 0 |
| call.36 | 0 |
| 33 | 0 |
| call.37 | 0 |

78 rows × 1 columns

Missing Values df_test

| | |
|---|---|
| Gene Description | 0 |
| Gene Accession Number | 0 |
| 39 | 0 |
| call | 0 |
| 40 | 0 |
| ... | ... |
| call.31 | 0 |
| 64 | 0 |
| call.32 | 0 |
| 62 | 0 |
| call.33 | 0 |

70 rows × 1 columns

# Data cleaning part 2

## Columns removal

```
# Removing 'call' columns from training and testing data frame
train_to_keep = [col for col in df_train.columns if "call" not in col]
test_to_keep = [col for col in df_test.columns if "call" not in col]

# Isolating only the usable features
df_train = df_train[train_to_keep]
df_test = df_test[test_to_keep]
```

## Columns arrangement

```
# Rearranging the order of columns in the data frame

train_column_order =  ['Gene Description', 'Gene Accession Number', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10',
       '11', '12', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22', '23', '24', '25',
       '26', '27', '28', '29', '30', '31', '32', '33', '34', '35', '36', '37', '38']

test_column_order = ['Gene Description', 'Gene Accession Number','39', '40', '41', '42', '43', '44', '45', '46',
       '47', '48', '49', '50', '51', '52', '53',  '54', '55', '56', '57', '58', '59',
       '60', '61', '62', '63', '64', '65', '66', '67', '68', '69', '70', '71', '72']

df_train = df_train.reindex(columns=train_column_order)
df_test = df_test.reindex(columns=test_column_order)
```

## Transposing columns

```
# Transposing the columns
X_train = df_train.T
X_test = df_test.T
```

## Data arrangement & re-formatting

| | Gene Description | Gene Accession Number | 1 | call | 2 | call.1 | 3 | call.2 | 4 | call.3 | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | AFFX-BioB-5_at (endogenous control) | AFFX-BioB-5_at | -214 | A | -139 | A | -76 | A | -135 | A | ... |
| 1 | AFFX-BioB-M_at (endogenous control) | AFFX-BioB-M_at | -153 | A | -73 | A | -49 | A | -114 | A | ... |
| 2 | AFFX-BioB-3_at (endogenous control) | AFFX-BioB-3_at | -58 | A | -1 | A | -307 | A | 265 | A | ... |

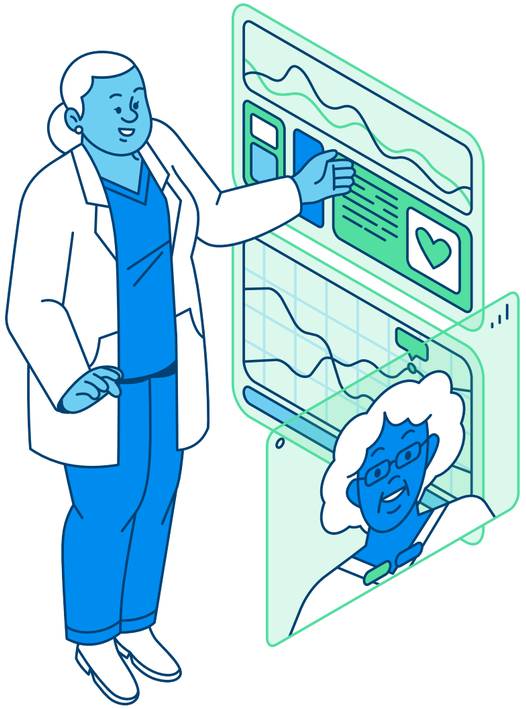
↓

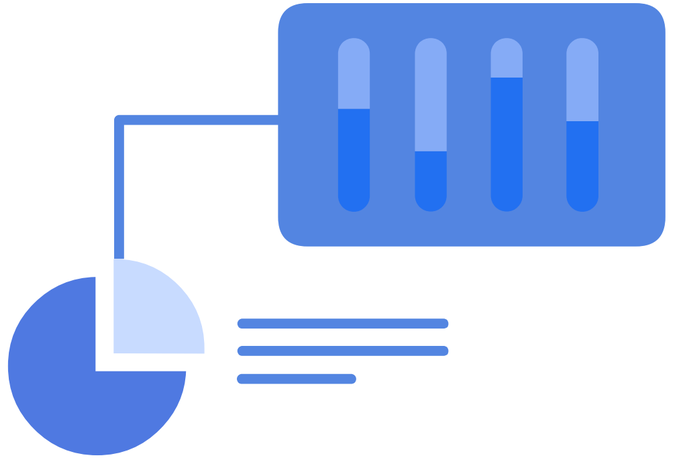
| | Gene Description | Gene Accession Number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | AFFX-BioB-5_at (endogenous control) | AFFX-BioB-5_at | -214 | -139 | -76 | -135 | -106 | -138 | -72 | -413 | ... |
| 1 | AFFX-BioB-M_at (endogenous control) | AFFX-BioB-M_at | -153 | -73 | -49 | -114 | -125 | -85 | -144 | -260 | ... |
| 2 | AFFX-BioB-3_at (endogenous control) | AFFX-BioB-3_at | -58 | -1 | -307 | 265 | -76 | 215 | 238 | 7 | ... |

↓

| Gene Accession Number | AFFX-BioB-5_at | AFFX-BioB-M_at | AFFX-BioB-3_at | AFFX-BioC-5_at | AFFX-BioC-3_at | AFFX-BioDn-5_at | AFFX-BioDn-3_at | AFFX-CreX-5_at | AFFX-CreX-3_at | AFFX-BioB-5_st | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -214 | -153 | -58 | 88 | -295 | -558 | 199 | -176 | 252 | 206 | ... |
| 2 | -139 | -73 | -1 | 283 | -264 | -400 | -330 | -168 | 101 | 74 | ... |
| 3 | -76 | -49 | -307 | 309 | -376 | -650 | 33 | -367 | 206 | -215 | ... |

# DATA CONSTRUCTION (MENGKONSTRUKSI DATA)





# Data Construction strategy

**Initial feature representative**

**Dataset characteristics**

- High-dimensional: Raw gene expression values (~7000 genes)
- Small sample size: ~72 patients
- Skewed (contains outliers)
- Imbalance dataset (~7:3)

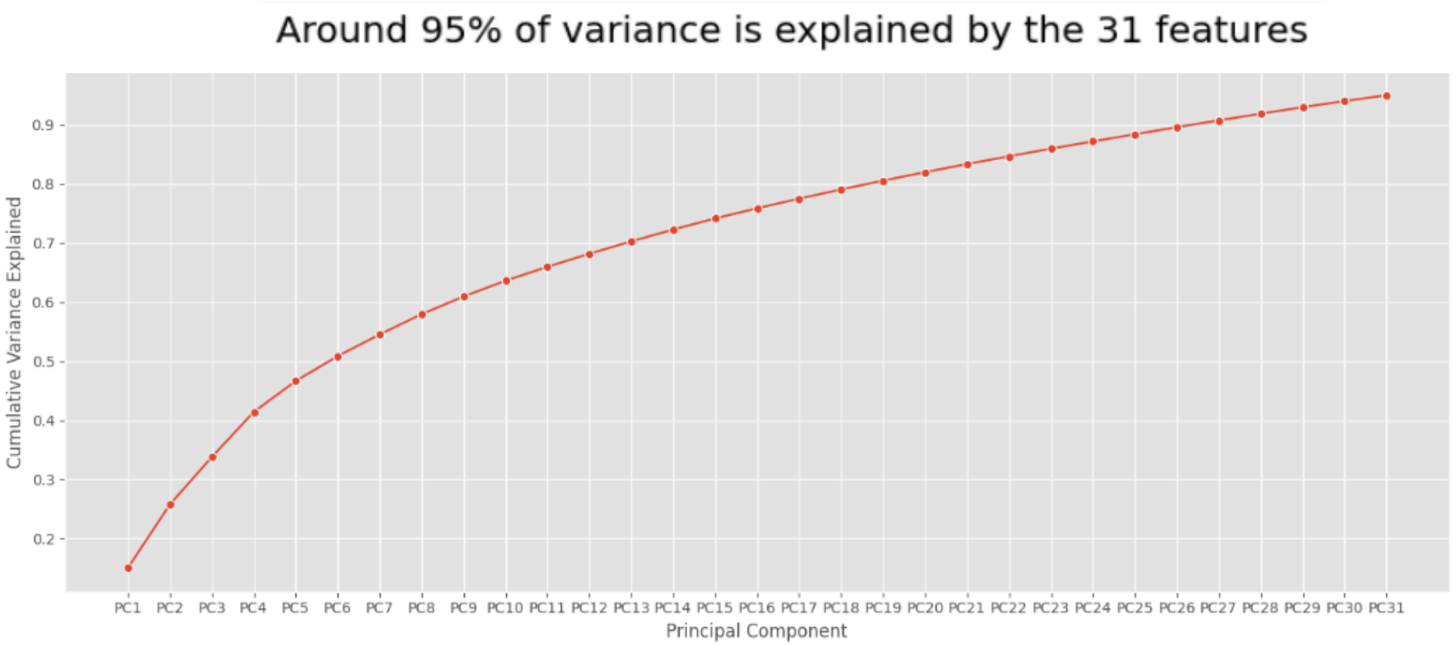
**Needs**

- Dimensionality reduction to avoid overfitting
- Outlier handling without data removal (prevent information loss)
- Feature scaling to improve model learning
- Resampling to handle class imbalance

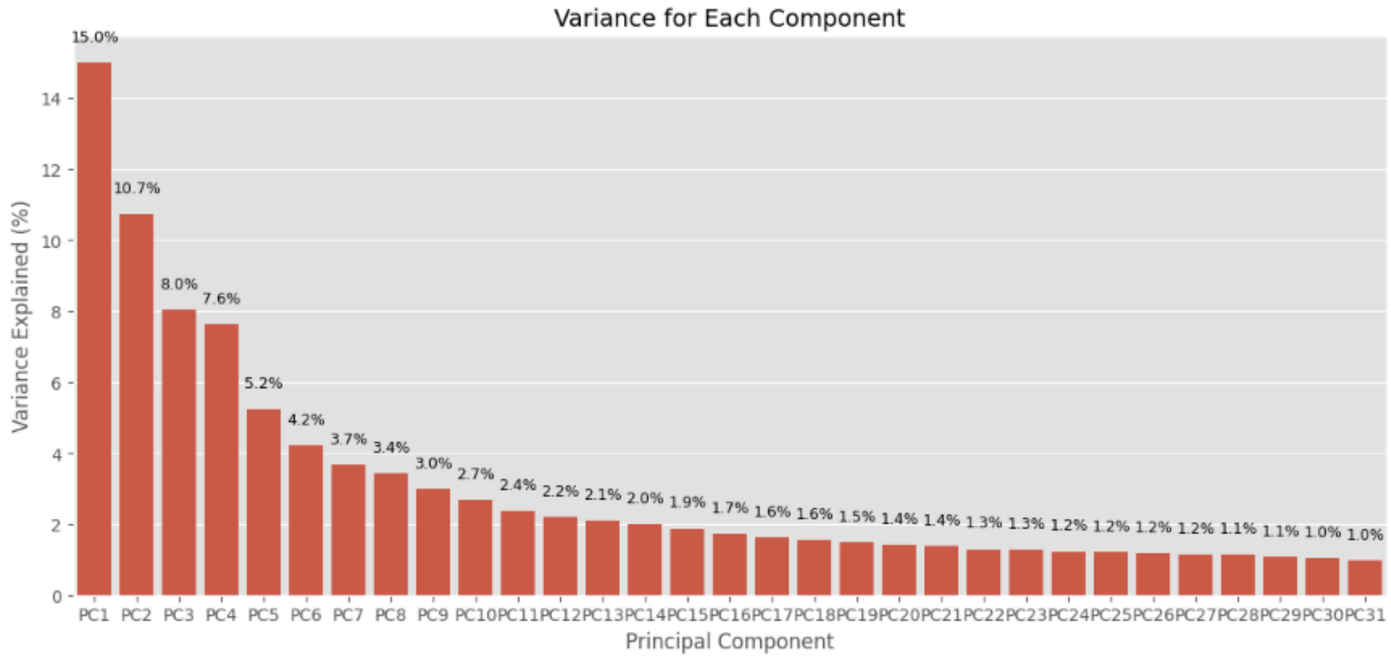
# Feature transformation: PCA construction

