Interpretability in the Frequentists and Bayesian Frameworks

Final Capstone Project Report

by

Marta Matosas Fonolleda

Sandra Valdés Salas

Index

1. Glossary of Terms and Acronyms

Accuracy: concerns the ability of a model to make correct predictions (Molnar, C. (2020).

BART: acronym for Bayesian Additive Regression Trees.

CART: acronym for Classification And Regression Tree.

Global interpretability refers to the “understanding of how the model makes decisions, based on a holistic view of its features and each of the learned components” (Molnar, C. (2020)).

HDI: acronym for Highest Density Interval.

Interpretability: concerns to what degree the model allows for human understanding (Molnar, C. (2020)).

Local interpretability: examines “what and why the model makes a certain prediction for a single instance.” (Molnar, C. (2020)).

MCMC: acronym for Markov Chain Monte Carlo method.

SHAP: acronym for Shapley Additive exPlanations.

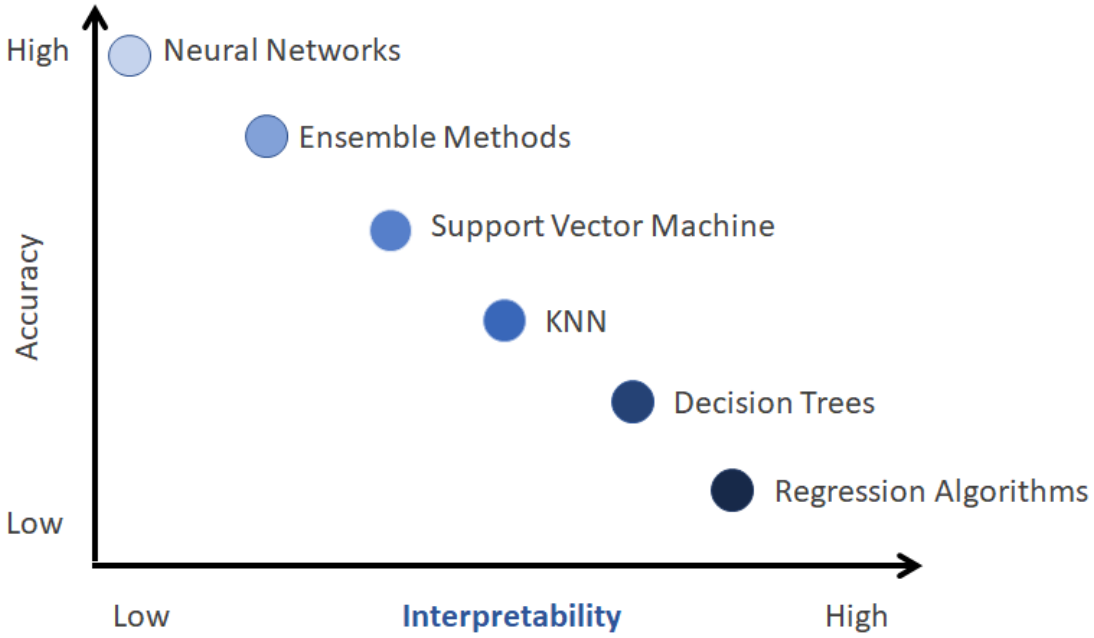
2. Introduction

2.1. Background

Interpretability is important because the higher the interpretability of a model, the easier it is to understand why decisions or predictions were made (Molnar, 2020). Model interpretability addresses issues related to trust, informativeness, fairness and ethical decision making (Lipton, 2017). For these reasons, interpretability is particularly relevant for high risk domains.

In this project we are concerned with the accuracy and the interpretability of predictive models. Accuracy concerns the ability of a model to make correct predictions, while interpretability concerns to what degree the model allows for human understanding.

Figure 1. The trade-off between accuracy and interpretability



More complex and opaque models many times exhibit high accuracy, while more interpretable models may lack the necessary accuracy.

Consequently, new methods have been explored to address this tension in both the frequentist and the Bayesian fields.

