The University of Economics, Prague

(Vysoká škola ekonomická v Praze)

Knowledge Discovery in Databases:

RapidMiner and BigML

Course **4IZ451**

Project 1 – May 18. 2025

Bc. Smaliak Danila

Table of contents

List of Figures and Tables

Figures

Tables

1. Introduction

The purpose of this project is to conduct a comparative analysis of basic classification algorithms using two popular AutoML tools — RapidMiner (Altair AI Studio) and BigML. These platforms provide accessible interfaces for building, training, and evaluating machine learning models, which makes them particularly suitable for students and users without extensive programming experience.

In this study, no data preprocessing or algorithm parameter tuning was applied. All models were trained using default settings provided by each platform. The goal is to assess the performance of a set of commonly used classification algorithms, namely: Logistic Regression, Naive Bayes, k-Nearest Neighbors (k-NN), Decision Tree, Random Forest (Ensemble method)

The models are compared based on several performance metrics, such as accuracy, confusion matrix, and class-specific measures like precision and recall. In addition to performance, the interpretability of each algorithm is also considered, as it plays a key role in practical applications such as medical diagnostics. Finally, the report includes a reflection on user experience with each platform to evaluate their usability and effectiveness for exploratory data analysis and model development.

2. Basic Information about the Dataset

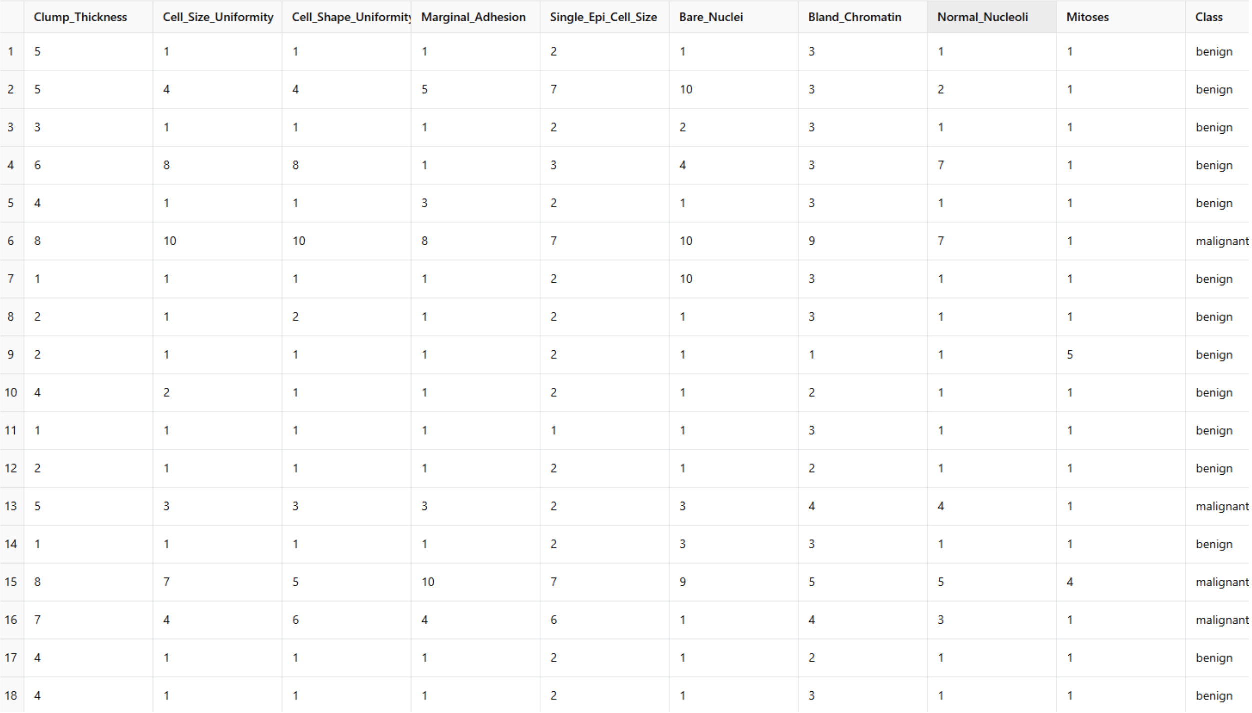
The dataset used in this study is related to medical diagnosis and is widely used in medical machine learning education due to its simplicity and clinical relevance. Each instance in the dataset represents numeric measurements taken from a fine needle aspirate (FNA) of a breast mass.

The key properties of the dataset (Figure 1.) are as follows:

- Name of the dataset: Breast Cancer Wisconsin (Original)
- Number of examples (instances): 699
- Number of attributes (features): 9 numeric attributes
- Target variable (class label): `Class`
- Type of prediction task: Binary classification
- Class values and distribution:
    - malignant — 240 examples (~34.3%)
    - benign — 459 examples (~65.7%)

All values in the dataset are complete — there are no missing values. Each attribute is a numeric score between 1 and 10, reflecting specific characteristics of the cell nuclei (e.g., thickness, uniformity, adhesion). These values are based on human-graded cytological features and are intended to help distinguish between benign and malignant tumor samples.

Figure 1. Dataset table



	Clump_Thickness	Cell_Size_Uniformity	Cell_Shape_Uniformity	Marginal_Adhesion	Single_Epi_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses	Class
1	5	1	1	1	2	1	3	1	1	benign
2	5	4	4	5	7	10	3	2	1	benign
3	3	1	1	1	2	2	3	1	1	benign
4	6	8	8	1	3	4	3	7	1	benign
5	4	1	1	3	2	1	3	1	1	benign
6	8	10	10	8	7	10	9	7	1	malignant
7	1	1	1	1	2	10	3	1	1	benign
8	2	1	2	1	2	1	3	1	1	benign
9	2	1	1	1	2	1	1	1	5	benign
10	4	2	1	1	2	1	2	1	1	benign
11	1	1	1	1	1	1	3	1	1	benign
12	2	1	1	1	2	1	2	1	1	benign
13	5	3	3	3	2	3	4	4	1	malignant
14	1	1	1	1	2	3	3	1	1	benign
15	8	7	5	10	7	9	5	5	4	malignant
16	7	4	6	4	6	1	4	3	1	malignant
17	4	1	1	1	2	1	2	1	1	benign
18	4	1	1	1	2	1	3	1	1	benign

3. RapidMiner – Experiment Setup and Rationale

In this part of the project, the AutoML platform RapidMiner (Altair AI Studio) is used to build and evaluate multiple classification models on the provided dataset. The aim is to test and compare a variety of commonly used classification algorithms under default settings and determine which performs best in terms of accuracy and interpretability.

RapidMiner was selected for its visual interface, which enables the design of machine learning workflows without writing code. It allows users to connect predefined building blocks (operators) to create a complete process — from loading data to training models and evaluating performance. The following classification algorithms were trained and evaluated using the default configuration (no hyperparameter tuning):

- Logistic Regression
- Naive Bayes
- k-Nearest Neighbors (k-NN)
- Decision Tree

- Random Forest (Ensemble method)

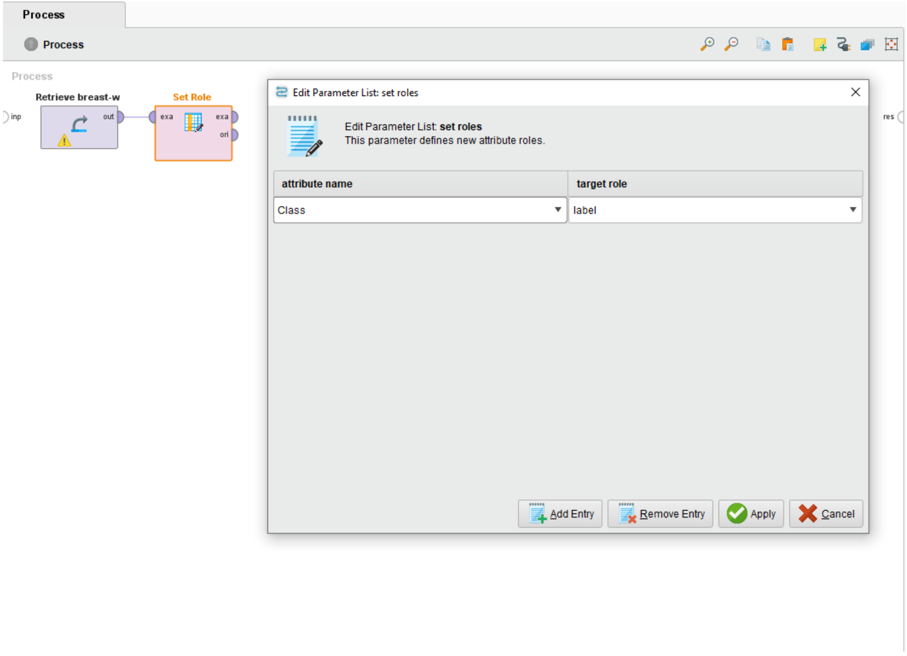
Each model was trained using a train/test split, and their performance was assessed based on overall accuracy, confusion matrix, and additional metrics such as precision and recall. The results are then compared to determine which algorithm offers the best balance between predictive performance and model transparency.

3.1. Dataset Preparation in RapidMiner

Before building classification models, the dataset was first imported into RapidMiner using the Retrieve operator. The dataset was previously saved in CSV format and loaded directly into the RapidMiner workflow.