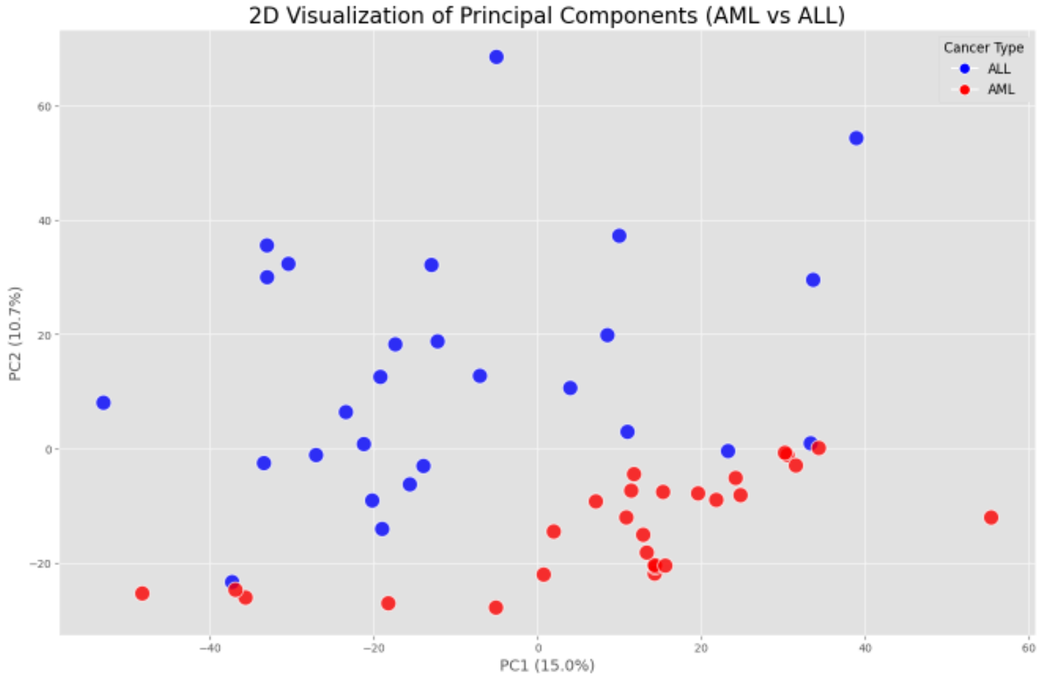
📉 **PCA Implementation**

- PCA replaces individual genes as features
- Reduced ~7,000+ to 31 Principal components
- Captures >95% of total variance
- Benefits:
  - Reduce overfitting: reduce the complexity of the data
  - Speed up computations: fewer dimensions → faster training time
  - Feature extraction: PCA can help to identify the most important underlying features in the data
  - Noise reduction: component with low variance → considered noise → discarded

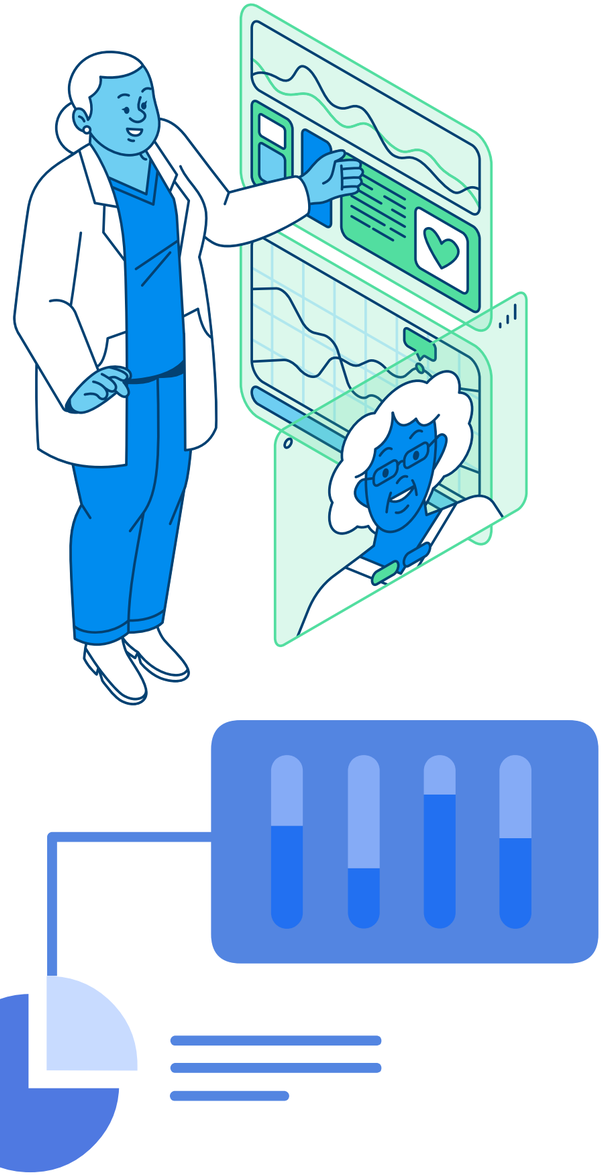


# Principle Component Analysis (PCA)



# BUILDING A MODEL SCENARIO (MEMBANGUN SKENARIO MODEL)



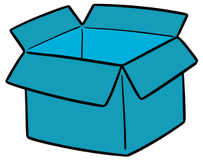
# Modeling approach

| Element | Description |
|---|---|
| Problem Type | Supervised Learning – Binary Classification |
| Target Variable | Leukemia_Type: AML (1), ALL (0) |
| Input Features | PCA-transformed gene expression data |
| ML Algorithms | 9 classifier Models |
| Validation Strategy | Stratified K-Fold Cross-Validation (e.g., 5 folds) on training set |
| Evaluation Metrics | Accuracy, ROC-AUC |