On the frequentist approach, Lundberg and Lee (2017) introduced the SHapley Additive exPlanations (SHAP) framework for interpreting predictions by assigning each feature an importance value for a particular prediction.

On the Bayesian framework, on the other hand, a method for interpretability was recently introduced by Afrabandpey, et. al. (2020). The authors proposed a general principle for interpretability based, first, on building a highly accurate and, second, using an interpretable proxy model to explain the predictive model. It is worth mentioning that the use of simple models as interpretable surrogates to highly predictive black-box models has been a common practice in Machine Learning. However, this approach has poorly been explored in the Bayesian framework (Afrabandpey, et. al., 2020).

Even though there is no consensus on the definition of interpretability and its scope, for the purpose of this project, the concepts of global and local interpretability as defined by Molnar, C. (2020) apply:

- global interpretability refers to the “understanding of how the model makes decisions, based on a holistic view of its features and each of the learned components”.

- local interpretability examines “what and why the model makes a certain prediction for a single instance.”

Finally, the performance metrics will be excluded from the local interpretability results since they are not differential to any model since they can always be obtained.

2.2. Problem Statement

Interpretability is most relevant in high-impact domains. Consequences derived from poor interpretability are lack of advancement of knowledge in the field, credibility and error identification.

Thus, it is key to optimize the trade-off between accuracy and interpretability.

This project explores the interpretability of predictive models that are generally classified as high accuracy models in both the Frequentist and the Bayesian frameworks.

2.3. Problem Elaboration

To explore the interpretability of a model generally classified as high accuracy, an ensemble method will be compared to a regression model.

In the frequentist framework, an ensemble model will be further explained with SHAP.

In the Bayesian framework, the ensemble model trained is BART, acronym for the Bayesian Additive Regression Trees.

The high impact domain chosen is healthcare; and the dataset used is the Breast Cancer Wisconsin**.** The target is the diagnosis of the tumor: benign or malignant.

The project is partially based on previous work where the dataset was preprocessed, 5 out of thirty features were selected and six models of the frequentist framework were optimized by hyperparameter tuning.

2.4. Motivation

The motivation for this project is that breast cancer is the most frequently diagnosed cancer in women, and the leading cause of death cancer in women worldwide. An estimated over half a million women died due to breast cancer, only in 2016. If breast cancer is detected early, there are more treatment options and a better chance for survival.

An increased interpretability of a prediction model may build its credibility and advance knowledge in the field further increasing the accuracy of diagnosis and the chance for survival.

2.5. Project Scope

The following activities will be executed within the scope of the project. They are broken down into two sections, one per each of the frameworks.

Frequentist framework:

- Identify the best combination of models to ensemble.
- Implement ensembling methods: voting and stacking.
- Apply SHAP to the most accurate ensembling model.
- Interpret the results globally (model) and locally (predictions).

Python and Sklearn and SHAP libraries will be used to train and explain the models.

Bayesian framework:

- Implement Bayesian Logistic Regression and Bayesian Additive Regression Trees (BART).
- Interpret the results globally (model) and locally (predictions).

The JAGS function and the MCMC and BART package in R will be used to train and explain the models.

3. Literature Review

Relevant research for this project was focused on the following elements: interpretability, SHAP and BART.

Molnar, C. (2020). Interpretable Machine Learning. Retrieved from https://christophm.github.io/interpretable-ml-book/

Lundberg, S. M. and Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. Retrieved from https://arxiv.org/abs/1705.07874

Doshi Velez, F. (2017) Towards a rigorous science of interpretable machine learning. Retrieved from https://arxiv.org/pdf/1702.08608.pdf

Afrabandpey, et. al. (2020). A Decision-Theoretic Approach for Model Interpretability in Bayesian Framework. Retrieved from https://arxiv.org/abs/1910.09358

Lipton, Z. C. (2017). The Mythos of Interpretability. Retrieved from https://arxiv.org/pdf/1606.03490.pdf

Chipman, H. A. et al. (2010). BART: Bayesian additive regression trees. Retrieved from https://projecteuclid.org/euclid.aoas/1273584455

4. Frequentist Framework Methodology

Logistic Regression

The Logistic Regression model to be interpreted is explained by equation 1.