The next important step was to define the target variable. This was done using the Set Role operator, where the attribute named Class was explicitly assigned the role of label, as shown in Figure 2. This step is essential for supervised learning, as it tells RapidMiner which attribute should be predicted by the model. This completed the initial setup of the dataset. Once the label was defined, the data was ready to be passed into various modeling operators for classification.

Figure 2. *Assigning the target role to the Class attribute*



Edit Parameter List: set roles

Edit Parameter List: **set roles**
This parameter defines new attribute roles.

attribute name	target role
Class	label

3.2. Model Evaluation Setup

To ensure objective and consistent evaluation of all classification models, a Cross Validation operator was used in RapidMiner. This method splits the dataset into multiple folds, where models are trained on a subset of the data and validated on the remaining part. This helps minimize the impact of random variation in train/test splits and provides a more reliable estimate of model performance.

As shown in Figure 3, the process includes multiple Cross Validation branches — one for each algorithm tested:

- Logistic Regression
- Naive Bayes
- k-Nearest Neighbors (k-NN)
- Decision Tree
- Random Forest

Each Cross Validation block receives the preprocessed dataset where the Class attribute is set as the label. Inside each block, the process follows the same logic:

1. Training phase — the model is trained on training folds using a specific algorithm (e.g., Decision Tree).
2. Testing phase — the trained model is applied to the validation fold.
3. Performance evaluation — metrics such as accuracy, precision, and recall are calculated using the Performance (Classification) operator.

Figure 4. is an example of model evaluation logic inside a Cross Validation block using a Decision Tree. The model is trained, applied to test data, and evaluated for classification accuracy. This modular design ensures that all models are evaluated under the same conditions, which makes their comparison fair and methodologically correct.

Figure 3. Basic set up

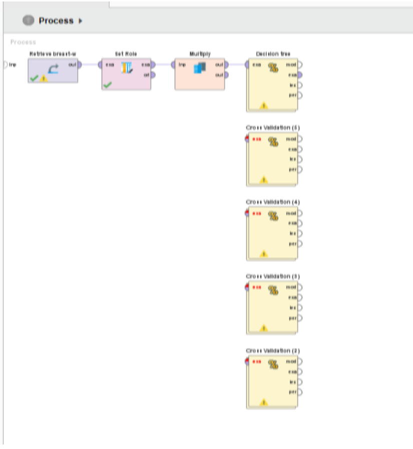
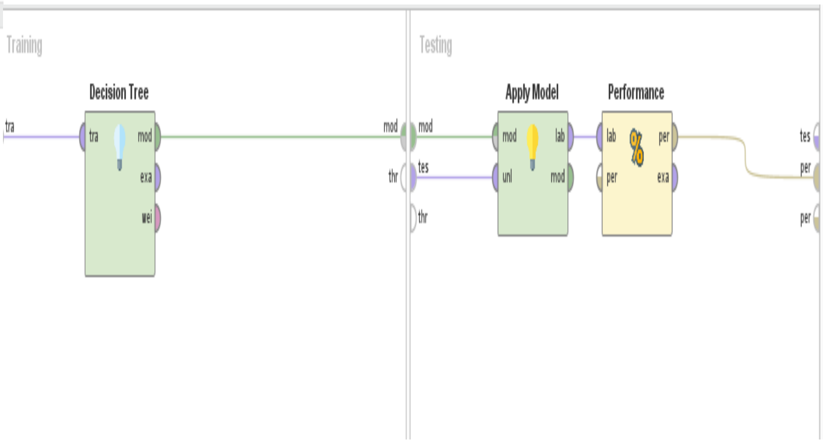


Figure 4. Inside logic



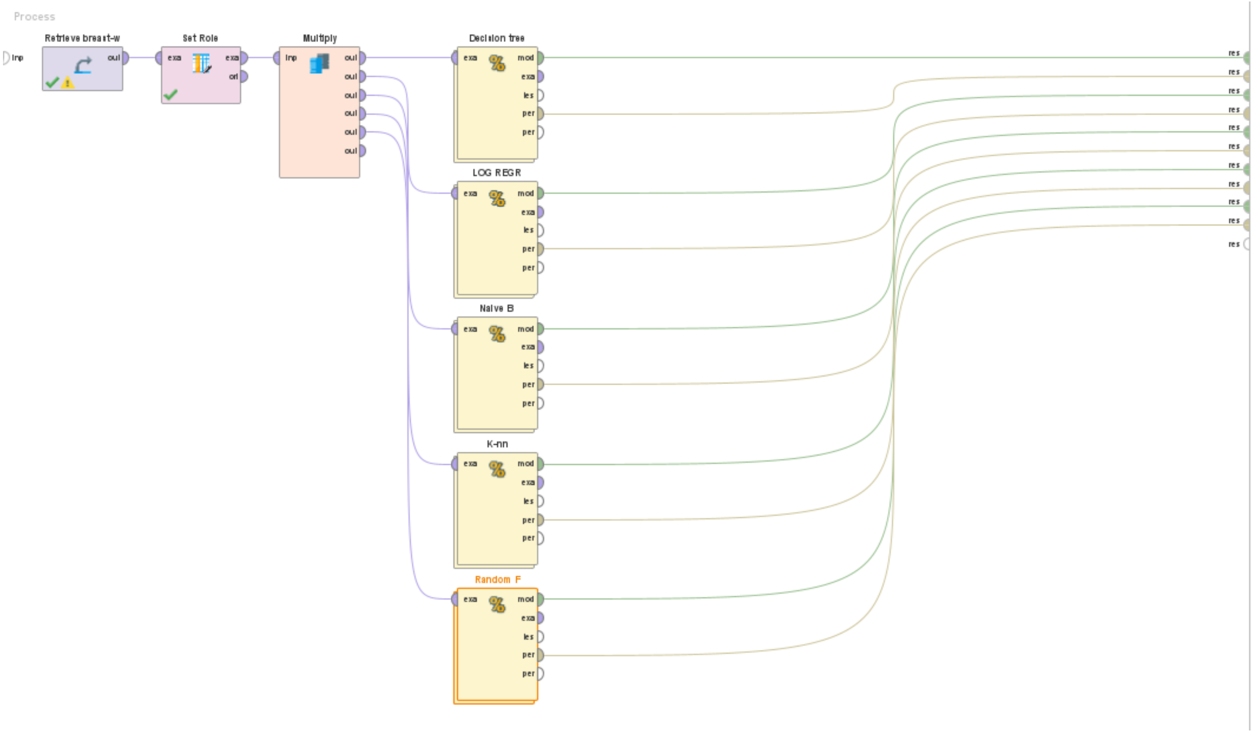
3.3. Final Experimental Design

After defining the data roles and setting up Cross Validation blocks for each model, all five classification algorithms were incorporated into a single workflow, as shown in Figure 4. The Multiply operator was used to create five identical copies of the dataset. Each copy was passed into a separate model block using one of the following algorithms:

- Decision Tree
- Logistic Regression
- Naive Bayes
- k-Nearest Neighbors (k-NN)
- Random Forest

Each model block includes a cross-validation loop with internal training, testing, and performance evaluation, ensuring consistency across all experiments. The output ports of each model are connected to the final result view, allowing for easy side-by-side comparison of their evaluation metrics. This design enables systematic and fair comparison of different models, making it easy to analyze which algorithm performs best on the breast cancer dataset.

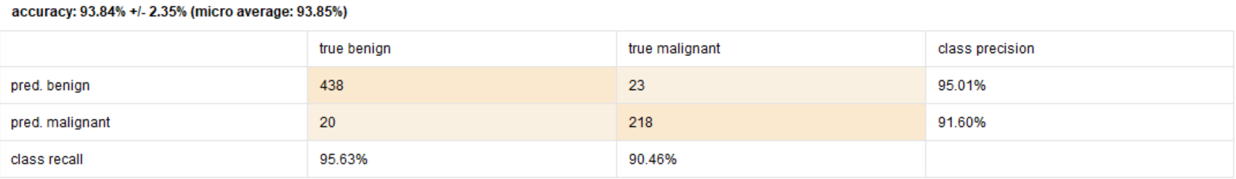
Figure 5. Final experimental workflow in RapidMiner



3.4. Decision Tree— Results and Interpretation

The decision tree model demonstrated a high overall accuracy of 93.84% with a deviation of ±2.35%, indicating stable performance under cross-validation. The classification quality is balanced: precision for benign reached 95.01% and for malignant 91.60%, while recall values were 95.63% and 90.46%, respectively, showing the model's solid ability to detect both classes. The number of misclassifications was relatively low—23 malignant cases were incorrectly labeled as benign and 20 benign cases as malignant, confirming a reasonable trade-off between sensitivity and specificity. These results make the model well suited for binary classification tasks with limited data and no preprocessing. Due to its transparency and the possibility of visualizing decision paths, the tree can also be useful in medical data analysis, where interpretability is important.