# Preprocessing

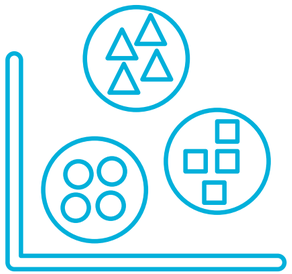
## Handling Outliers

### Winsorization



## Dimensional reduction

### PCA



## Handling Imbalance

## Resampling

### SMOTE

## Scaling

### RobustScaler

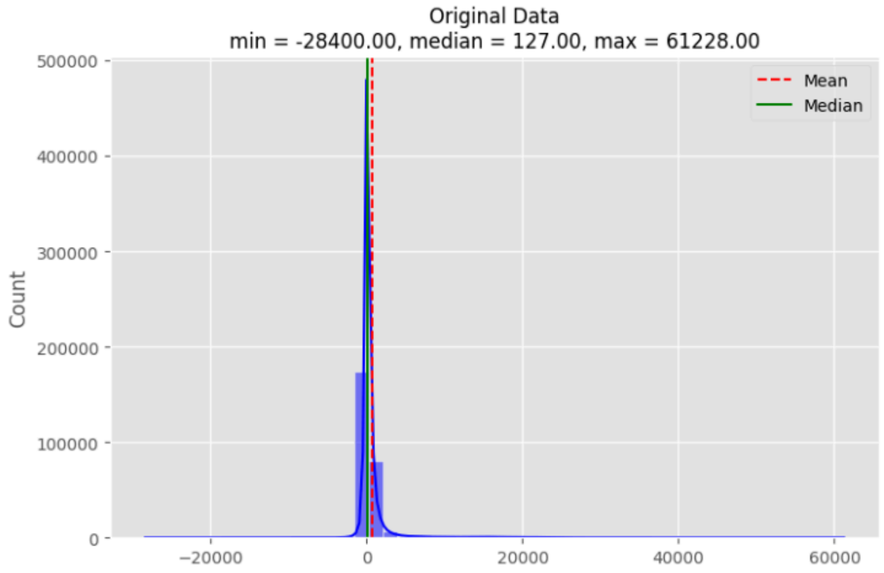
# Outlier handling

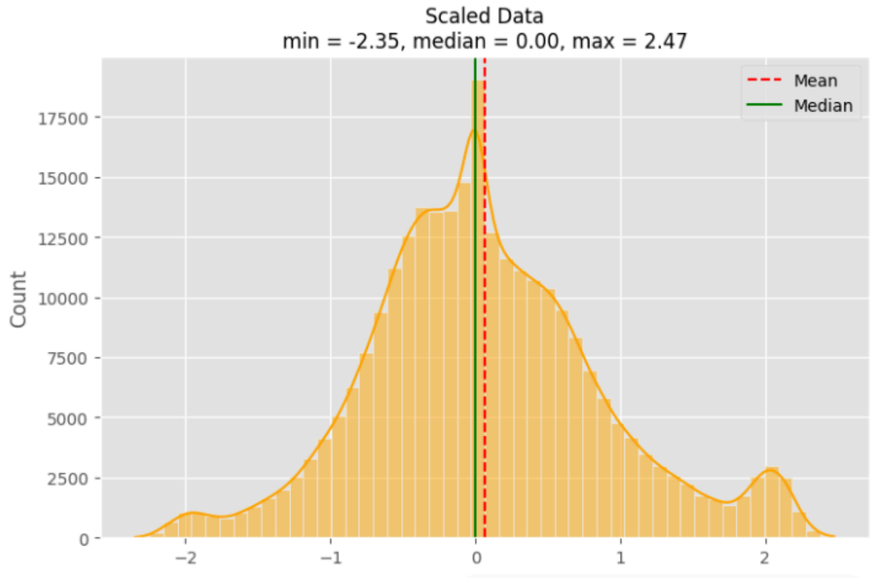
**Winsorization**

- Capping method: IQR
- Tail: Both
- Fold: 1.5

# Scaling

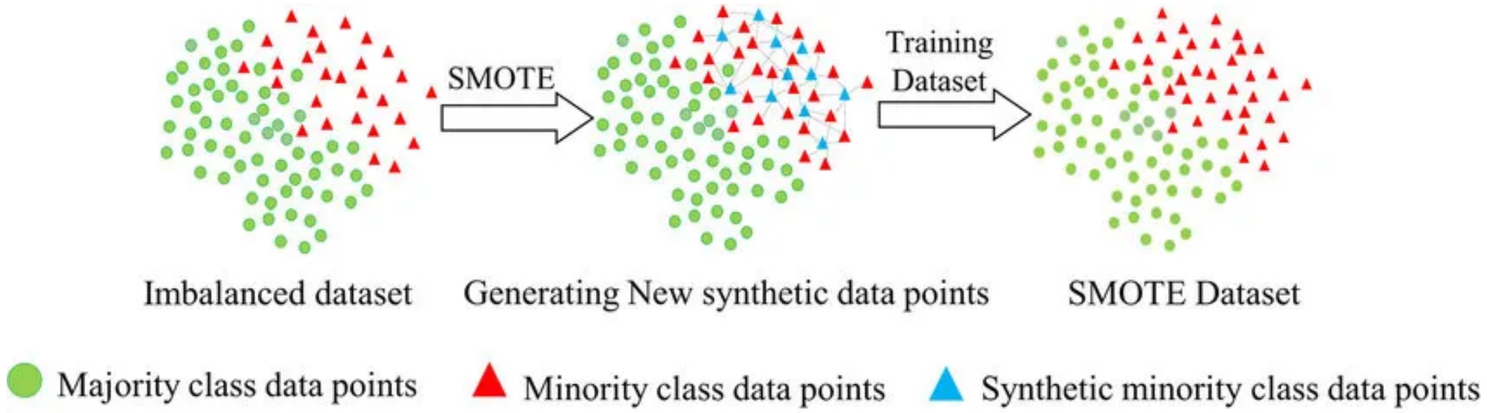
**Scaling → Robust scaler**



# Resampling: SMOTE

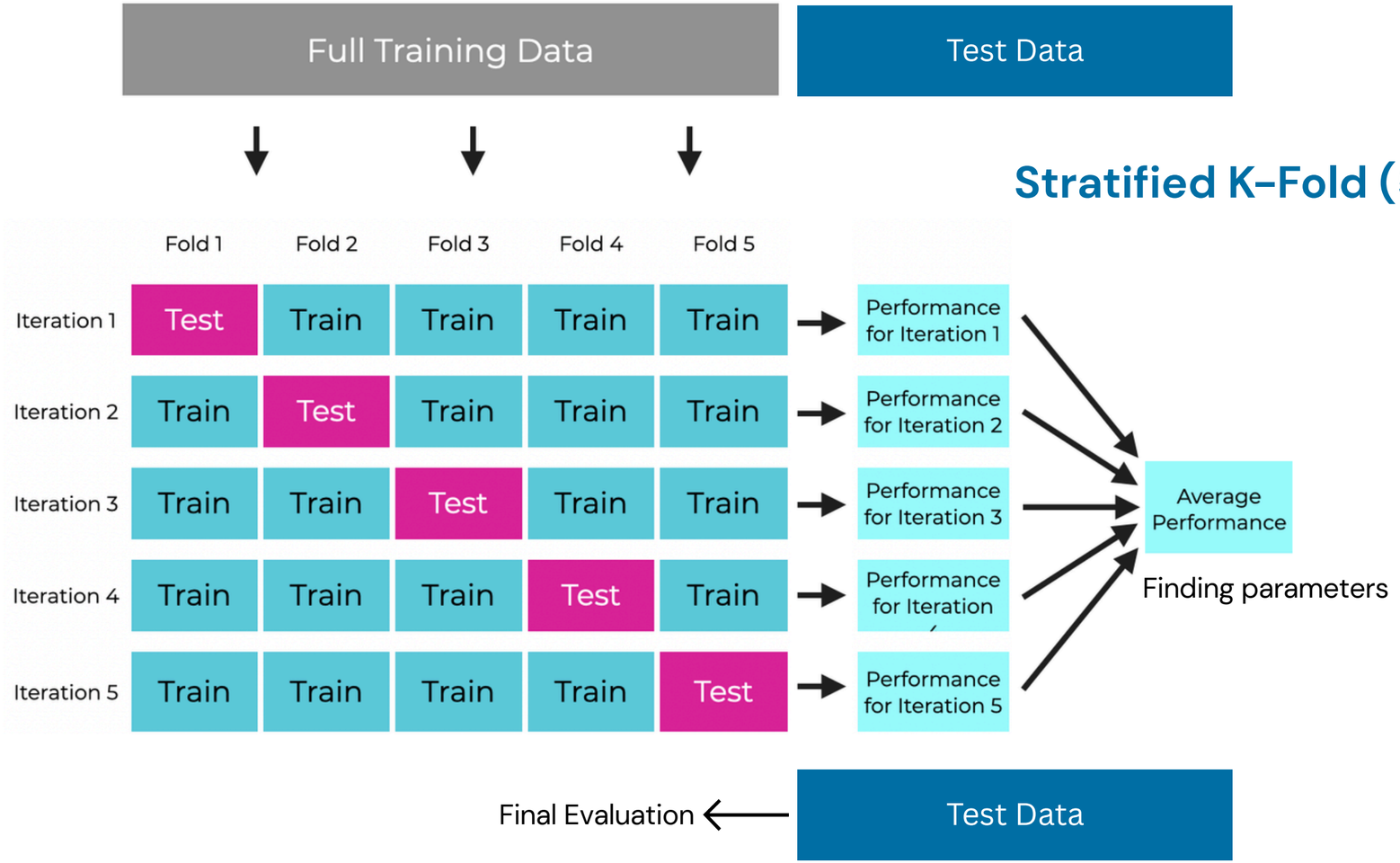
## Synthetic Minority Over-Sampling Technique (SMOTE)




## Class Distribution

```
Before Upsampling:
Counter({0: 27, 1: 11})
After Upsampling:
Counter({0: 27, 1: 27})
```

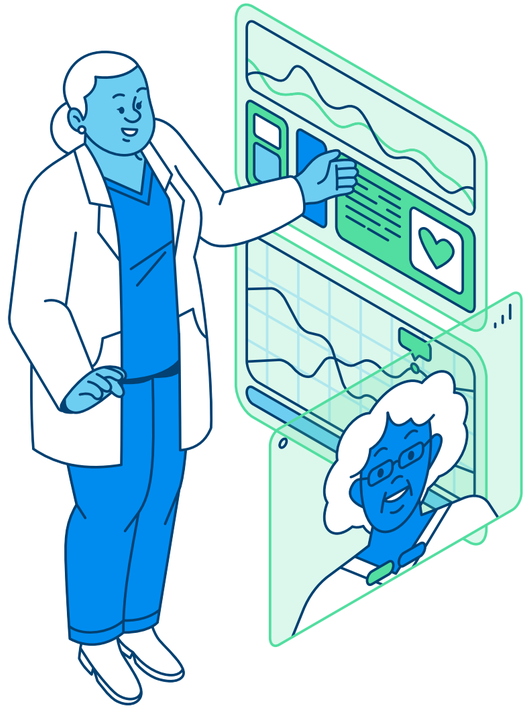
# Cross validation

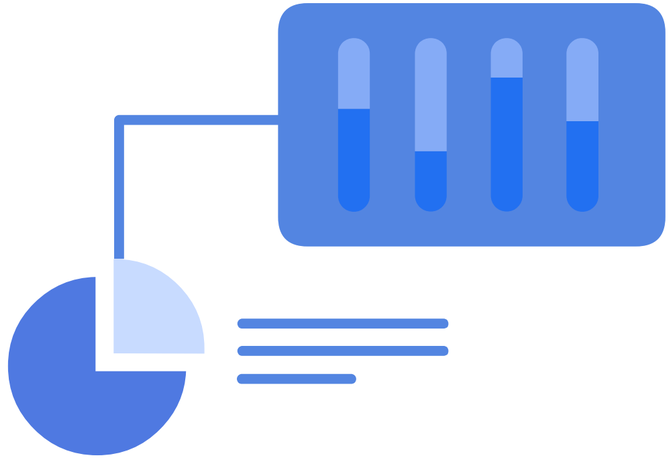


# Benchmark Models

| No | Model | Type | Characteristics |
|---|---|---|---|
| 1 | Logistic Regression | Linear | Fast, interpretable, but limited for complex/nonlinear patterns |
| 2 | KNN | Distance-based | Classification based on the closest training examples |
| 3 | Decision Tree (DT) | Tree | Simple, interpretable, high-variance (prone to overfitting) |
| 4 | Random Forest | Tree + Bagging | Ensemble of DT trained on random subsets of data and features |
| 5 | AdaBoost | Tree + Boosting | Early boosting algorithm using weighted weak learners |
| 6 | Gradient Boosting | Tree + Boosting | Generic gradient boosting method |
| 7 | XGBoost | Tree + Boosting | Optimized gradient boosting with regularization and pruning |
| 8 | LightGBM | Tree + Boosting | Faster, memory-efficient version of gradient boosting. |
| 9 | CatBoost | Tree + Boosting | Specialized for categorical features, but still effective for numeri |

# MODEL BUILDING (MEMBANGUN MODEL)



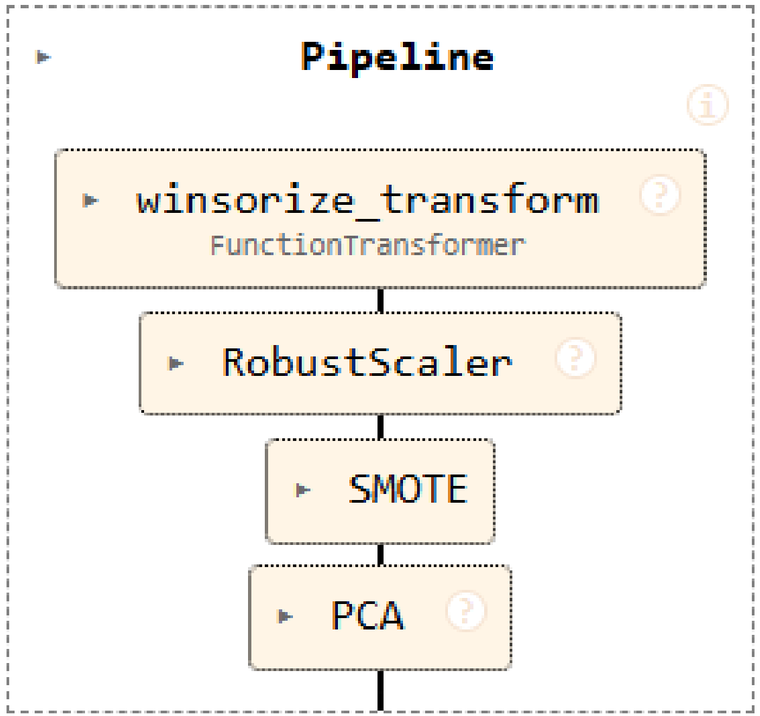


# Modeling

## Modeling strategy

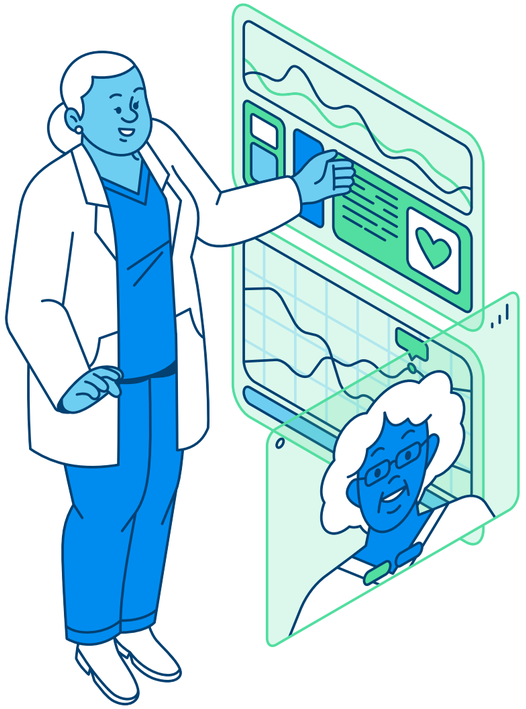
### Pipeline

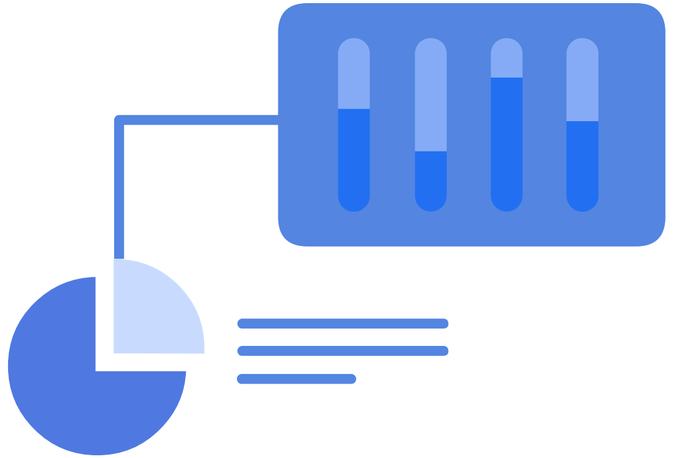
```
pipe_model = Pipeline([
    ('handling outlier', winsorizer),
    ('scaler', scaler),
    ('resamplers', smote),
    ('pca', pca),
    ('algo', model)
])
```




