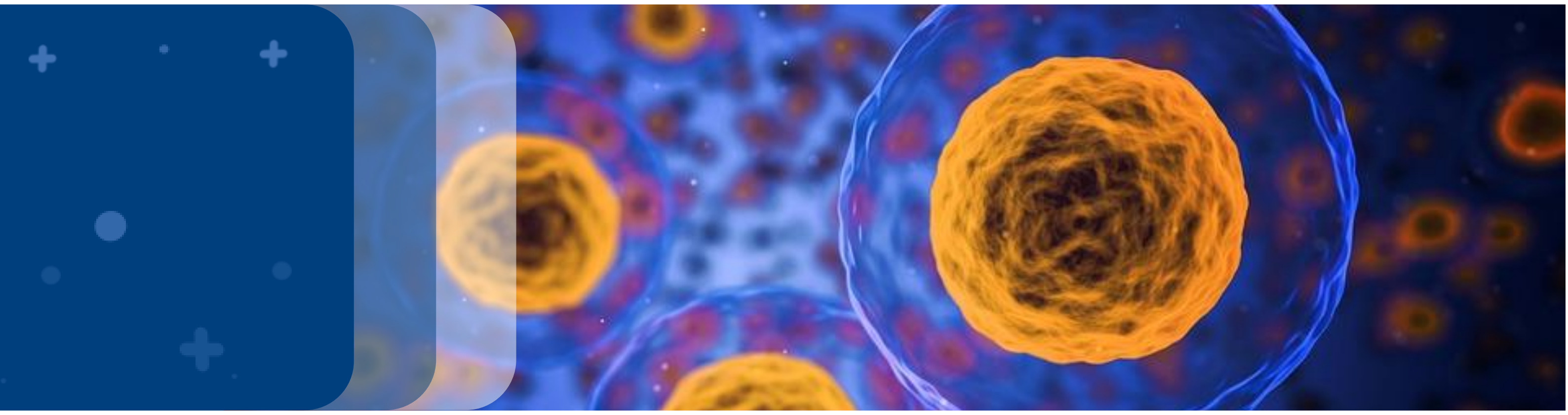
supper&supper



Breast Tumor Quality Prediction With The Help Of dataiku

Project Goal

Breast cancer is the second most common type of cancer and the first among women worldwide with more than 2 million new cases diagnosed in 2018. Early diagnosis improves the survival drastically. The correct clustering of tumors into benign and malignant protects patients from taking unnecessary drugs and decreases potential harm.

The project goal was to develop a machine learning model to predict breast tumor quality.

Provided Data

Two different data sets from a Wisconsin hospital with 570 and 700 cases, respectively, were used to address the problem. The first data set was based exclusively on the cellular level. For each image ten cell features with the corresponding mean, standard error and "worst" (mean of the three largest values) were computed. The second data set provided discrete values from one to ten of cell attributes and mitosis stage.

Challenges

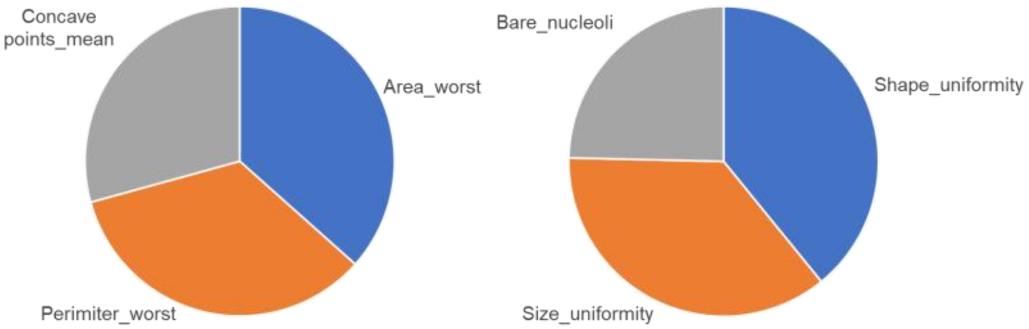
Both data sets contained heterogenous distribution of benign and malignant groups.

Applied Methods

Taking into consideration, that the output groups were known, we implemented supervised learning algorithms like random forest and logistic regression. Accuracy was chosen as a metric to obtain a comparison between algorithms performances. These steps were carried out with Dataiku, the platform democratizing access to data and enabling enterprises to build their own path to AI.