Figure 6. Decision tree matrix



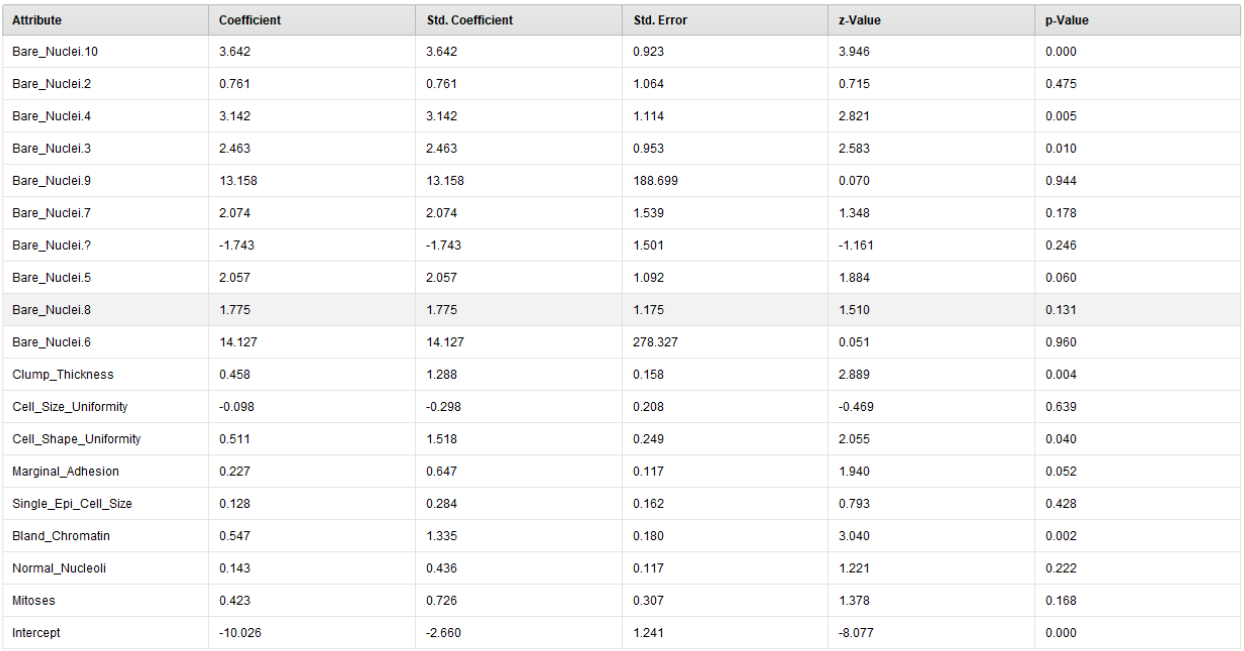
accuracy: 93.84% +/- 2.35% (micro average: 93.85%)

	true benign	true malignant	class precision
pred. benign	438	23	95.01%
pred. malignant	20	218	91.60%
class recall	95.63%	90.46%	

3.5 Logistic Regression – Results and Interpretation

The logistic regression coefficient table enables assessment of the importance and contribution of individual features, making the model not only predictive but also analytical. The most statistically significant predictors ($p < 0.05$) were Bare_Nuclei.10, .4, .3, Clump_Thickness, Bland_Chromatin, and the Intercept, as indicated by high z-values and low p-values. These variables contribute most substantially to the model and influence the likelihood of a case being classified as malignant. Features with high coefficients and significant p-values can be considered key diagnostic factors within the model's interpretation. Therefore, logistic regression offers not only effective classification but also transparency, allowing its use in identifying the most influential predictors in the dataset.

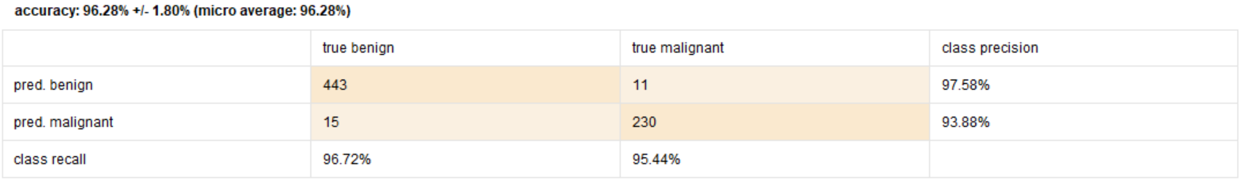
Figure 7. Logistic regression outputs



Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Bare_Nuclei.10	3.642	3.642	0.923	3.946	0.000
Bare_Nuclei.2	0.761	0.761	1.064	0.715	0.475
Bare_Nuclei.4	3.142	3.142	1.114	2.821	0.005
Bare_Nuclei.3	2.463	2.463	0.953	2.583	0.010
Bare_Nuclei.9	13.158	13.158	188.699	0.070	0.944
Bare_Nuclei.7	2.074	2.074	1.539	1.348	0.178
Bare_Nuclei.?	-1.743	-1.743	1.501	-1.161	0.246
Bare_Nuclei.5	2.057	2.057	1.092	1.884	0.060
Bare_Nuclei.8	1.775	1.775	1.175	1.510	0.131
Bare_Nuclei.6	14.127	14.127	278.327	0.051	0.960
Clump_Thickness	0.458	1.288	0.158	2.889	0.004
Cell_Size_Uniformity	-0.098	-0.298	0.208	-0.469	0.639
Cell_Shape_Uniformity	0.511	1.518	0.249	2.055	0.040
Marginal_Adhesion	0.227	0.647	0.117	1.940	0.052
Single_Epi_Cell_Size	0.128	0.284	0.162	0.793	0.428
Bland_Chromatin	0.547	1.335	0.180	3.040	0.002
Normal_Nucleoli	0.143	0.436	0.117	1.221	0.222
Mitoses	0.423	0.726	0.307	1.378	0.168
Intercept	-10.026	-2.660	1.241	-8.077	0.000

The logistic regression model delivered excellent results with an accuracy of 96.28% and a deviation of ±1.80%, indicating high stability under cross-validation. The classification quality was well balanced: precision was 97.58% for benign and 93.88% for malignant, while recall reached 96.72% and 95.44%, respectively, confirming the model's strong ability to detect both classes while minimizing errors. The number of misclassifications was very low, with only 11 false positives and 15 false negatives, reflecting the model's reliability. This combination of precision, sensitivity, and consistency makes logistic regression fully suitable for binary classification tasks. Given these properties, the model can also be applied in medical contexts, especially where it is essential to clearly distinguish safe cases from potentially dangerous ones.

Figure 8. Logistic regression matrix



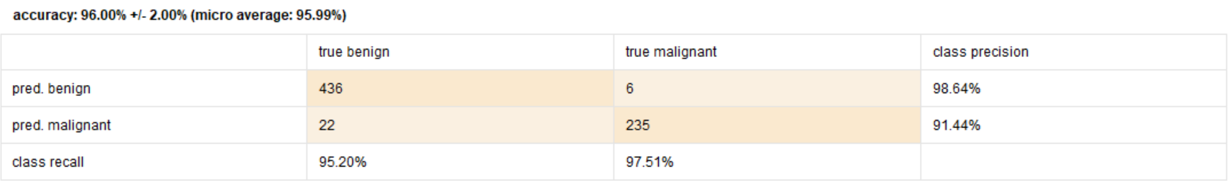
accuracy: 96.28% +/- 1.80% (micro average: 96.28%)

	true benign	true malignant	class precision
pred. benign	443	11	97.58%
pred. malignant	15	230	93.88%
class recall	96.72%	95.44%	

3.6. Naive Bayes – Results and Interpretation

The Naive Bayes model demonstrated a high overall accuracy of 96.00% with a deviation of ±2.00%, indicating stable results under repeated validation. Precision reached 98.64% for benign and 91.44% for malignant, while recall was 95.20% and 97.51%, respectively, showing that the model effectively detects both classes and performs especially well in identifying malignant cases. The number of classification errors was relatively low: 6 malignant instances were incorrectly predicted as benign and 22 benign cases as malignant, indicating an acceptable level of false classifications and moderate risk of missing dangerous cases. These metrics confirm that Naive Bayes, despite its simplicity and the assumption of feature independence, handles the classification task reliably and can be applied in resource-limited scenarios. Due to its speed, robustness, and acceptable accuracy, it can also be considered in applied medical contexts where sensitivity to malignant outcomes is critical.

Figure 9. NB matrix



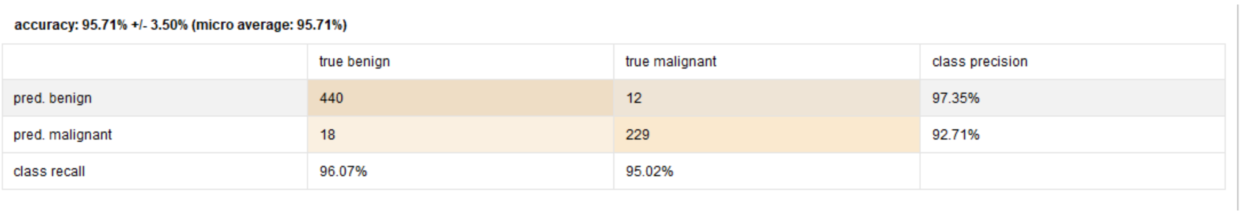
accuracy: 96.00% +/- 2.00% (micro average: 95.99%)

	true benign	true malignant	class precision
pred. benign	436	6	98.64%
pred. malignant	22	235	91.44%
class recall	95.20%	97.51%	

3.7. k-Nearest Neighbors (k-NN) – Results and Interpretation

The k-nearest neighbors model achieved an overall accuracy of 95.71% with a deviation of ±3.50%, reflecting solid classification performance, though with slightly higher variability between validation folds compared to other models. Precision reached 97.35% for benign and 92.71% for malignant, while recall values were 96.07% and 95.02%, indicating the model's strong ability to correctly identify both benign and malignant cases. Misclassifications were moderate: 12 malignant instances were incorrectly predicted as benign and 18 benign cases as malignant, which remains within an acceptable margin. Given that the k-NN algorithm is sensitive to feature scaling and sample size, such results are considered quite successful for unprocessed data. The model may be valuable in scenarios where interpretability is not a priority and raw classification accuracy is more important, including certain practical medical contexts.