- **Training set & testing set:** 38 patients data for training, 34 patients data for testing
- **Feature set**: 31 Principal components (PC), capturing 95% variance
- **Resampling applied**: SMOTE to balance class distribution in training data
- **Cross-validation:** Stratified 5-Fold Cross-Validation

# MODEL EVALUATION (MENGEVALUASI HASIL PEMODELAN)





# Metrics used for experiments

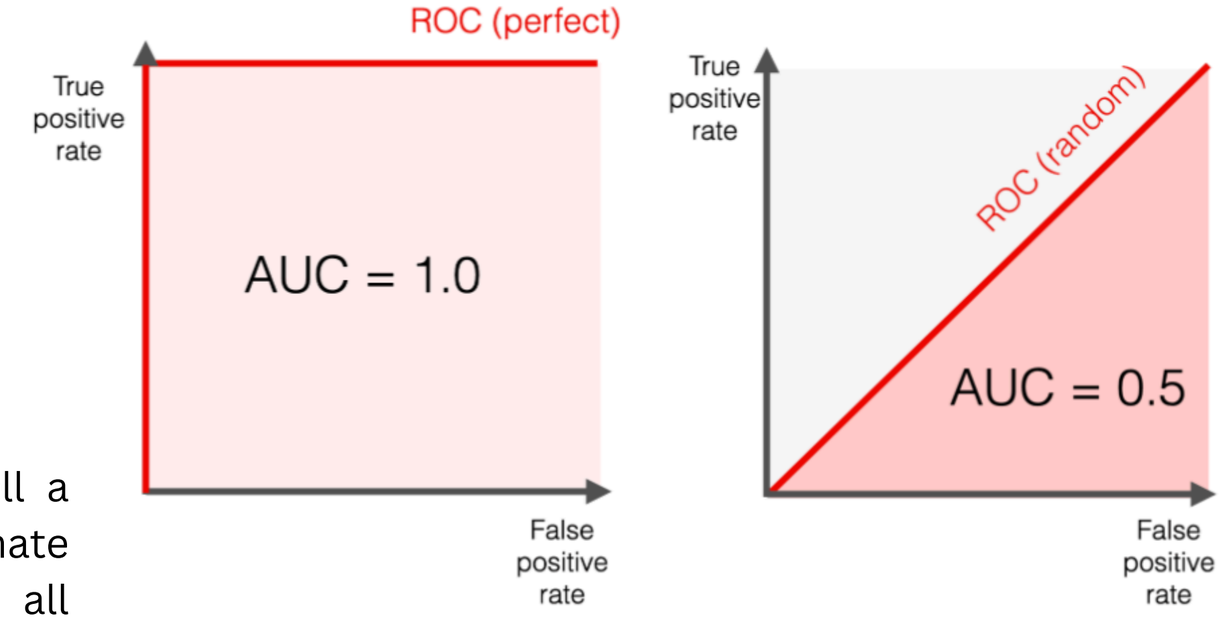
$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy shows how often a classification ML model is correct overall.

evidentlyai.com

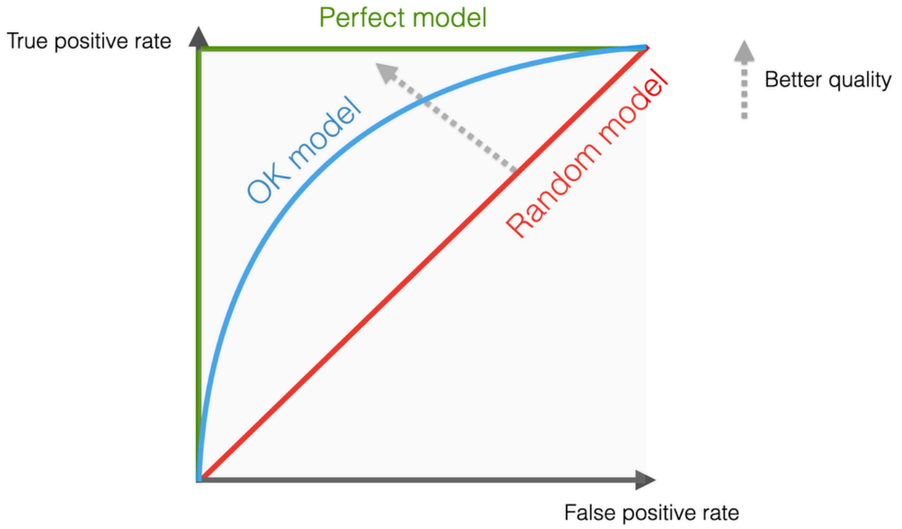
# Metrics : ROC AUC score




**ROC AUC Score**

- The area under the ROC curve. It sums up how well a model can produce relative scores to discriminate between positive or negative instances across all classification thresholds.
- The ROC AUC score **ranges from 0 to 1**,
  - 0.5 → Random guessing
  - 0.8 → Good Performance
  - 0.9 → Great performance
  - 1 → Perfect performance.

evidentlyai.com



# Model Performance

## Accuracy score

| | model | accuracy mean (PCA) | accuracy std (PCA) |
|---|---|---|---|
| 0 | Logistic Regression | 1.000 | 0.000 |
| 1 | KNN | 0.943 | 0.114 |
| 8 | LightGBM | 0.896 | 0.147 |
| 3 | Random Forest | 0.893 | 0.096 |
| 2 | Decision Tree | 0.893 | 0.054 |
| 4 | AdaBoost | 0.868 | 0.080 |
| 5 | Gradient Boosting | 0.868 | 0.080 |
| 6 | CatBoost | 0.868 | 0.080 |
| 7 | XGBoost | 0.868 | 0.080 |

## ROC AUC score

| | model | ROC_AUC mean (PCA) | ROC_AUC std (PCA) |
|---|---|---|---|
| 0 | Logistic Regression | 1.000 | 0.000 |
| 1 | KNN | 1.000 | 0.000 |
| 6 | CatBoost | 1.000 | 0.000 |
| 4 | AdaBoost | 0.980 | 0.040 |
| 8 | LightGBM | 0.960 | 0.080 |
| 3 | Random Forest | 0.953 | 0.093 |
| 7 | XGBoost | 0.933 | 0.076 |
| 5 | Gradient Boosting | 0.930 | 0.098 |
| 2 | Decision Tree | 0.880 | 0.084 |

# Model Performance – Testing dataset

## Accuracy (Training VS Testing)

| | model | Mean Accuracy - Training Dataset | Mean Accuracy - Testing Dataset |
|---|---|---|---|
| **7** | XGBoost | 0.868 | 0.971 |
| **8** | LightGBM | 0.896 | 0.971 |
| **0** | Logistic Regression | 1.000 | 0.941 |
| **3** | Random Forest | 0.893 | 0.941 |
| **4** | AdaBoost | 0.868 | 0.912 |
| **6** | CatBoost | 0.868 | 0.912 |
| **2** | Decision Tree | 0.893 | 0.882 |
| **5** | GradienBoost | 0.868 | 0.882 |
| **1** | KNN | 0.943 | 0.824 |

# Model Performance – Testing dataset

## ROC-AUC score (Training vs testing)

| | model | ROC-AUC Score - Train data | ROC-AUC Score - Test data |
|---|---|---|---|
| **7** | XGBoost | 0.933 | 0.989 |
| **8** | LightGBM | 0.960 | 0.989 |
| **3** | Random Forest | 0.953 | 0.986 |
| **0** | Logistic Regression | 1.000 | 0.982 |
| **6** | CatBoost | 1.000 | 0.975 |
| **1** | KNN | 1.000 | 0.966 |
| **4** | AdaBoost | 0.980 | 0.961 |
| **5** | GradienBoost | 0.930 | 0.930 |
| **2** | Decision Tree | 0.880 | 0.889 |

# Model Performance – Testing dataset

## Accuracy vs ROC-AUC score

| | model | Accuracy score - Test data | ROC-AUC Score - Test data |
|---|---|---|---|
| 7 | XGBoost | 0.971 | 0.989 |
| 8 | LightGBM | 0.971 | 0.989 |
| 0 | Logistic Regression | 0.941 | 0.982 |
| 3 | Random Forest | 0.941 | 0.986 |
| 4 | AdaBoost | 0.912 | 0.961 |
| 6 | CatBoost | 0.912 | 0.975 |
| 2 | Decision Tree | 0.882 | 0.889 |
| 5 | GradienBoost | 0.882 | 0.930 |
| 1 | KNN | 0.824 | 0.966 |