Equation 1

$$x = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + \beta 4x4 + \beta 5x5$$

Ensemble Method and SHAP

The ensemble method to be explained was created by stacking the Logistic Regression and Random Forest models that were optimized and trained in previous work (see Appendix for general information about the modelling of previous work).

The criteria applied to choose the models to be stacked is the following:

1- choose top performing model as a benchmark: Logistic Regression
2- identify models that have learned differently:

   a) inspect missed predictions

   b) calculate intersection with respect to benchmark

   c) select the two models with the smallest intersection.

Next, SHAP was applied to it to further explain it. The SHAP explanation method implemented is the KernelSHAP: this method is model agnostic; there is not a SHAP method specific to ensemble methods.

SHAP was chosen because today is the state-of-the-art method for interpretability in machine learning. SHAP was proposed as a unified framework to interpret model predictions at the 31st Conference on Neural Information Processing Systems (2017), a top conference in the machine learning community. This new class unifies six additive feature explanation methods, including: LIME, DeepLift, Shapley regression values, among others. This is notable because several recent methods in the class lack the proposed desirable properties found in SHAP.

5. Frequentist Framework Analysis and Results

Logistic Regression Global Interpretability

The results of the logistic regression model are the beta coefficients:

Table 1. Beta coefficients of the Bayesian Logistic Regression model

Feature	***Beta coefficient***
texture_mean	0.30351628
area_worst	0.0278593
smoothness_worst	3.81442389
area_mean	-0.02193761
concavity_mean	8.29138305

Following equations 2 and 3 that are derived from the model, changes in the coefficients can be traced to changes in the odds ratio and the probability, respectively.

Equation 2

$$\frac{P(y=1)}{1-P(y=1)} = odds = exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right)$$

Equation 3

$$P(y^{(i)}=1) = \frac{1}{1+exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

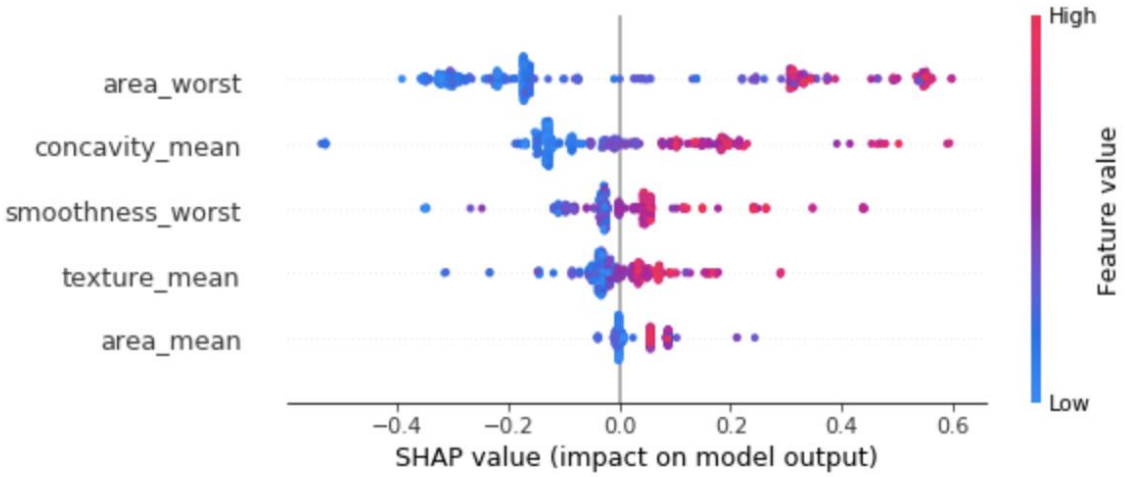
Logistic Regression Local Interpretability

The beta coefficients calculated are point estimates that can provide the following information at the instance level:

1. the probability of the prediction of a single test instance
2. how a change in a feature by one unit changes the odds ratio by a factor of exponential beta
3. how a change in a feature by one unit changes the probability.