Figure 10. K-NN matrix



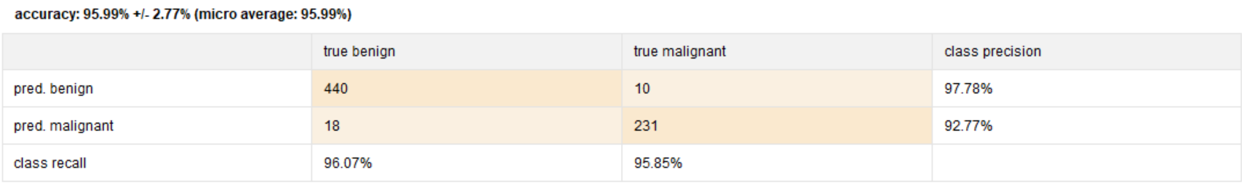
accuracy: 95.71% +/- 3.50% (micro average: 95.71%)

	true benign	true malignant	class precision
pred. benign	440	12	97.35%
pred. malignant	18	229	92.71%
class recall	96.07%	95.02%	

3.8. Random Forest – Results and Interpretation

The Random Forest model demonstrated very high overall accuracy of 95.99% with a deviation of ±2.77%, indicating consistent performance across cross-validation folds. Precision reached 97.78% for benign and 92.77% for malignant, while recall was 96.07% and 95.85%, suggesting strong performance in identifying both classes, with a slight tendency toward benign detection. Classification errors were minimal, with 10 malignant cases incorrectly predicted as benign and 18 benign cases as malignant, confirming a stable balance between sensitivity and specificity. Given that Random Forest relies on an ensemble approach and reduces the risk of overfitting, it proved particularly effective on the raw dataset. This model is well-suited for tasks where high accuracy must be maintained along with reliability, including applied medical contexts where minimizing false negatives is essential without compromising overall prediction quality.

Figure 11. Random Forest matrix



accuracy: 95.99% +/- 2.77% (micro average: 95.99%)

	true benign	true malignant	class precision
pred. benign	440	10	97.78%
pred. malignant	18	231	92.77%
class recall	96.07%	95.85%	

3.9. Comparison of Models

Brief analysis:

Highest accuracy: Logistic Regression (96.28%) with the lowest deviation.

Best sensitivity to malignant cases (recall): Naive Bayes (97.51%), which is important for reducing the risk of missing malignant cases.

Fewest total errors: Logistic Regression (only 26 errors: 11 false positives + 15 false negatives).

Lowest number of false negatives (malignant → benign): Random Forest and k-NN — only 10–12 errors.

Most stable result (lowest standard deviation): again, Logistic Regression.

Table 1. Model comparison

Model	Accuracy	Precision (Benign / Malignant)	Recall (Benign / Malignant)	FP	FN
Decision Tree	93.84% ± 2.35%	95.01% / 91.60%	95.63% / 90.46%	23	20
Log. Regression	96.28% ± 1.80%	97.58% / 93.88%	96.72% / 95.44%	11	15
Naive Bayes	96.00% ± 2.00%	98.64% / 91.44%	95.20% / 97.51%	6	22
k-NN	95.71% ± 3.50%	97.35% / 92.71%	96.07% / 95.02%	12	18
Random Forest	95.99% ± 2.77%	97.78% / 92.77%	96.07% / 95.85%	10	18

The comparative analysis of five classification models showed that logistic regression delivered the best overall performance, achieving the highest accuracy (96.28%) with the lowest deviation (±1.80%) and the fewest classification errors, along with balanced precision and recall for both classes. Closely following were Random Forest and Naive Bayes, which both achieved nearly equivalent accuracy (~96%), with Naive Bayes showing the highest sensitivity to malignant cases and Random Forest offering the best combination of robustness and precision without a notable increase in errors. The k-NN model also performed reliably, though with higher variability, while the Decision Tree, despite its interpretability, was the least stable and accurate among the five. Given these characteristics, logistic regression can be recommended as a baseline solution for binary classification tasks without preprocessing, particularly when both reliability and consistency are required. Random Forest is well-suited for scenarios where minimizing false negatives is critical, especially in clinical contexts where missing a malignant case can have serious consequences. Naive Bayes may be appropriate when sensitivity to malignant outcomes is prioritized, even if its assumptions of feature independence may not fully hold. Overall, all models demonstrated practical applicability, but the choice of algorithm should consider not only accuracy but also the specific medical use case, balancing reliability, interpretability, and patient safety.

4. BigML: Model Implementation and Evaluation

After completing the experimental setup and evaluation in RapidMiner, the analysis was extended using a second platform: BigML. This allowed for a comparison of not only model performance but also user experience between a desktop-based and a cloud-based machine learning environment.

BigML is an online, interactive machine learning tool that provides a highly visual and accessible interface. Its cloud-based architecture makes it convenient for conducting quick experiments without requiring installation or programming skills. In contrast to RapidMiner's workflow-based logic, BigML emphasizes simplicity and automation, making it particularly useful for testing standard algorithms in educational and exploratory contexts.

For consistency, the same classification task was carried out: predicting whether a breast tumor is benign or malignant based on various cell features. The following models were tested:

- Logistic Regression, a commonly used probabilistic classifier, especially suitable for medical data due to its transparency.
- Decision Tree, which offers an interpretable set of decision rules derived from recursive feature splitting.
- Ensemble Model (Random Forest), which enhances accuracy and robustness by combining predictions from multiple decision trees.

All models were trained on 70% of the dataset and evaluated on the remaining 30% using BigML's standard train-test split. The evaluation included key metrics such as accuracy, precision, recall, F1-score, and a breakdown of error types, with a specific focus on minimizing false negatives, which are especially critical in the context of medical diagnostics.

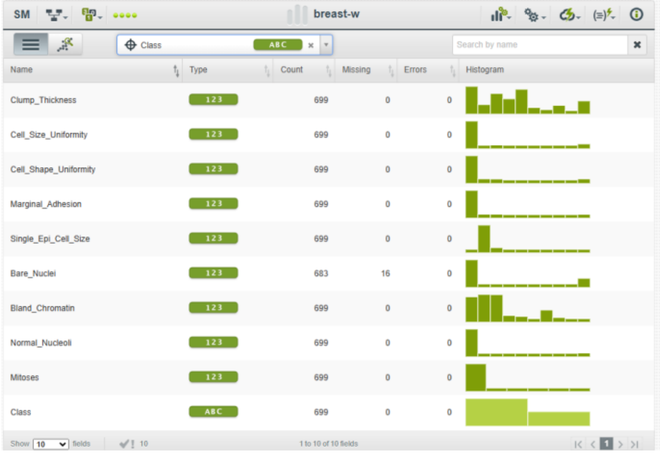
4.1. Data Upload and Preparation

The dataset used in this study is based on a well-known medical dataset related to breast cancer diagnosis. The data was uploaded into BigML in .csv format under the name breast-w.csv. It includes 699 patient records with 10 attributes describing characteristics of cell nuclei obtained via fine-needle aspiration.

Attributes include: Clump_Thickness, Cell_Size_Uniformity, Cell_Shape_Uniformity, Bare_Nuclei, Bland_Chromatin, Normal_Nucleoli, and others.

BigML automatically identified the data types: nine of the attributes were detected as numeric, and one field, Class, was recognized as categorical. This Class variable contains the diagnosis outcome for each case, labeled as either benign or malignant.

Figure 12. Dataset in Big ML



Name	Type	Count	Missing	Errors	Histogram
Clump_Thickness	123	699	0	0	
Cell_Size_Uniformity	123	699	0	0	
Cell_Shape_Uniformity	123	699	0	0	
Marginal_Adhesion	123	699	0	0	
Single_Epi_Cell_Size	123	699	0	0	
Bare_Nuclei	123	683	16	0	
Bland_Chromatin	123	699	0	0	
Normal_Nucleoli	123	699	0	0	
Mitoses	123	699	0	0	
Class	ABC	699	0	0	

Show 10 fields ✓! 10 1 to 10 of 10 fields

4.2. Target Variable and Data Splitting

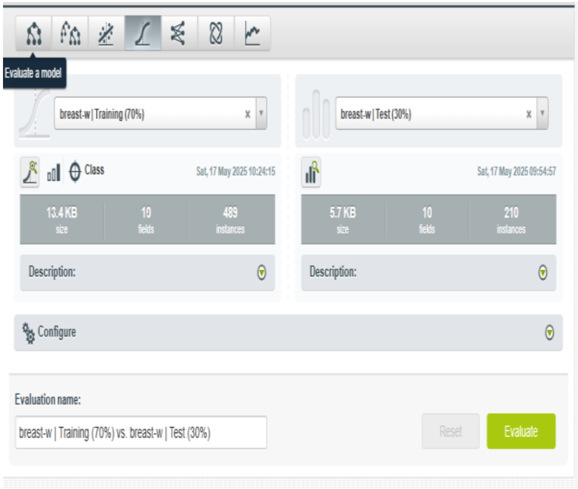
After uploading, BigML automatically selected the Class variable as the Objective Field, since it was the only categorical attribute and thus a suitable target for classification. This field represents the medical diagnosis and is the variable we aim to predict using the models.