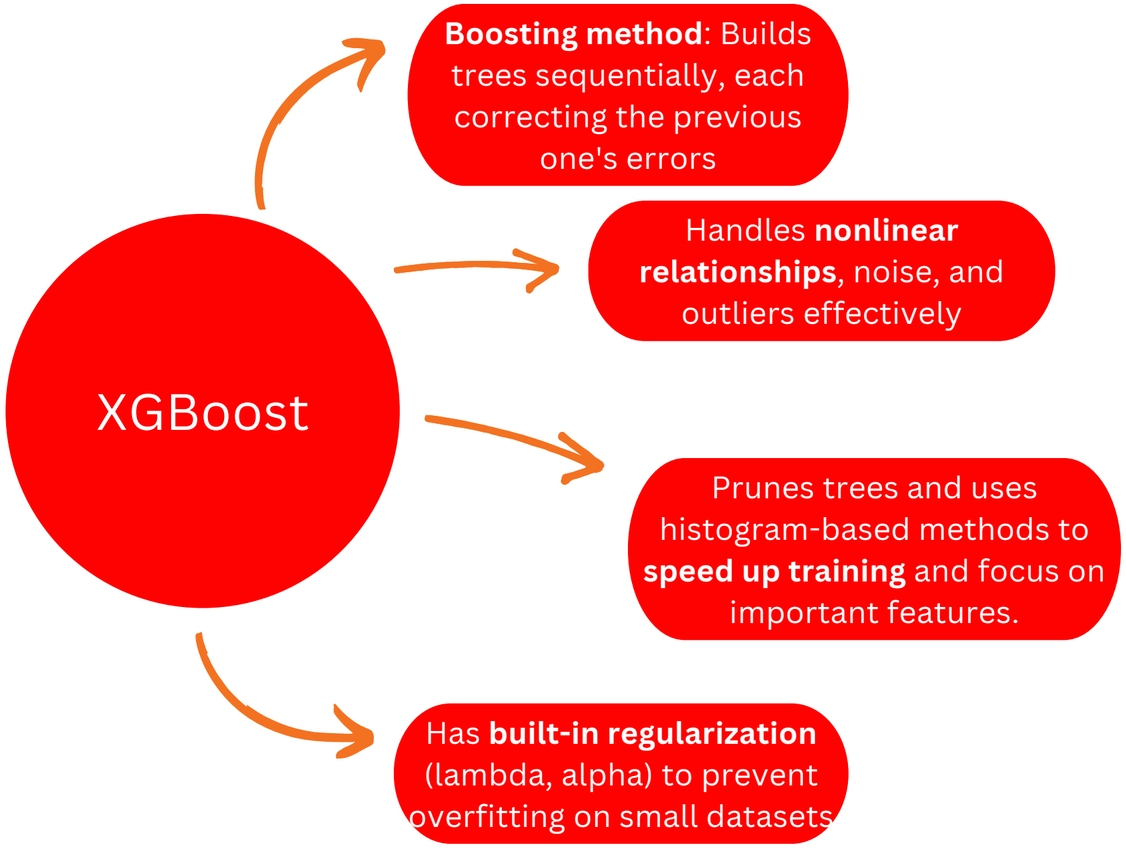

XGBoost


LightGBM

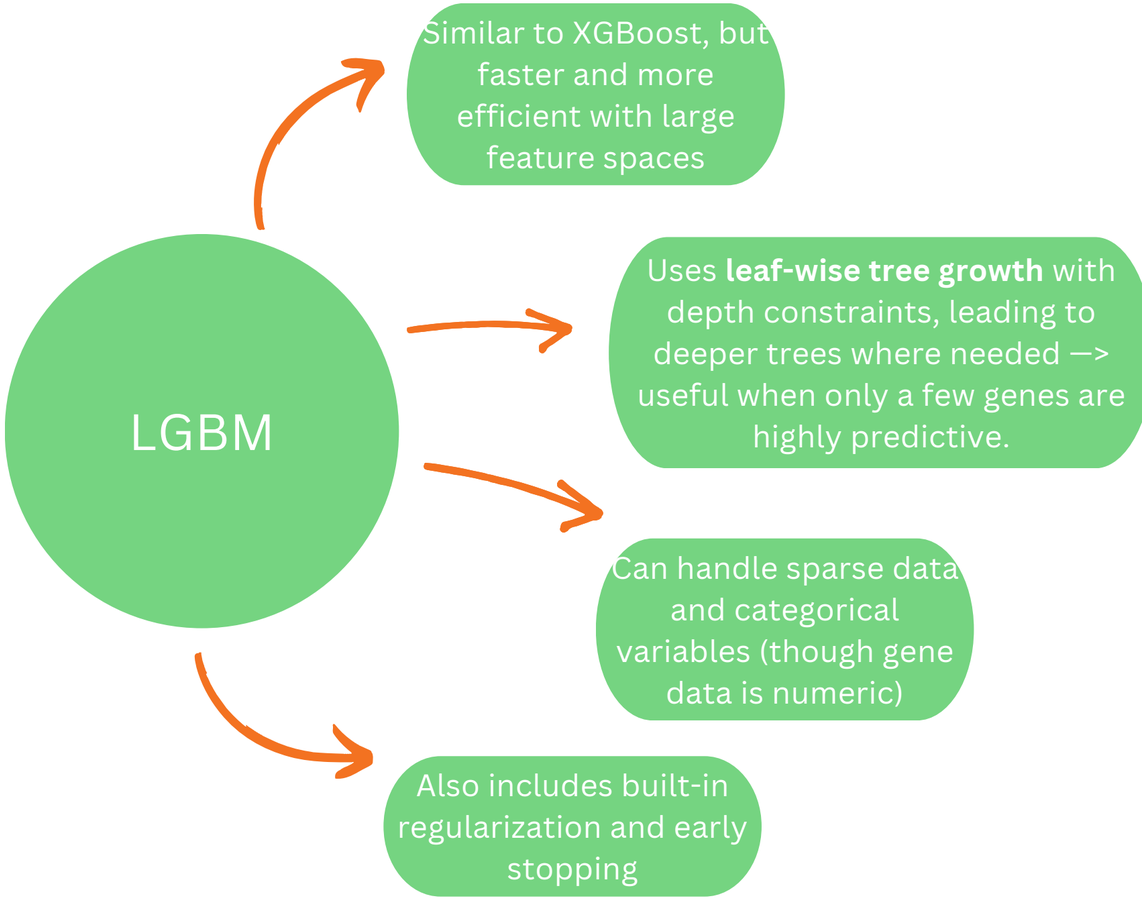

Random Forest

# Top Models

## EXtreme Gradient Boosting (XGBoost)

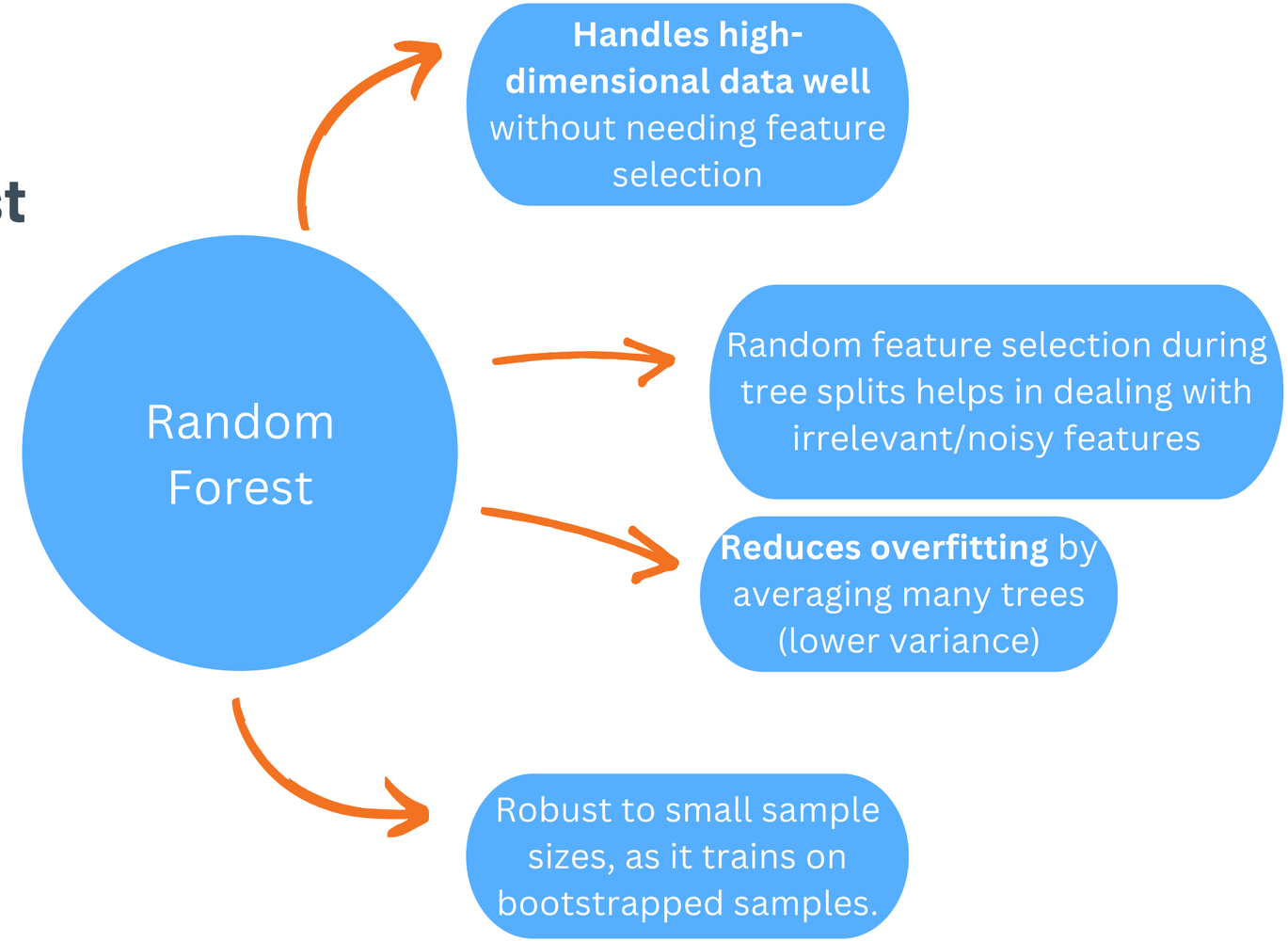



## LightGBM (LGBM)



# Top Models

## Random Forest



# Hyperparameter tuning

## XGBoost

```
hyperparam_space_xgboost = [{
    'model__max_depth': [3, 5, 7, 9],  # Expanding search space
    'model__learning_rate': [0.1, 0.05, 0.01, 0.001],  # Including smaller learning rates
    'model__n_estimators': [100, 150, 200, 250],  # Higher n_estimators for lower learning rates
    'model__subsample': [0.8, 0.9, 1.0],  # Exploring subsample rates
    'model__colsample_bytree': [0.8, 0.9, 1.0],  # Exploring column sampling
    'model__gamma': [0, 0.1, 0.2, 0.5],  # Adding gamma for regularization
    'model__min_child_weight': [1, 2, 3, 4],  # Adding min_child_weight
}]
```

## LightGBM

```
hyperparam_space_lgbm = [{
    'model__max_depth': [3, 5, 7],  # maximum_depth of trees
    'model__learning_rate': [0.1, 0.05, 0.01],  # More basic learning rates
    'model__n_estimators': [100, 150, 200],  # Reduced number of estimators
    'model__subsample': [0.8, 1.0],  # Fewer options for subsample
    'model__colsample_bytree': [0.8, 1.0],  # Fewer options for colsample_bytree
    'model__num_leaves': [31, 50],  # LightGBM-specific parameter for leaves
    'model__min_child_samples': [10, 20, 30],  # LightGBM-specific parameter for minimum samples in a leaf
}]
```

## Random Forest

```
hyperparam_space_rf = [{
    'model__n_estimators': [100, 200, 300],  # Number of trees
    'model__max_depth': [None, 10, 20, 30],  # Maximum depth of trees
    'model__min_samples_split': [2, 5, 10],  # Minimum number of samples to split an internal node
    'model__min_samples_leaf': [1, 2, 4],  # Minimum number of samples per leaf node
    'model__max_features': ['sqrt', 'log2'],  # Number of features to consider at each split
    'model__bootstrap': [True, False],  # Whether bootstrap samples are used
}]
```

# Hyperparameter Tuning

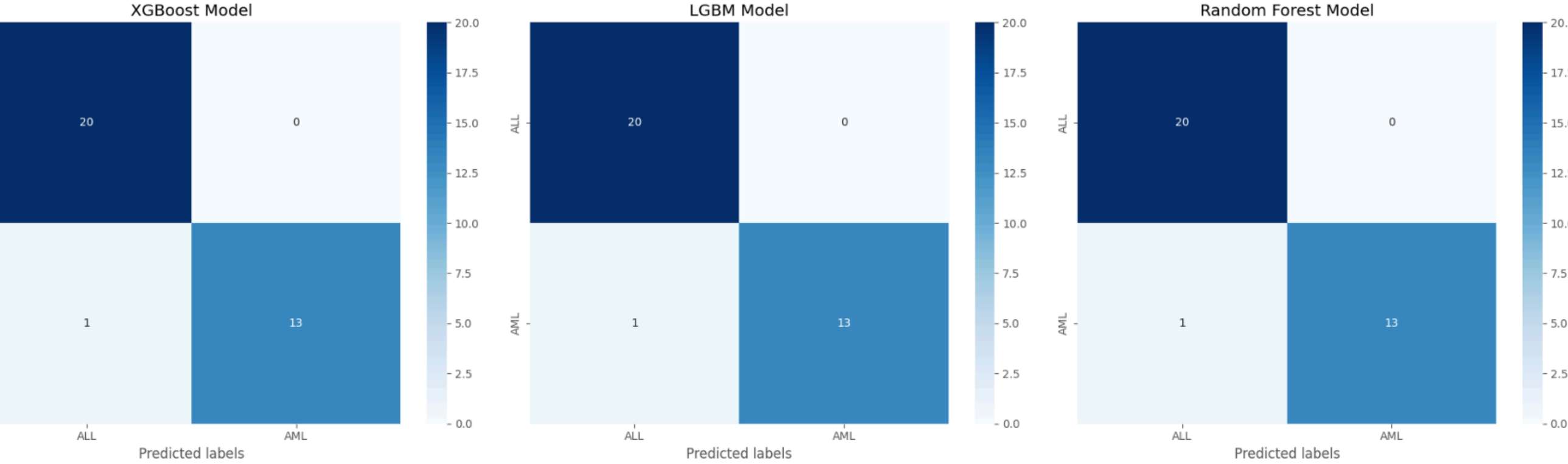
## Overal Accuracy score comparison

| Model | Conditions | Train score | Test score |
|---|---|---|---|
| XGBoost Classifier | Before Tuning | 0.868 | 0.971 |
| XGBoost Classifier | After Tuning | 0.893 | 0.971 |
| LGBM Classifier | Before Tuning | 0.896 | 0.971 |
| LGBM Forest Classifier | After Tuning | 0.893 | 0.971 |
| Random Forest Classifier | Before Tuning | 0.871 | 0.941 |
| Random Forest Classifier | After Tuning | 0.975 | 0.971 |