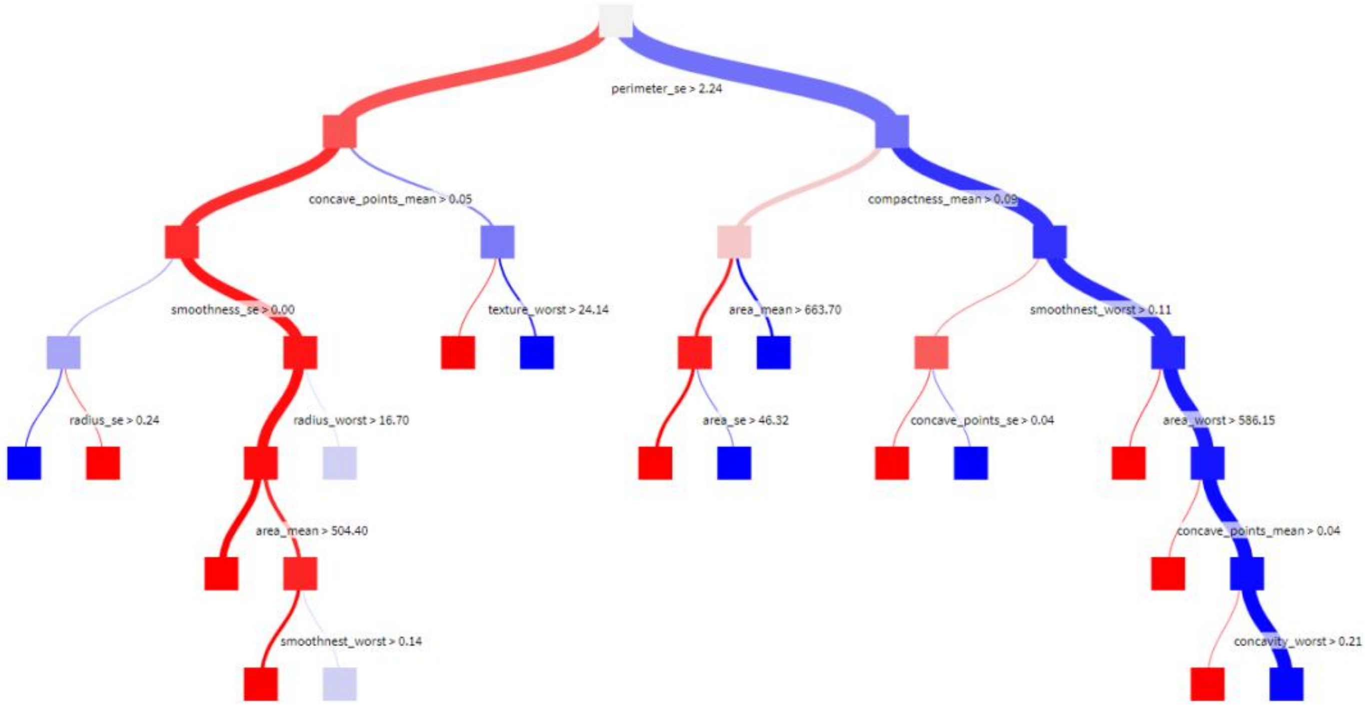
Project Outcome

Random forest approach performed with an accuracy of 99,7 % and of 99,3% respectively.



Variables importance of the two data sets according to the random forest approach

Using random forest method we obtained the importance of used variables of both data sets. For the first data set the means of the three largest values were critical for making predictions. It was notable that mitosis stage of the second date set played no role for the prediction.



Decision tree of the first data set. Red corresponds to the benign tumors and blue to the malignant ones. Each parent node includes probability of child nodes being benign or malignant.

Logistic regression approach performed with an accuracy of 99.6% in both cases. The confusion matrices based on the optimized F1-score were stored. The false prediction of benign tumor instead of malignant was penalized higher.

The threshold of the confusion matrices corresponded to the number beyond which the prediction was considered positive, the values were set to 0.475 and 0.25.

	Predicted M	Predicted B	Total
Actually M	37	2	39
Actually B	1	81	82
Total	38	83	121

	Predicted M	Predicted B	Total
Actually M	46	0	46
Actually B	4	89	93
Total	50	89	139

Further applications

The results of the predictions can be used in medicine to automate the estimation of breast tumors with high accuracy.

Since the possible therapy depends on the risk of the recurrence event, the algorithm can address the problem as well.