To properly evaluate the models' generalization performance, the dataset was randomly split into two parts using BigML's built-in Split Dataset function:

- 70% of the data (489 records) was used to train the models,
- and 30% (210 records) was held out for evaluation and testing purposes.

Random splitting ensures that both classes are proportionally represented in the training and testing sets. This step is crucial to avoid biased evaluation and to simulate the model's behavior on unseen data.

Figure 13. Train and test split



breast-w | Training (70%)

breast-w | Test (30%)

Class — Sat, 17 May 2025 10:24:15

13.4 KB size	10 fields	489 instances

Sat, 17 May 2025 09:54:57

5.7 KB size	10 fields	210 instances

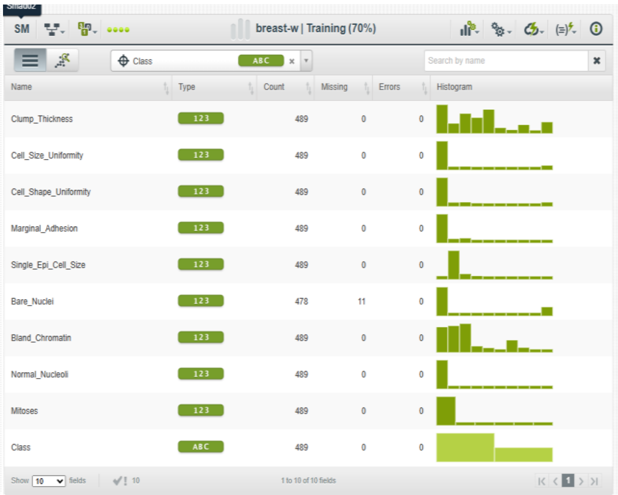
Description:

Description:

Evaluation name:

breast-w | Training (70%) vs. breast-w | Test (30%)

Figure 14. Training dataset



breast-w | Training (70%)

Class ABC

Name	Type	Count	Missing	Errors	Histogram
Clump_Thickness	123	489	0	0	
Cell_Size_Uniformity	123	489	0	0	
Cell_Shape_Uniformity	123	489	0	0	
Marginal_Adhesion	123	489	0	0	
Single_Epi_Cell_Size	123	489	0	0	
Bare_Nuclei	123	478	11	0	
Bland_Chromatin	123	489	0	0	
Normal_Nucleoli	123	489	0	0	
Mitoses	123	489	0	0	
Class	ABC	489	0	0	

Show 10 fields ✓! 10 1 to 10 of 10 fields 1

4.3. Logistic Regression Model

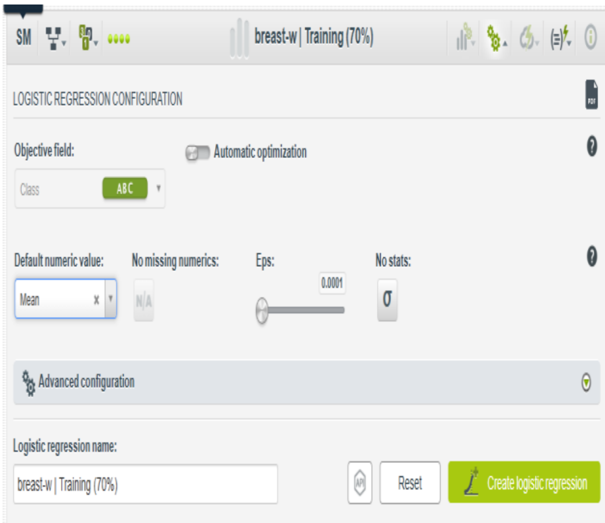
The first model developed was a Logistic Regression, trained using the 70% training dataset. Logistic regression is a statistical model that estimates the probability of a binary outcome using a logistic (sigmoid) function applied to a linear combination of input features.

In BigML, the logistic regression model was generated automatically after selecting the training dataset and confirming the target field (Class). Missing values if had been missing, could have been po handled by assigning the mean of each feature as the default replacement value. The platform provided coefficients for each feature, showing how they contribute to increasing or decreasing the predicted probability of malignancy.

For instance, attributes such as Bare_Nuclei, Clump_Thickness, and Normal_Nucleoli had strong positive coefficients, indicating that higher values of these features increase the likelihood of the tumor being malignant. Conversely, features with negative coefficients slightly decreased that likelihood.

The resulting model provides both probabilistic outputs (e.g., "probability of malignant: 0.84") and binary class predictions, making it suitable for threshold-based decision-making in clinical contexts.

Figure 15. Creating Logistic regression



breast-w | Training (70%)

LOGISTIC REGRESSION CONFIGURATION

Objective field: Class ABC

○ Automatic optimization

Default numeric value: Mean

No missing numerics: N/A

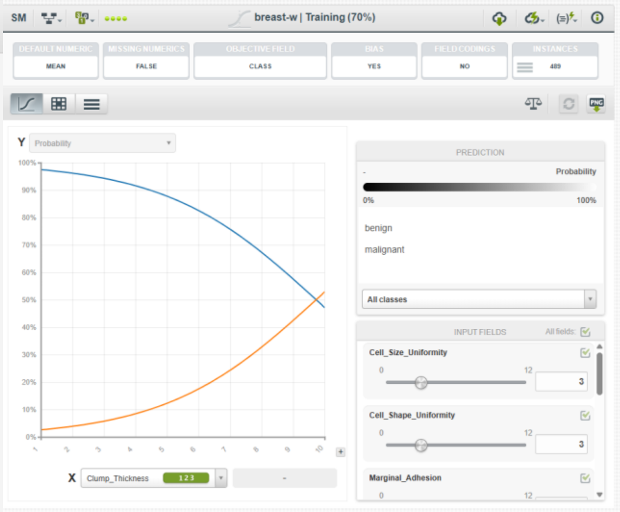
Eps: 0.0001

No stats: σ

Advanced configuration

Logistic regression name: breast-w | Training (70%)

Figure 16. Logistic regreression statistics



breast-w | Training (70%)

DEFAULT NUMERIC	MISSING NUMERICS	OBJECTIVE FIELD	BIAS	FIELD CODINGS	INSTANCES
MEAN	FALSE	CLASS	YES	NO	489

Y Probability

X Clump_Thickness 123

PREDICTION

Probability

0% 100%

benign

malignant

All classes

INPUT FIELDS All fields: ☑

Cell_Size_Uniformity ☑ 0 12 3

Cell_Shape_Uniformity ☑ 0 12 3

Marginal_Adhesion ☑ 0 12

Once the logistic regression model was trained, it was evaluated on the test dataset (30%) using BigML’s built-in Evaluate tool. This step simulates real-world application by testing the model’s performance on previously unseen data.

The results were highly satisfactory:

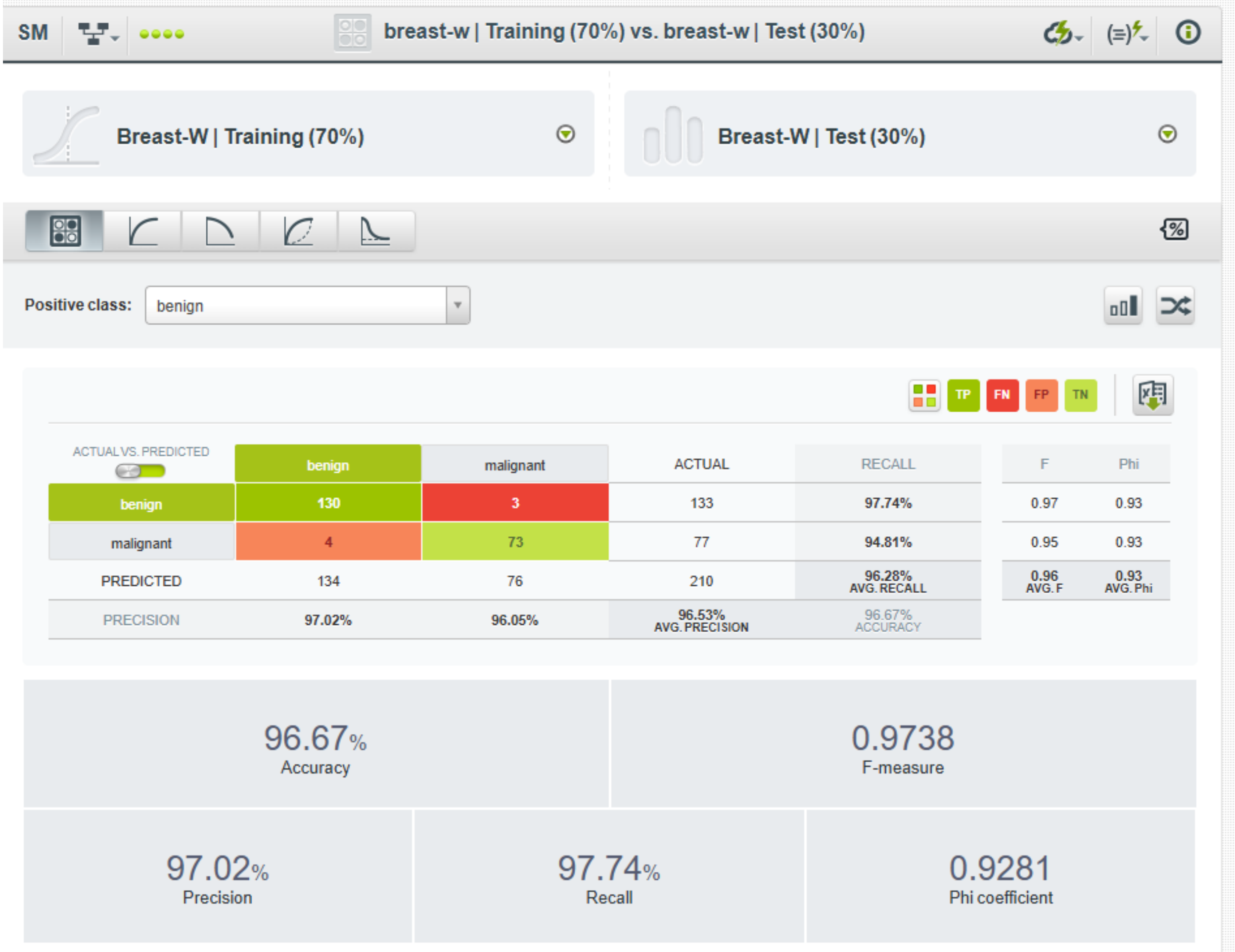
- Accuracy: 96.67%
- Recall (Sensitivity): 97.74% for benign cases, 94.81% for malignant cases
- Precision: 97.02%
- F1-Score: 0.9738