## Overall ROC AUC score comparison

| Model | Before Tunning | After Tunning |
|---|---|---|
| XGBoost Classifier | 0.989 | 0.989 |
| LGBM Classifier | 0.989 | 0.982 |
| Random Forest Classifier | 0.986 | 0.993 |

# Confusion matrix



# Classification report

### XGBoost

The Classification Report of XGBoost Classifier

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 1.00 | 0.98 | 20 |
| 1 | 1.00 | 0.93 | 0.96 | 14 |
| accuracy | | | 0.97 | 34 |
| macro avg | 0.98 | 0.96 | 0.97 | 34 |
| weighted avg | 0.97 | 0.97 | 0.97 | 34 |

### Random Forest

The Classification Report of Random Forest Classifier

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 1.00 | 0.98 | 20 |
| 1 | 1.00 | 0.93 | 0.96 | 14 |
| accuracy | | | 0.97 | 34 |
| macro avg | 0.98 | 0.96 | 0.97 | 34 |
| weighted avg | 0.97 | 0.97 | 0.97 | 34 |

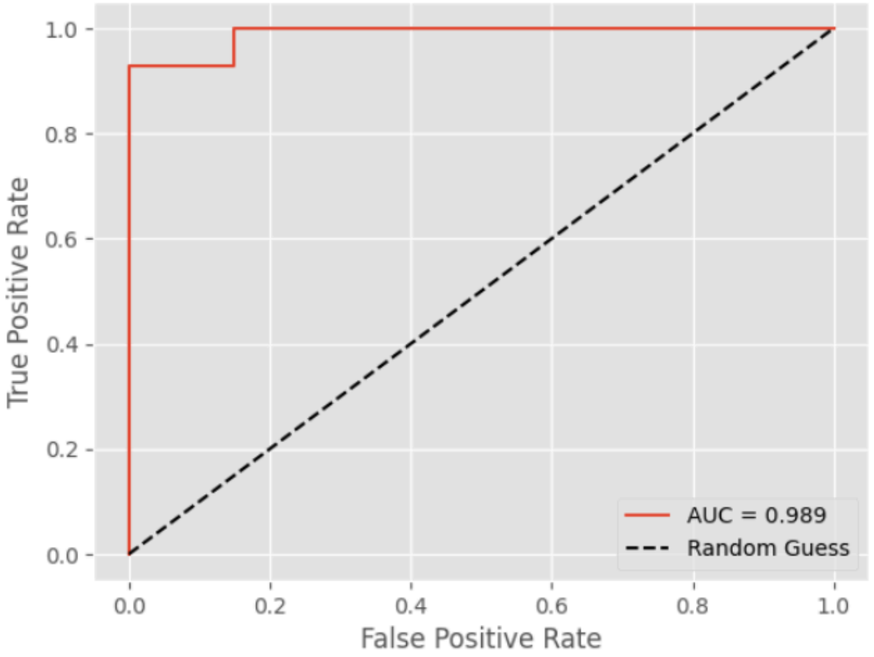
### LGBM

The Classification Report of LGBM Classifier

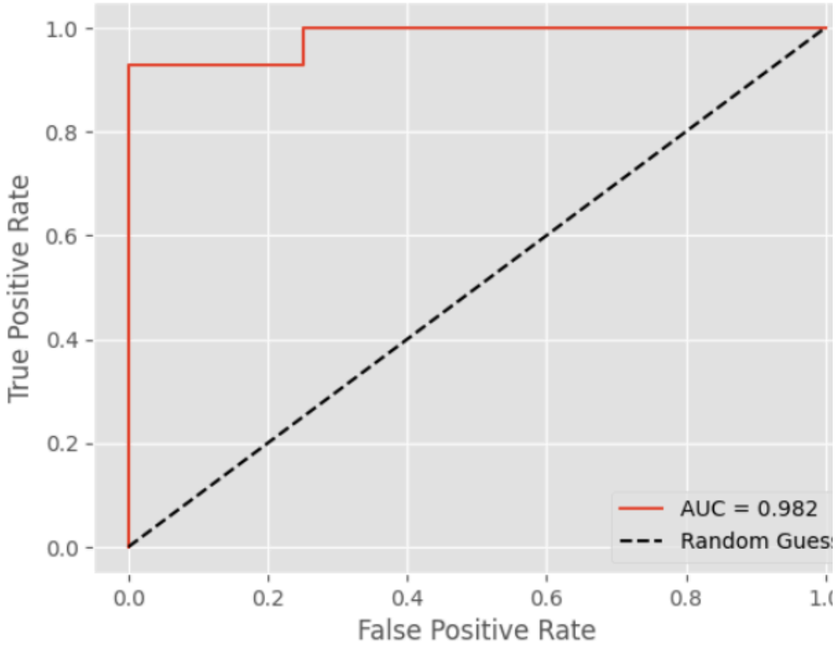
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 1.00 | 0.98 | 20 |
| 1 | 1.00 | 0.93 | 0.96 | 14 |
| accuracy | | | 0.97 | 34 |
| macro avg | 0.98 | 0.96 | 0.97 | 34 |
| weighted avg | 0.97 | 0.97 | 0.97 | 34 |