Ensemble Method and SHAP Global Interpretability

Figure 2. Summary plot of SHAP values on outcome "malignant"



The elements of Interpretation of the summary plot of SHAP values are three:

1- Feature importance: variables are ranked in descending order of importance. This provides an intuition of their relative importance.
2- Impact: the horizontal location shows whether the effect of that value is associated with a higher or lower prediction.
3- Original value: color shows whether that variable is high (in pink) or low (in blue) for that observation.

Thus, the takeaways of SHAP values of outcome “malignant” (figure 2) are:

1. area_worst is the most important and area_mean the least important among these 5 features.
2. A high value of “area_worst” has a high and positive impact on the outcome “malignant”. The “high” impact comes from the pink color, and the “positive” impact is shown on the X-axis. There is no direct link between a variation of the value of the feature and the probability of the diagnosis “malignant”.
3. Low value of area worst (in blue) can also lead to a positive impact on the outcome “malignant”.
4. In the other features we can also find a few low values having a positive impact on the outcome ‘malignant”. It seems there is a stronger correlation between high values and positive impact on outcome than in area worst.

Ensemble Method and SHAP Local Interpretability

Each observation gets a raw prediction and its own set of SHAP values by class.

The results of a False Negative instance are the following:

- Raw prediction: {benign: 0.94874832, malignant: 0.05125168}
- Individual SHAP values, outcome benign: {texture_mean: -0.0166343, area_worst: 0.09247941, smoothness_worst: 0.09247941, area_mean: 0.02068463, concavity_mean: 0.094231}

This instance is of great interest since it refers to a “malignant” tumor miss classified as “benign”.

The raw prediction shows the probability assigned to each label: the highest probability is assigned to the label “benign”, explaining why it has been misclassified.

The individual SHAP values provide an intuition of the contributions of the predictors to a prediction. They interpret the impact of having a certain value for a given feature in comparison to the prediction we would make if that feature took the expected value, a baseline value.

The individual SHAP values of this particular instance are represented in figure 3:

Figure 3. Individual SHAP values of a False Negative

The length of the bars represents the magnitude of the SHAP value, the contribution to the outcome, and the color indicates the sign: positive in pink and negative in blue.

The takeaways are the following:

1. The base value is 0.64 whereas the predicted value is 0.95.
2. The biggest impact of this sample being misclassified as benign comes from smoothness worst, it is the largest bar in pink.
3. The only feature decreasing it is texture mean, in blue.
4. Finally, note that the SHAP values of all features sum up to the difference between the base value and the predicted value, explaining the prediction of the instance.

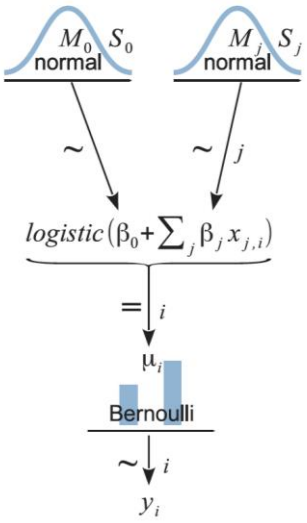
6. Bayesian Framework Methodology

The method used for producing approximations to the Bayesian posterior distributions was Markov Chain Monte Carlo (MCMC). All the features, except for the target, were standardized to reduce autocorrelation in the MCMC chains and to improve the efficiency of the algorithm.

Bayesian Logistic Regression

The first model proposed to explore interpretability in the Bayesian framework is a Bayesian Logistic Regression model, which represents a highly interpretable model. Figure 4 summarizes the Bayesian logistic regression model by displaying a dependency diagram of the model.