In terms of classification errors:

- 4 false negatives occurred (malignant tumors incorrectly predicted as benign),
- and 3 false positives (benign tumors incorrectly predicted as malignant).

These results demonstrate a strong balance between sensitivity and specificity, with low error rates and high predictive power. From a medical standpoint, the low number of false negatives is particularly important, as missing a malignant case could have severe consequences. Thus, logistic regression proved to be a robust and clinically relevant model for this task.

Figure 17. Logistic Regression Evaluation



breast-w | Training (70%) vs. breast-w | Test (30%)

Breast-W | Training (70%)

Breast-W | Test (30%)

Positive class: benign

ACTUAL VS. PREDICTED	benign	malignant	ACTUAL	RECALL	F	Phi
benign	130	3	133	97.74%	0.97	0.93
malignant	4	73	77	94.81%	0.95	0.93
PREDICTED	134	76	210	96.28% AVG. RECALL	0.96 AVG. F	0.93 AVG. Phi
PRECISION	97.02%	96.05%	96.53% AVG. PRECISION	96.67% ACCURACY		

96.67% Accuracy

0.9738 F-measure

97.02% Precision

97.74% Recall

0.9281 Phi coefficient

4.4. Decision Tree model

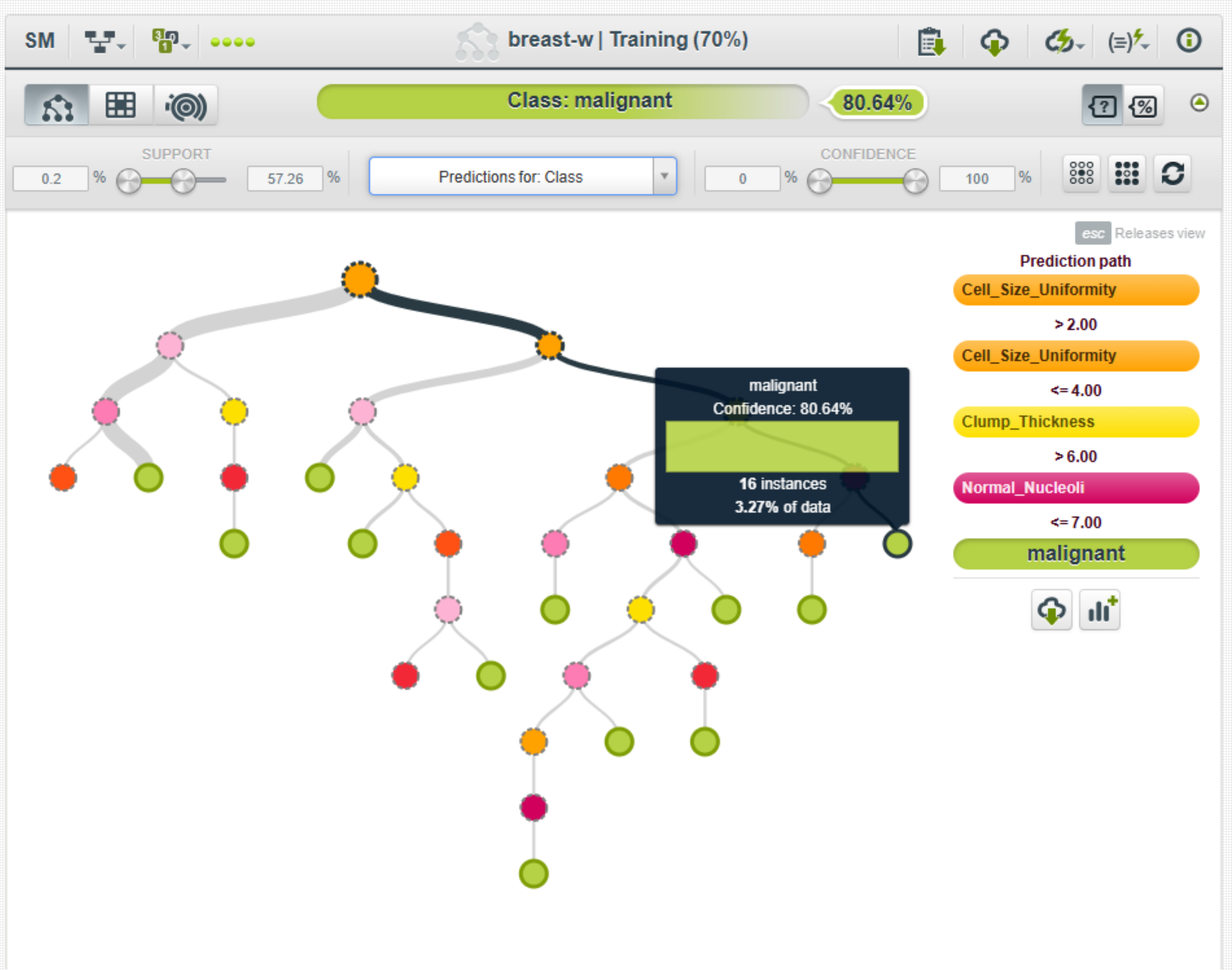
The Decision Tree was built using the 70% training subset. The model splits the dataset recursively based on conditions applied to the most predictive attributes. Each internal node represents a decision rule based on one attribute, while each terminal node (leaf) provides a class prediction and the model's confidence.

For example (Figure 18.), one decision path for predicting a malignant tumor involves the following conditions:

- Cell_Size_Uniformity between 2 and 4,
- Clump_Thickness greater than 6,
- and Normal_Nucleoli less than or equal to 7.

Cases that follow this path represent 3.27% of the training data, and are classified as malignant with 80.64% confidence. Such rule-based structures make Decision Trees highly interpretable, which is valuable in medical settings where transparency of predictions is essential.

Figure 18. Decision tree path



breast-w | Training (70%)

