# Data dictionary

**leukemia.csv**

The dataset contains gene expression of various leukemia patients and contains gene expression of various leukemia patients on 39 selected locations of the human genome. These genome positions refer to the genes NPM1, RUNX1, HOXA1, . . ., HOXA11, HOXA13. These genes are commonly known to be relevant for leukemia. This genomic data is the basis on which doctors obtain their diagnosis of whether a patient has leukemia.

Your task is to build an SVM classifier that decides for each patient whether or not they have blood cancer.

|  |  |
| --- | --- |
| **Name of the variable** | **Description** |
| Patient\_ID | Identification number of the patient |
| number\_at | Gene expression on the particular location on the human genome. These position identifiers consist of a cryptic number and “at” at the end. For example, “1563591\_at”. There are 39 different positions in the dataset |
| Leukemia | The type of leukemia of the patient. One of the following: AML, ALL, CML, CLL, Nonleukemia |

## List of options for fine-tuning the kernel function parameter in the SVM node

|  |  |  |
| --- | --- | --- |
| kernel Type | Parameter | Description |
| Linear | Regularization C | Controls this trade-off between a wide margin and a small number of misclassified points for which the optimal balance has to be found. A typical value should lie between 1 and 10, with 10 being the default. Increasing the regularization parameter improves the accuracy. However, this can also lead to overfitting. |
| RBF | Gamma | Typically, the value should be between 3/k and 6/k, where k is the number of input fields. Increasing the value improves the accuracy. However, this can also lead to overfitting. |
| Polynomial  Sigmoid | Gamma | Increasing this value improves the accuracy. However, this can also lead to overfitting. |
| Bias | Defines the constant value in the kernel function. The default value is 0, which is suitable in most cases. |
| Polynomial | Degree | Defines the degree of the polynomial kernel. This parameter controls the dimension of the mapping space and is normally less or equal to 10 |