Figure 4. Dependency diagram of the Logistic Regression model



This model aims at predicting the outcome *y*, which follows a Bernoulli distribution where *y*=1 is a malignant tumor and *y*=0 is a benign tumor. In this model, a linear combination of five numeric predictors (β_j) is mapped to a probability value via the logistic function. As there is not additional information about the prior distribution of the parameters, in this model, it is assumed that the intercept (β_0) and parameters (β_j**)** follow a prior normal distribution.

The software used for the implementation of this model was JAGS in R. The code of the model was obtained from Kruschke (2015).

Bayesian Additive Regression Trees (BART)

The second model proposed to explore interpretability in the Bayesian framework is Bayesian Additive Regression Trees (BART), which represents a highly accurate but opaque model. BART is a Bayesian "sum-of-trees" model; fitting and inference are accomplished via an iterative Bayesian backfitting MCMC algorithm.

Logit BART for dichotomous outcomes was implemented (equation 4).

Equation 4

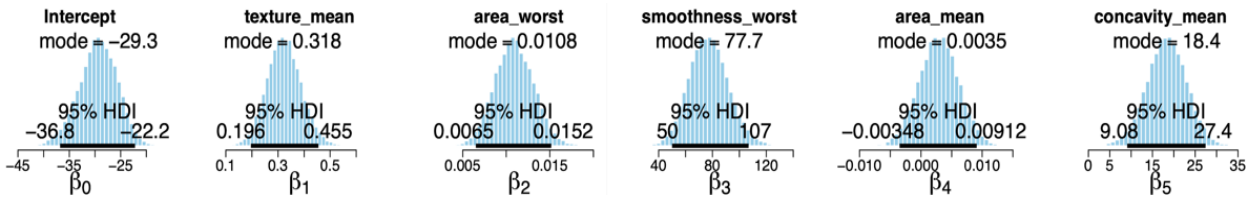
$$Y = f(x) + e = E(Y|x) \quad \Rightarrow \quad P(Y=1|x) = F(f(x))$$

Where F denotes the logit link and $f(x)$ is the sum of *m* regression trees, in this implementation m=1,000.

7. Bayesian Framework Analysis and Results

Bayesian Logistic Regression Global Interpretability

Figure 5. Posterior distributions of the Bayesian Logistic Regression model



The results of the Bayesian Logistic Regression model are the posterior distributions of the coefficients (figure 5).

The mode values indicate the most probable value of the beta coefficients; note that all the mode values of the beta coefficients of the features are positive.

The HDI indicates which points of a distribution are most credible for each beta coefficient. The HDI of the feature “area_mean”is distributed around zero, meaning that the most credible value of “area_mean” might be given by chance and, therefore, concluding that this parameter is not statistically significant in the model.

The width of the HDI provides a relative measure of the degree of certainty of the predictive power of the parameter. The more narrow the HDI width, the more certain are the values in the HDI. This means, for example, that we are more confident of beta 2, parameter of area_worst, than beta 1, parameter of texture_mean.

This does not indicate feature importance, rather a relative degree of confidence in their predictive power.

Bayesian Logistic Regression Local Interpretability

The mode values of the beta coefficients (figure 5) enable the following calculations providing interpretability at the instance level:

1. the outcome for a specific sample using the most credible values of the beta coefficients (the mode values).
2. how a change in a feature by one unit changes the odds ratio by a factor of exp(beta), following equation 2.
3. how a change in a feature by one unit changes the probability, following equation 3.

BART Global Interpretability

BART provides the number of times each feature is chosen in a tree decision rule by each regression tree. The mean was calculated to provide an intuition of how features are used in the decision rules used across all trees (table 2).

Table 2. Mean count of times a feature is chosen in a decision rule

Feature	Mean count
area_worst	27.83
concavity_mean	23.29
texture_mean	21.23
smoothness_worst	20.16
area_mean	18.65

Contrary to Bayesian Logistic Regression, there is no off-the-shelf method to obtain Bart's posterior distributions of the beta coefficients.

BART Local Interpretability

BART outputs a matrix of probabilities of outcome malignant with number of rows equal to number of posterior draws and each column corresponding to a test sample.