Class: malignant 80.64%

SUPPORT 0.2 % 57.26 %

Predictions for: Class

CONFIDENCE 0 % 100 %

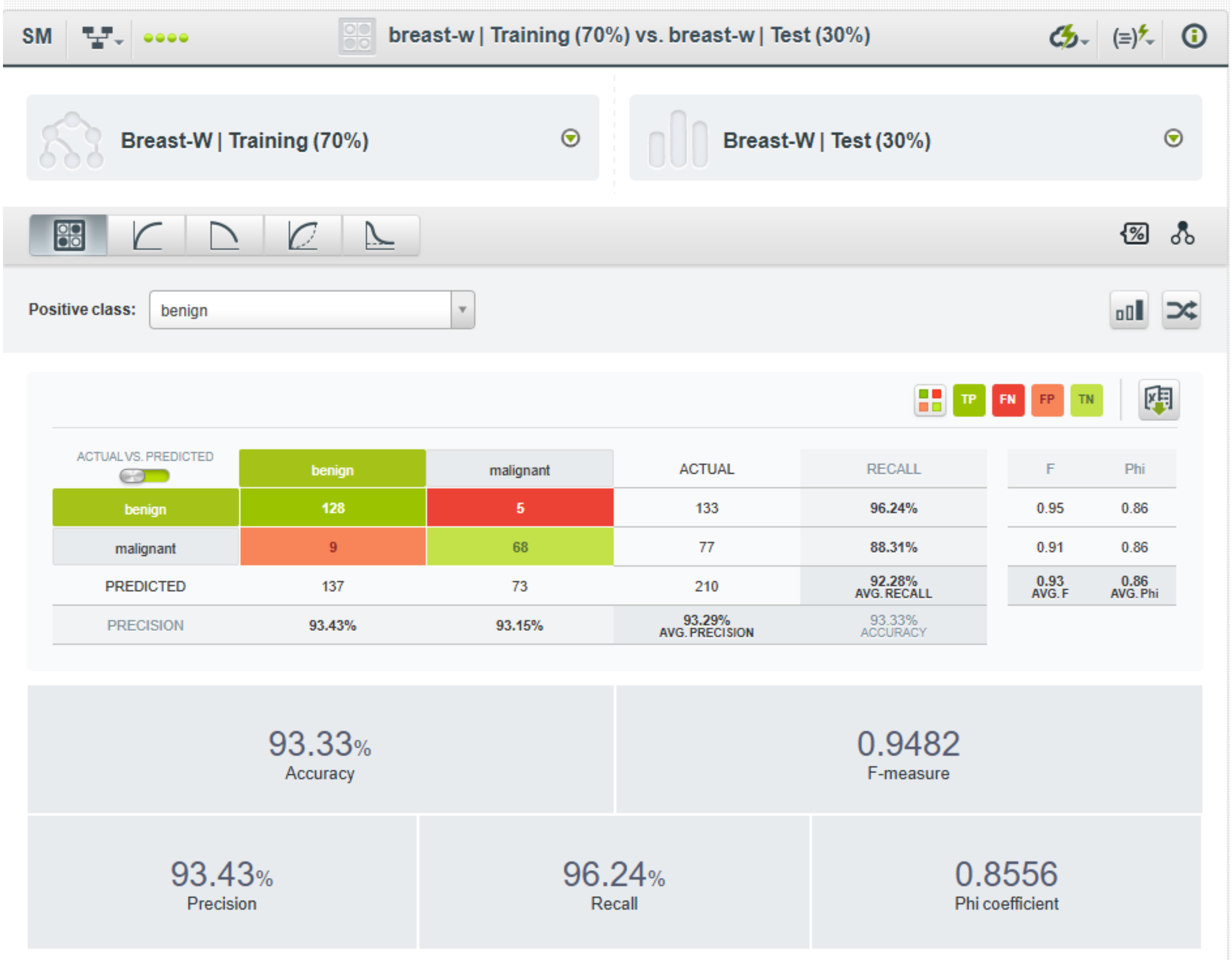
malignant
Confidence: 80.64%
16 instances
3.27% of data

Prediction path
Cell_Size_Uniformity
> 2.00
Cell_Size_Uniformity
<= 4.00
Clump_Thickness
> 6.00
Normal_Nucleoli
<= 7.00
malignant

The decision tree model was evaluated on the test set (30% of the dataset), yielding an overall accuracy of 93.33%. The recall for benign cases reached 96.24%, while recall for malignant cases was slightly lower at 88.31%, indicating that the model was somewhat more likely to miss malignant diagnoses. The model made 9 false negative errors (malignant tumors predicted as benign), which is a critical concern in medical contexts. Precision remained solid at 93.43%, and the F-measure was 0.9482, reflecting a good balance between precision and recall.

Although the performance was slightly lower than that of the logistic regression model, the decision tree offers higher interpretability. Clinicians can visually examine the decision path and understand which features led to a specific diagnosis, making this model particularly useful in scenarios where model transparency is essential.

Figure 19. Decision tree Evaluation



breast-w | Training (70%) vs. breast-w | Test (30%)

Breast-W | Training (70%)

Breast-W | Test (30%)

Positive class: benign

ACTUAL VS. PREDICTED	benign	malignant	ACTUAL	RECALL	F	Phi
benign	128	5	133	96.24%	0.95	0.86
malignant	9	68	77	88.31%	0.91	0.86
PREDICTED	137	73	210	92.28% AVG. RECALL	0.93 AVG. F	0.86 AVG. Phi
PRECISION	93.43%	93.15%	93.29% AVG. PRECISION	93.33% ACCURACY		

93.33% Accuracy

0.9482 F-measure

93.43% Precision

96.24% Recall

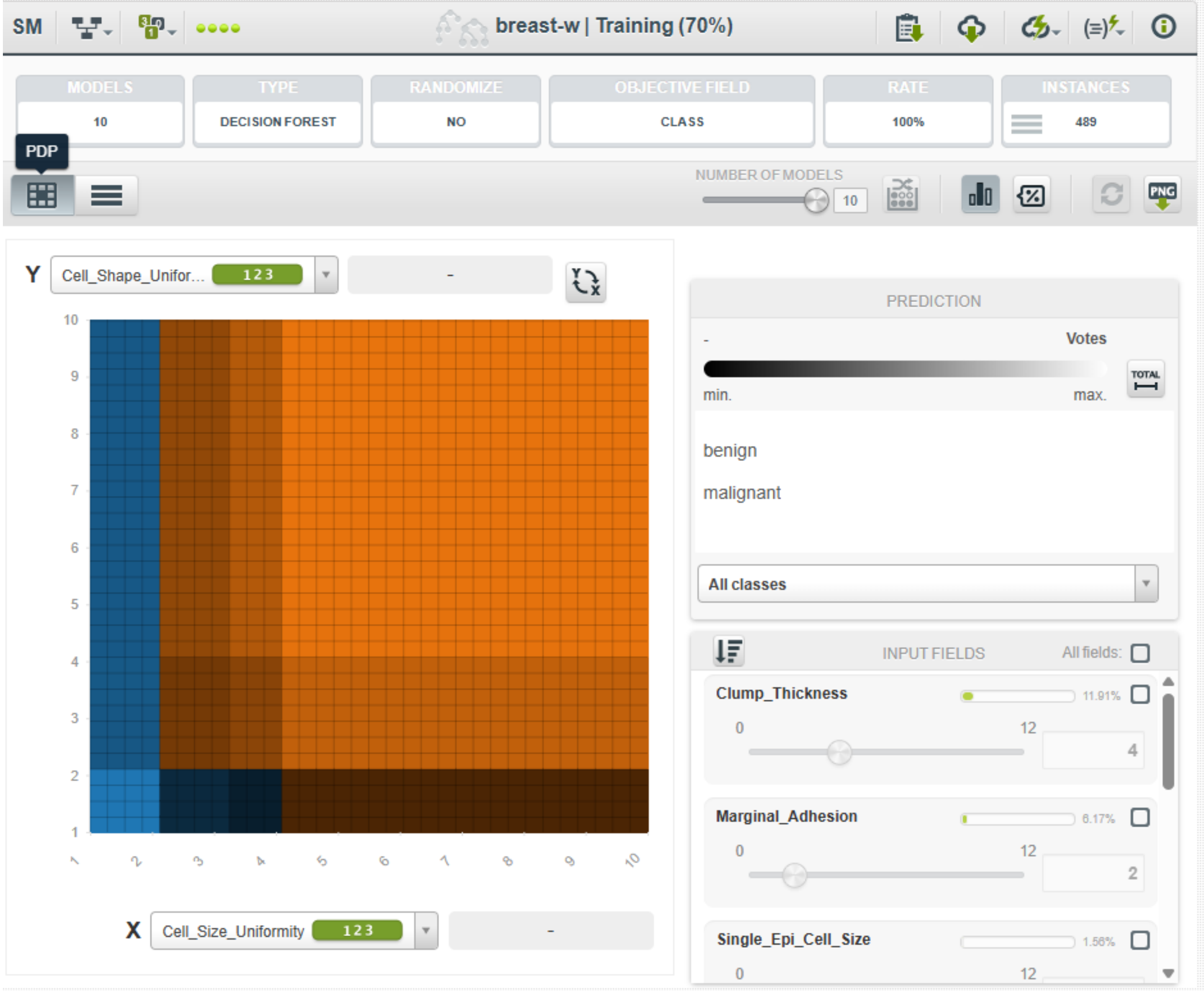
0.8556 Phi coefficient

4.5. Ensemble model

The final model trained was an Ensemble Model based on the Random Decision Forest algorithm. This technique builds multiple decision trees using random subsets of the training data and features, and aggregates their outputs to produce a more stable and accurate prediction. In this project, an ensemble of 10 decision trees was constructed using BigML's default parameters. The model was trained on the 70% training split, with Class as the objective field.

Ensemble models are particularly valuable in reducing overfitting and variance that may arise from using a single decision tree. Although interpretability is somewhat reduced compared to a standalone tree, the gain in predictive performance often justifies their use, especially in high-stakes fields like medical diagnostics.