# ROC-AUC Curve

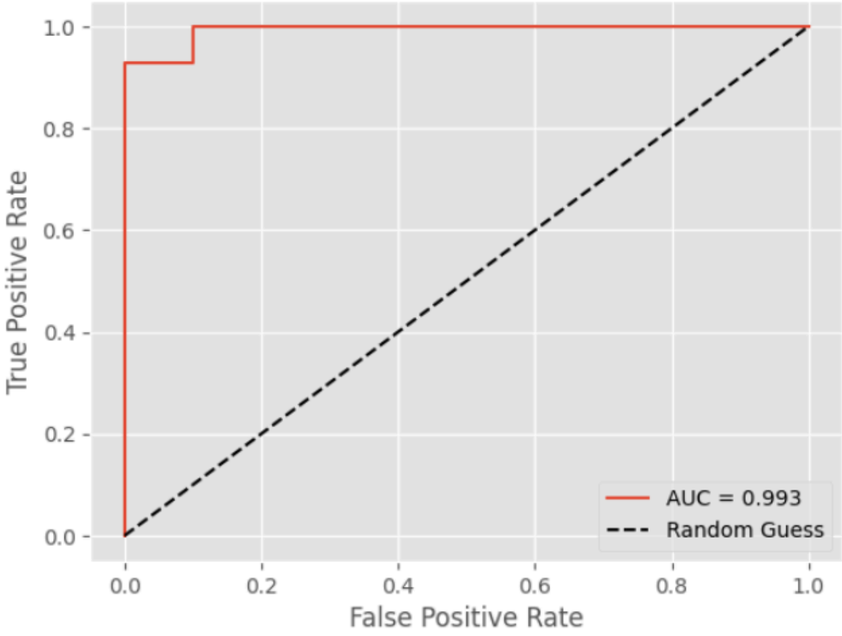
## XGBoost



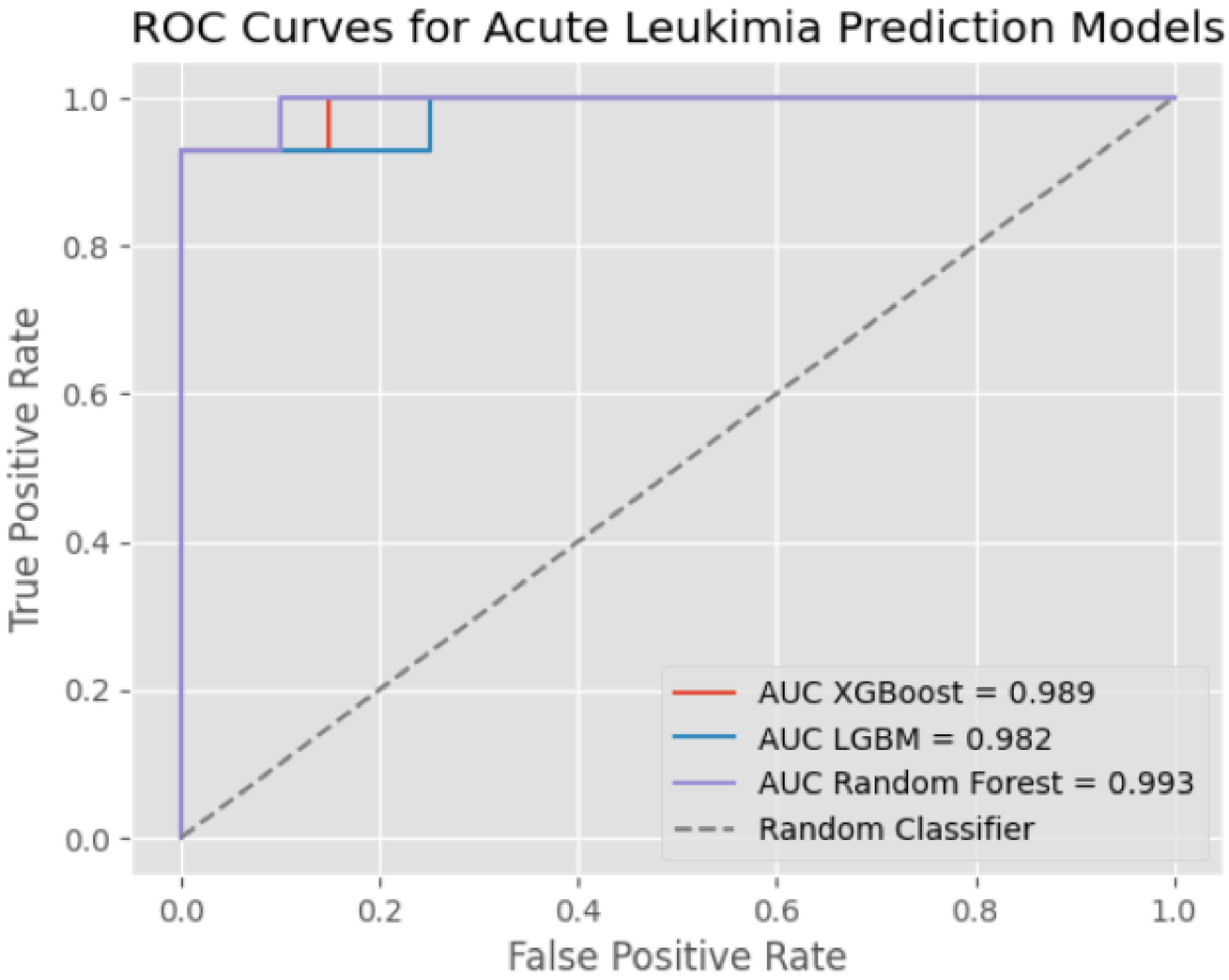

## LGBM




## Random Forest



# ROC-AUC Curve



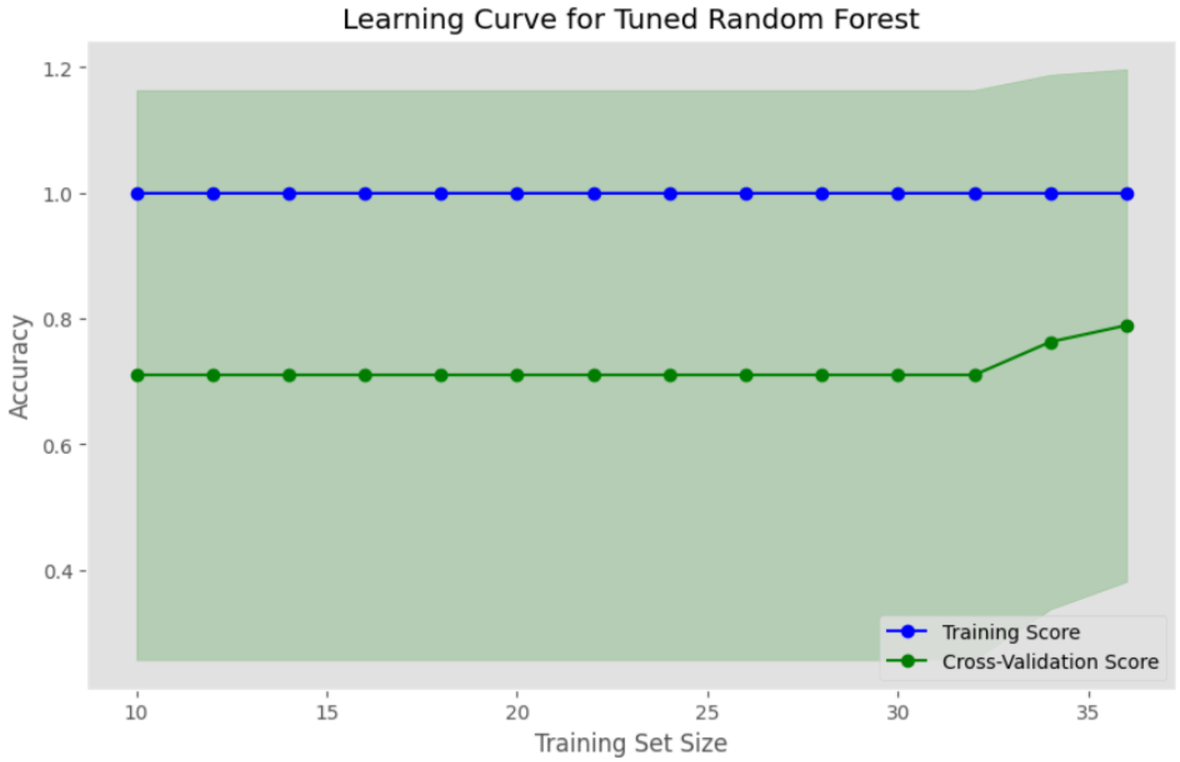
# Selected Model

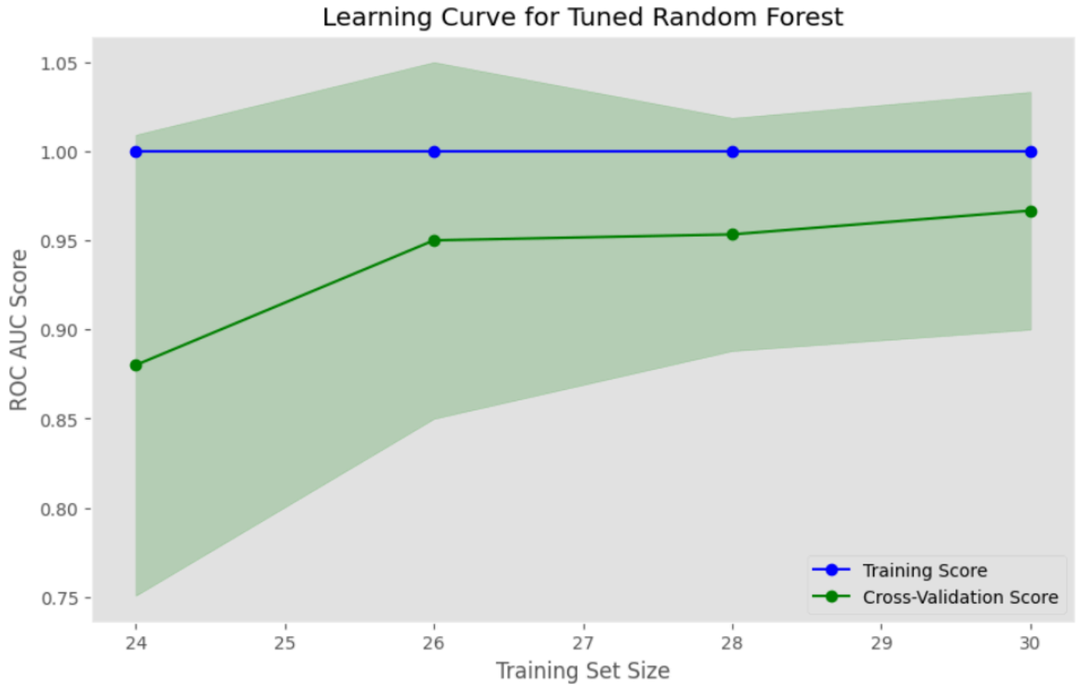
Tuned Random Forest is the best model for predicting acute leukemia subtype in this scenario, due to its high accuracy of 0.971 and best ROC AUC score of 0.993.

Best model parameters:

- n_estimators: 300
- min_samples_split: 5
- min_samples_leaf: 1
- max_features': 'log2'
- max_depth': 30
- model__bootstrap': True