See figure 6 for a subset of the first six features (columns) and 5 of the 1,000 regression trees (rows).

Figure 6. Subset of probabilities matrix

```
          [,1]        [,2]      [,3]         [,4]        [,5]         [,6]
[1,] 0.9987685 0.093989987 0.8957154 0.0046249450 0.073034358 0.0607259220
[2,] 0.9981552 0.050956615 0.9234520 0.0017679130 0.151630538 0.0535965026
[3,] 0.9989266 0.061282635 0.9673051 0.0026289646 0.173552564 0.0685120872
[4,] 0.9963945 0.074026386 0.9591167 0.0016901353 0.167630423 0.0259843965
[5,] 0.9981236 0.066364822 0.9078386 0.0017100410 0.209568556 0.0156350212
```

Contrary to Bayesian Logistic Regression, there is no off-the-shelf method to obtain Bart's posterior distributions of the beta coefficients.

8. Conclusion

8.1. Conclusions

Frequentist Framework

In the frequentist framework, Shapley Additive Explanations (SHAP) proved to be a powerful method to increase interpretability of opaque models, in this case, ensemble methods.

The main advantage of SHAP is that it provides an intuition about feature importance. For instance, in this project, area_worst seems to be the most relevant feature whereas area_mean is the least. However, the SHAP values are not meaningful because they do not provide traceability to the probability.

Bayesian Framework

The main advantage of the Bayesian framework is that it provides posterior distributions that give the following three pieces of information.

First, the statistical significance of the features: according to the Bayesian Logistic Regression model implemented in this project, area_mean seems to be not statistically significant because its 95% High Density Function (HDI) is distributed around zero, suggesting that the value of area_mean might be given by chance. The rest of the features are statistically significant.

Second, a ranking of confidence in the predictive power of the features: according to the Bayesian Logistic Regression model, area_worst has the narrowest 95% HDI, thus it is the first feature in the ranking of confidence. A narrow HDI indicates certainty about the values of the beta coefficients.

Finally, meaningful beta coefficients that provide traceability between the change in the probability and the change in the feature values.

Nonetheless, it is worth mentioning that this information is not available via an off-the-shelf method to interpret opaque models in the Bayesian framework, as is the case with the Bayesian Additive Regression Trees (BART).

8.2. Project Limitations

The following project limitations have been identified:

1. KernelSHAP is slow and, therefore, impractical to use when you want to compute Shapley values for many instances. This impacts the computation of SHAP feature importance since it entails computing Shapley values for a lot of instances.

2. KernelSHAP ignores feature dependence as do most other permutation based interpretation methods. The values from random instances that replace feature values are sampled from the marginal distribution. This leads to putting too much weight on unlikely data points if features are dependent.

3. MCMC can be computationally expensive.

4. The choice of the prior distribution in the Bayesian data modelling is subjective.

5. A single Bayesian ensemble method was implemented and explored.

8.3. Future Research

Building on the last project limitation identified, two lines of future work have been defined.

1. Implement additional Bayesian methods generally classified as high accuracy.

2. Implement and explore tools and methods to increase the interpretability of high accuracy Bayesian models.

Specifically, the models and tools to be explored are:

1. CART (Classification And Regression Tree).
   1.1. The advantages provided for our project are the following:

- it provides most model interpretability because they are simply series of if-else conditions,
- it can handle both numerical and categorical data, and
- nonlinear relationships among features do not affect the performance of the decision trees.

1.2. The disadvantages to be managed are:

- a small change in the dataset can make the tree structure unstable which can cause variance, and
- decision tree learners create underfit trees if some classes are imbalanced. It is therefore recommended to balance the data set prior to fitting with the decision tree.

2. BART-Machine: a new package in R implementing Bayesian additive regression trees (BART). The package introduces many new features for data analysis using BART such as variable selection, interaction detection, model diagnostic plots, incorporation of missing data and the ability to save trees for future prediction. Additionally, this package is capable of handling both large sample sizes and high-dimensional data. Thus, it is significantly faster than the current implementation of BART.