Figure 20. Ensemble model



breast-w | Training (70%)

MODELS	TYPE	RANDOMIZE	OBJECTIVE FIELD	RATE	INSTANCES
10	DECISION FOREST	NO	CLASS	100%	489

NUMBER OF MODELS 10

Y Cell_Shape_Unifor... 123 -

X Cell_Size_Uniformity 123 -

PREDICTION

- Votes

min. max.

benign

malignant

All classes

INPUT FIELDS All fields: □

Clump_Thickness 11.91% □ 0 12 4

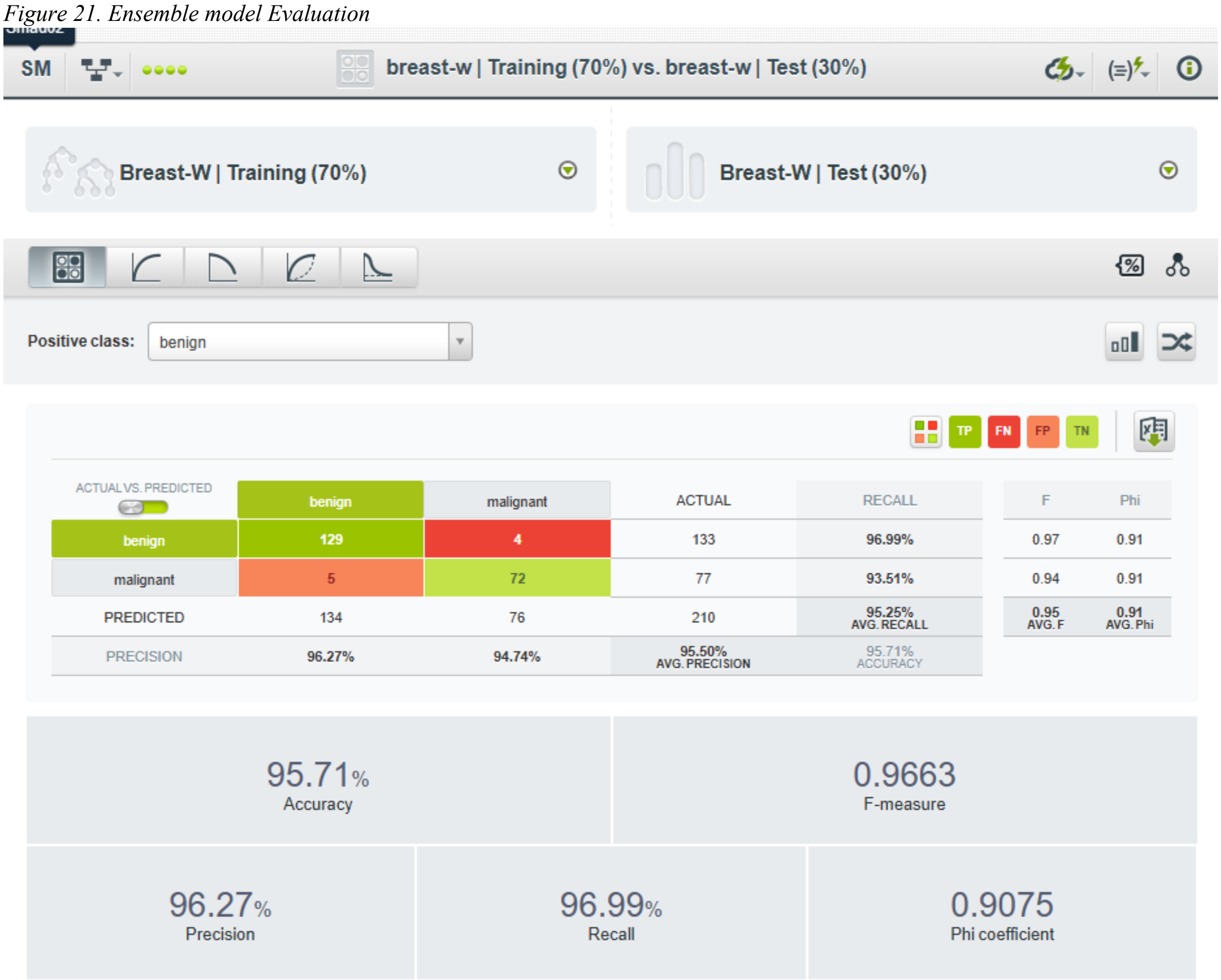
Marginal_Adhesion 6.17% □ 0 12 2

Single_Epi_Cell_Size 1.56% □ 0 12

The ensemble model, based on a random decision forest with 10 trees, was evaluated on the 30% test dataset. It achieved the highest overall performance among all models tested, with an accuracy of 95.71%, precision of 96.27%, and an outstanding recall of 96.99% for benign cases. Malignant case recall was also strong at 93.51%, resulting in only 5 false negatives and 4 false positives.

The ensemble method outperformed the individual decision tree by reducing variance and increasing generalization. Although it is less interpretable than a single tree, the boost in accuracy and reduced error rates make it a compelling choice, especially in high-stakes domains like healthcare where diagnostic precision is essential. Its balanced performance across all metrics demonstrates robustness and reliability.

Figure 21. Ensemble model Evaluation


SM
breast-w | Training (70%) vs. breast-w | Test (30%)
Breast-W | Training (70%)
Breast-W | Test (30%)
TP
FN
FP
TN

Positive class: benign

ACTUAL VS. PREDICTED	benign	malignant	ACTUAL	RECALL	F	Phi
benign	129	4	133	96.99%	0.97	0.91
malignant	5	72	77	93.51%	0.94	0.91
PREDICTED	134	76	210	95.25% AVG. RECALL	0.95 AVG. F	0.91 AVG. Phi
PRECISION	96.27%	94.74%	95.50% AVG. PRECISION	95.71% ACCURACY		

95.71% Accuracy

0.9663 F-measure

96.27% Precision

96.99% Recall

0.9075 Phi coefficient

4.6. BigML Model Comparison and Interpretation

The three classification models implemented in BigML—Logistic Regression, Decision Tree, and Ensemble (Random Forest)—each demonstrated high accuracy, but with distinct trade-offs:

Table 2. BigML Comparison

Model	Accuracy	Precision	Recall (Malignant)	F-measure	Phi Coefficient	False Positives (FP)	False Negatives (FN)
Logistic Regression	96.67%	97.02%	94.81%	0.9738	0.9281	3	4
Decision Tree	93.33%	93.43%	88.31%	0.9482	0.8556	5	9
Ensemble (Random Forest)	95.71%	96.27%	93.51%	0.9663	0.9075	4	5

As shown in Table 2, all three models achieved high predictive performance, but with varying strengths:

- Logistic Regression outperformed the others across nearly all metrics. It achieved the highest accuracy (96.67%), strong precision (97.02%), and excellent recall (94.81%) for malignant cases. It also had the highest F-measure (0.9738) and Phi coefficient (0.9281), indicating both consistency and strong correlation between predictions and true labels.

With only 3 false positives and 4 false negatives, it represents the most balanced and reliable model in this comparison.

- Decision Tree provided the most interpretable model but had the lowest recall (88.31%) for malignant cases and the highest number of false negatives (9). Despite its solid precision (93.43%) and a decent F-measure (0.9482), the risk of misclassifying malignant tumors limits its practical application in clinical scenarios. However, its simplicity and transparency may still justify its use where interpretability is critical.
- Ensemble Model (Random Forest) offered a strong middle ground. With accuracy of 95.71%, high precision (96.27%), and recall (93.51%), it demonstrated robust performance while maintaining relatively low error rates (4 FP, 5 FN). It also achieved a high F-measure (0.9663) and Phi coefficient (0.9075), confirming its consistency and suitability for medical diagnostic tasks where both accuracy and reliability are required.

Overall, while all models proved usable, Logistic Regression provided the best balance of precision, recall, and low error rates. Random Forest followed closely, and Decision Tree remains valuable for its explainability.

5. User Experience: RapidMiner vs. BigML

From a usability perspective, RapidMiner and BigML offer distinct user experiences:

- RapidMiner provides a highly flexible visual workflow builder with many customizable options. It's well-suited for advanced users or scenarios requiring control over each step (e.g., cross-validation, branching, parameter tuning). However, it may feel overwhelming to beginners due to the large number of operators and interface complexity.
- BigML, in contrast, offers a streamlined and guided experience, automatically identifying target variables and performing train-test splits. It is ideal for quick experimentation and educational purposes. Though it lacks the deep customization of RapidMiner, it excels in clarity and speed.

In short:

- RapidMiner = best for detailed workflow control and custom setups
- BigML = best for simplicity, speed, and accessibility

Conclusion

This project compared five machine learning algorithms across two platforms—RapidMiner and BigML—on a breast cancer classification task. Logistic Regression consistently emerged as the most balanced model in both environments, delivering high accuracy and minimal error rates. Ensemble methods (Random Forest) also demonstrated excellent performance, particularly in reducing false negatives.

Decision Trees were slightly less accurate but offered the highest transparency, which is valuable in medical decision-making. BigML proved to be faster and more intuitive for beginners, while RapidMiner enabled deeper customization and reproducibility through structured workflows.

Overall, the choice of tool and algorithm should align with the specific goals of the analysis—whether interpretability, precision, or ease of use. For medical applications, minimizing false negatives while maintaining model reliability remains the top priority.

References

Altair. (n.d.). Altair RapidMiner. Retrieved from https://altair.com/altair-rapidminer

BigML. (2016, September 23). BigML Logistic Regressions: The 6 Steps to Predictions. Retrieved from https://blog.bigml.com/2016/09/23/bigml-logistic-regressions-the-6-steps-to-predictions/

BigML. (n.d.). What is an ensemble? Retrieved from https://support.bigml.com/hc/en-us/articles/208204385-What-is-an-ensemble

RapidMiner. (n.d.). Cross Validation. RapidMiner Documentation. Retrieved from https://docs.rapidminer.com/9.9/studio/operators/validation/cross_validation.html

BigML. (n.d.). BigML Evaluation Measures. Retrieved from https://bigml.com/evaluate

BigML. (n.d.). Random Decision Forests in BigML. Retrieved from https://bigml.com/gallery/models/random-decision-forest

BigML. (n.d.). BigML Decision Tree Visualizations. Retrieved from https://bigml.com/gallery/models/decision-tree

BigML. (n.d.). BigML Dashboard Overview. Retrieved from https://bigml.com/dashboard