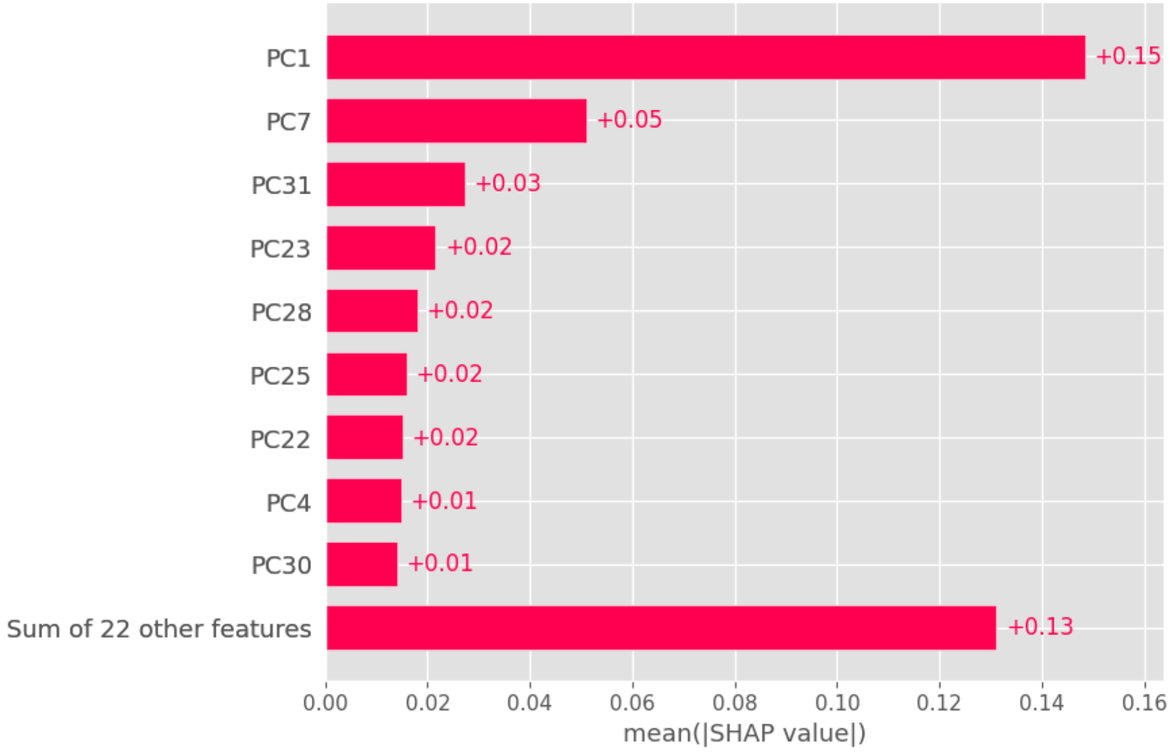
# Tuned Random Forest Learning Curve



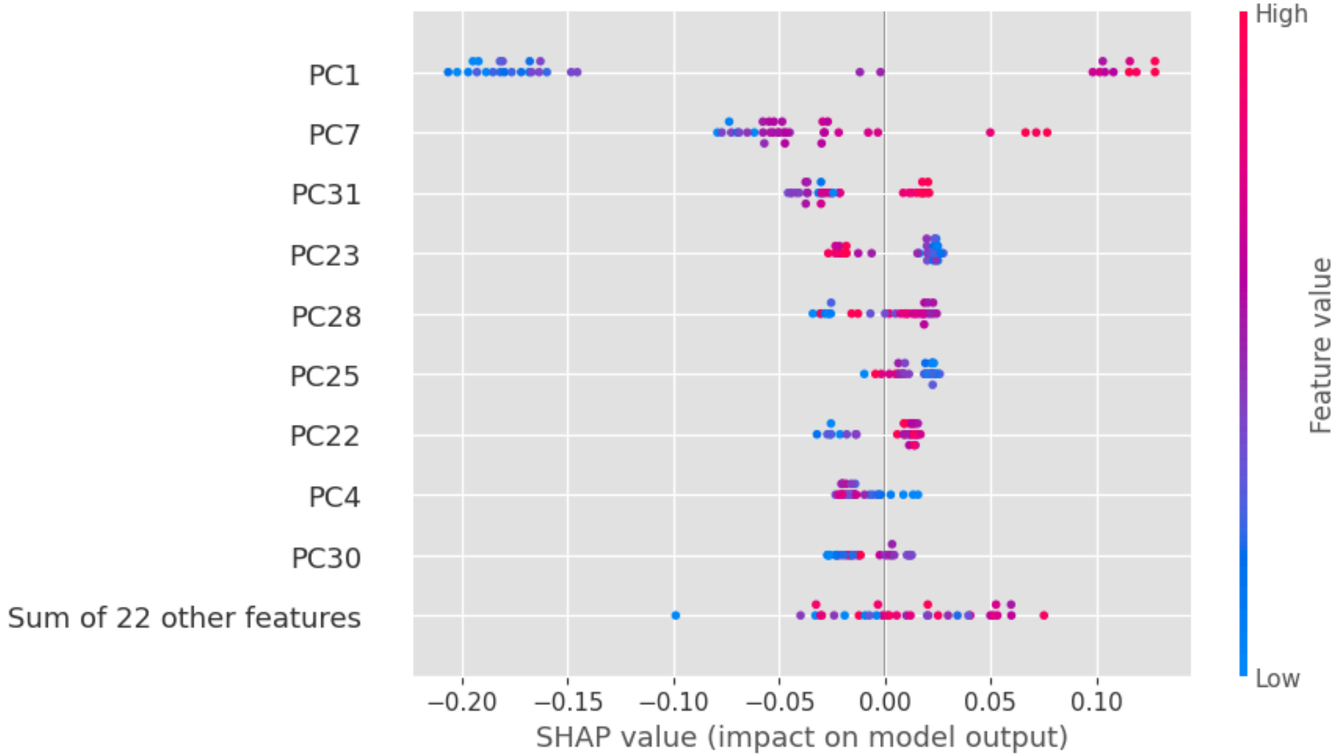



# SHAP Tuned RFModel (Shapley Additive exPlanations)

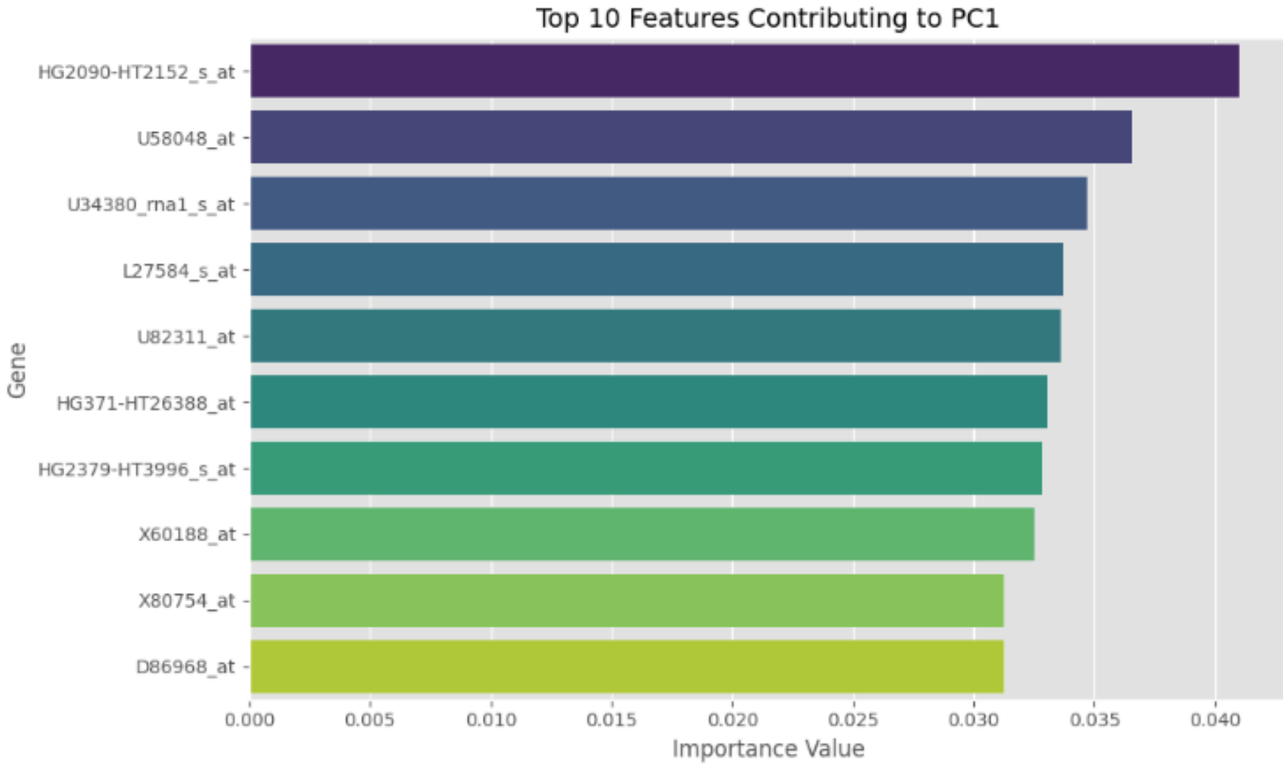
## SHAP Bar plot

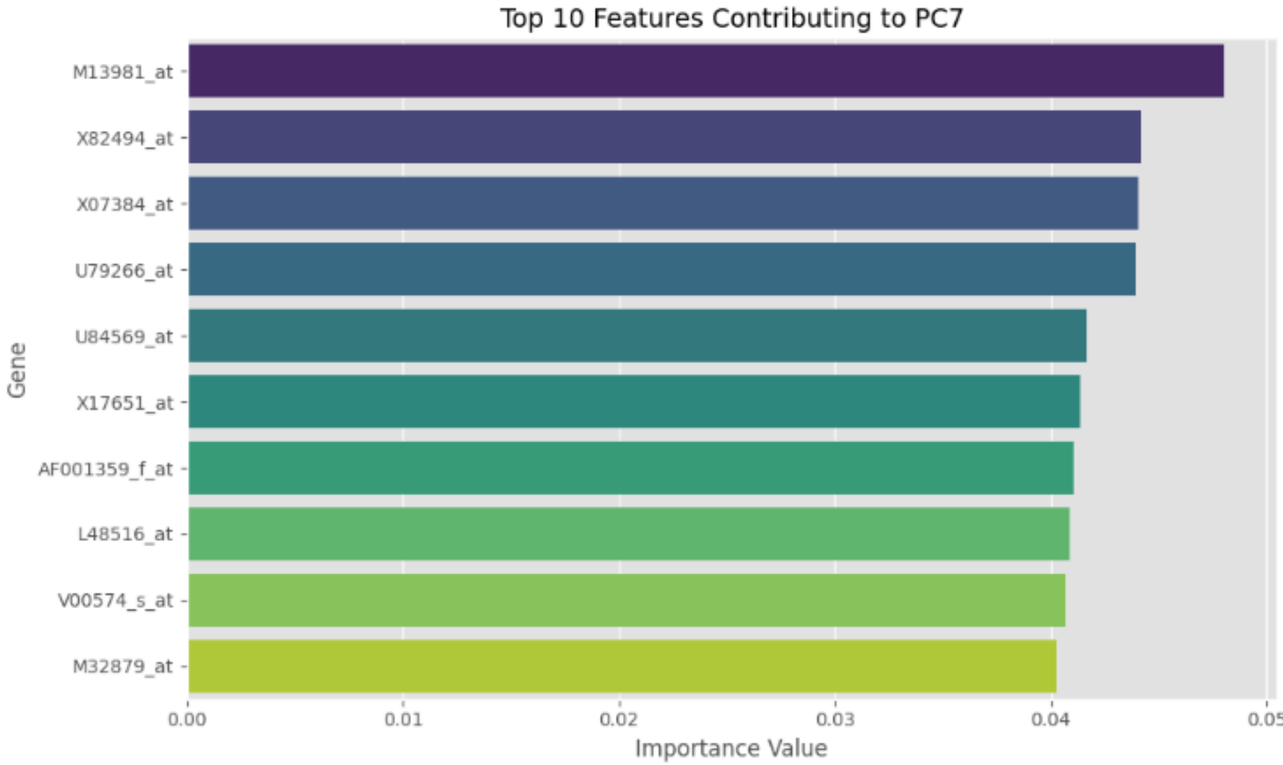



## SHAP Bee Swarm plot



# Contributor genes for Tuned RF



# Contributor genes & Roles in Leukemia/Cancer

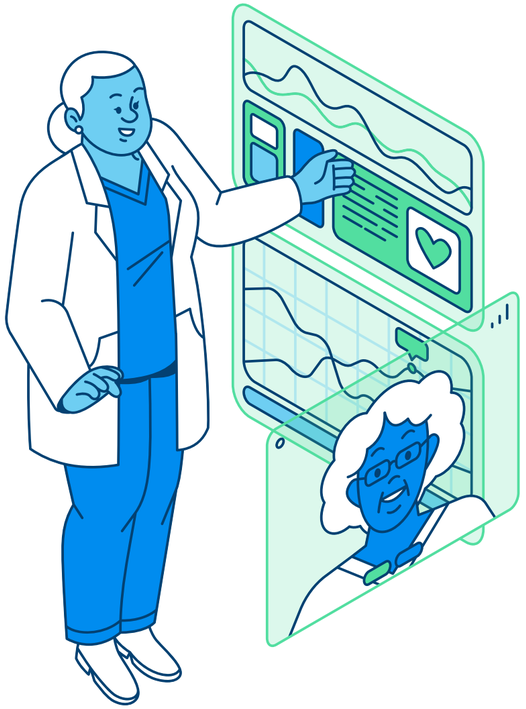
## PC1 contributor genes (sample)

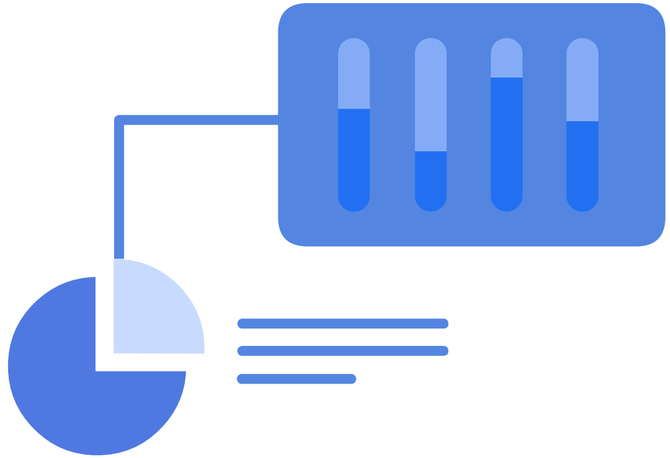
| Gene Code | Gene Name | Gene Function | Role in Model Prediction |
|---|---|---|---|
| HG2090-HT2152_s_at | External Membrane Protein, 130 KDa | Involved in cellular membrane processes. | Cellular interactions and signaling. |
| U58048_at | PRSM1 Metallopeptidase 1 (33 kD) | Enzyme involved in protein degradation and turnover. | Protein turnover in cancer cells. |
| U34380_rna1_s_at | TEC (Tyrosine Kinase) Gene | Tyrosine kinase implicated in signal transduction pathways. | Cancer cell signaling. |

## PC7 contributor genes (sample)

| Gene Code | Gene Name | Gene Function | Role in Model Prediction |
|---|---|---|---|
| M13981_at | INHA Inhibin, alpha | Regulates hormone production and inhibits cell growth, linked to cancer progression | Regulation of cell growth in leukemia |
| X82494_at | FBLN2 Fibulin 2 | Involved in extracellular matrix organization and cell adhesion | ECM remodeling relevant to cancer cells |
| X07384_at | GLI Glioma-associated oncogene | Zinc finger transcription factor, regulates cell proliferation and differentiation | Oncogene associated with leukemia |

# REVIEWING THE MODELING PROCESS (MELAKUKAN REVIEW PEMODELAN)





# Review Modeling process

| Aspect | Review Result |
|---|---|
| Validity | All steps are justified and documented |
| Reproducibility | Pipeline used from preprocessing to prediction, can be reproduced |
| Metrics evaluation | Model Accuracy ≥ 95% (RF = 97.1%) and ROC-AUC ≥ 0.95 (RF = 0.993) → Meet the requirement |
| Generalization | No significant difference between training and testing (Acc: 97.5% → 97.1%, ROC: 0.986 → 0.993) |
| Interpretability | Dimensionality reduced using PCA → PCA components contain genes that relevant to Leukemia |

# Review Modeling process

**Opportunities for Improvement**

- More hyperparameter tuning for top models.
- Experiment with other types of models that are good for handling a high-dimensional dataset
- Experiment with SMOTE Variants and compare resampling vs class weighting

**Potential next steps**

- Apply the model to a new biomedical dataset
- Build a web-based prediction interface (demo)

# THANK YOU


My Contact

https://github.com/harishmuh

www.linkedin.com/in/harish-muhammad-7b600b102/

harishmuh@gmail.com