9. References

http://www.r-5.org/files/books/computers/algo-list/statistics/data-mining/John_K_Kruschke-Doing_Bayesian_Data_Analysis-EN.pdf

https://www.researchgate.net/figure/Global-burden-of-breast-cancer-by-continents-per-100-000-women-per-yearbased-on-Ferlay_fig2_323991514

https://www.breastcancer.org/symptoms/understand_bc/statistics

https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf

https://www.ascopost.com/News/60293

https://www.ncbi.nlm.nih.gov/books/NBK470268/

https://www.archivesofpathology.org/doi/pdf/10.5858/arpa.2018-0463-RA

https://www.europadonna.org/breast-cancer-facts/

https://cran.r-project.org/web/packages/bartMachine/vignettes/bartMachine.pdf

https://www.researchgate.net/post/What_is_the_difference_between_bart_and_cart

https://towardsdatascience.com/decision-trees-d07e0f420175

W. Nick Street, William H. Wollberg and O.L. Mangasarian, “Nuclear Feature Extraction For Breast Tumor Diagnosis”

10. Appendix

For reference and clarity, information the previous work is included in this appendix. The information provided is a selection of information that is mentioned in previous sections of this report, is not intended to be an exhaustive description of the previous work.

Data Description

The dataset includes 569 samples and 30 numerical four-digit features.

Those 30 features are divided in 3 sets of the same 10 features. The sets correspond to the calculated values: the mean, the largest or worst and the standard error.

The 10 extracted features are:

1. Radius (mean of distances from center to points on the perimeter)
2. Texture (standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (perimeter^2 / area - 1.0)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension ("coastline approximation" - 1)

Upon inspection of the dataset, the key findings are that there are no missing values and no evidence of noisy or inconsistent data.

Finally, the target is the diagnosis of the breast tissue: malignant or benign. Since the target is discrete and binary, this is a binary classification problem.

Dataset retrieved from: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Data collection

When a fine needle aspirate biopsy is performed, a small drop of fluid is obtained and the sample is evaluated by a pathologist. In this case, the sample is expressed onto a glass slide and stained. The image for digital analysis is generated by a color video camera mounted on top of a microscope. To successfully analyze the digital image, a user selects a set of nuclei and specifies each location of each cell nucleus boundary through a graphical user interface. The system then analyzes and computes features relative to the size, shape and texture of those nuclei. Specifically, it calculates the mean value, the largest value and the standard error of each feature over the range of cells selected.

Link to paper accompanying the dataset:

https://www.researchgate.net/publication/2512520_Nuclear_Feature_Extraction_For_Breast_Tumor_Diagnosis

Modelling

Five of the thirty features were selected following feature importance results of the Decision Tree algorithm: area_worst, texture_mean, smoothness_worst, concavity_mean and area_mean.

The following classifiers were trained and optimized via hyper parameter tuning: Logistic Regression, Support Vector Machine (SVM), XGBoost, K-Nearest Neighbors, Random Forest and Decision Tree. This was an iterative exercise: the hyper parameters range were updated and the tuning process was re-executed until the overall accuracy was stabilized. The train and test sets were split by 70% and 30%, respectively, and the pipeline included standardization of all the features. GridSearchCV was used to methodically build and evaluate all the models for each combination of algorithm parameters specified in each of their respective parameter grids.

Finally, given the unbalance of the class labels (62.7% benign, 37.3% malignant) and the limited data sample (569), stratified k-fold cross-validation was used as a cross-validation method because it generally results in a less biased or less optimistic estimate of the model skill than other methods. Specifically, k was set to 10 following a general recommendation.

Best parameters of the Logistic Regression and Random Forest classifiers:

- Best hyper parameters of Logistic Regression:
  {'C': 10, 'multi_class': 'ovr', 'solver': 'newton-cg'}
- Best hyper parameters of Random Forest:
  {'max_depth': 5, 'min_samples_leaf': 5, 'min_samples_split': 2, '_n_estimators': 100}