# Individual Assignment INF554

Max Rehman Linder

October 2023

## Question 1

An impurity measure is a function $\phi : P \rightarrow \mathbb{R}$ where:

$$P = \{\sum_{k=1}^{K} p_k = 1, p_k \in [0,1]\} \tag{1}$$

And:

- $\phi$ has a unique maxima when all classifications are equally likely, $p_1 = p_2 = .... = p_K$.
  This means that the node is the least pure when when the data within that node is equally likely to assume any of its K possible classifications.
- $\phi$ has a unique minimum when one $p_k = 1$ and all others are equal to 0.
  Which corresponds to all data within that node with certainty belonging to one specific class. This means that all data is correctly classified at that node.
- $\phi$ is symmetric.
  The function being symmetric in this setting corresponds to the node not being any more or less pure depending on in which order the data is fed into the tree.

With K=1, we have $p_1 = x$, $p_2 = 1 - x$ and $x \in [0,1]$. The gini-index can then be written as $1 - x(1-(1-x)) + (1-x)(1-x) = -2x^2 + 2x$
We can now take the derivative of x to find the maxima:
$\frac{d}{dx}(-2x^2 + 2x = -4x + 2)$, setting the equation equal to 0 gives us $x = 0.5$
Deriving another time gives us $\frac{d^2}{dx^2}(-2x^2 + 2x) = -4$, which tells us that the function is strictly concave, and therefore that the extreme point is a maxima that occurs when $p_1 = p_2 = 0.5 \Rightarrow (\frac{1}{K}, \frac{1}{K})$.

As for the minima, $-2x^2 + 2x = 0$ gives us $x = \{0, 1\}$, which is at the edges of the the set $x \in [0,1]$ and corresponds to $(1,0)$ and $(0,1)$

For symmetry, shift x to be at $0.5$.
$-2(x+0.5)^2 + 2(x+0.5) = -2x^2 - 2x - 0.5 + 2x + 1 = -2x^2 + 0.5$. Since all terms are symmetric, the function itself is symmetric around $x = 0.5$, which is the maxima of the function.

## Question 2

For a given observation out of a set of size N, the chance of that observation not being sampled is $1 - (\frac{N-1}{N})^N$.
Let's compute : $(\frac{N-1}{N})^N$ as N goes to $\infty$

$$lim_{N\rightarrow\infty}(\frac{N-1}{N})^N = lim_{N\rightarrow\infty}(\frac{N}{N-1})^{-N} = lim_{N\rightarrow\infty}(\frac{N+1}{N})^{-N}$$
$$= lim_{N\rightarrow\infty}(1 + \frac{1}{N})^{-N} = lim_{N\rightarrow\infty}(1 + \frac{1}{N})^{N^{-1}} = e^{-1} = 1/e$$

So the chance for a observation not being sampled is $1 - 1/e$.

# Question 3

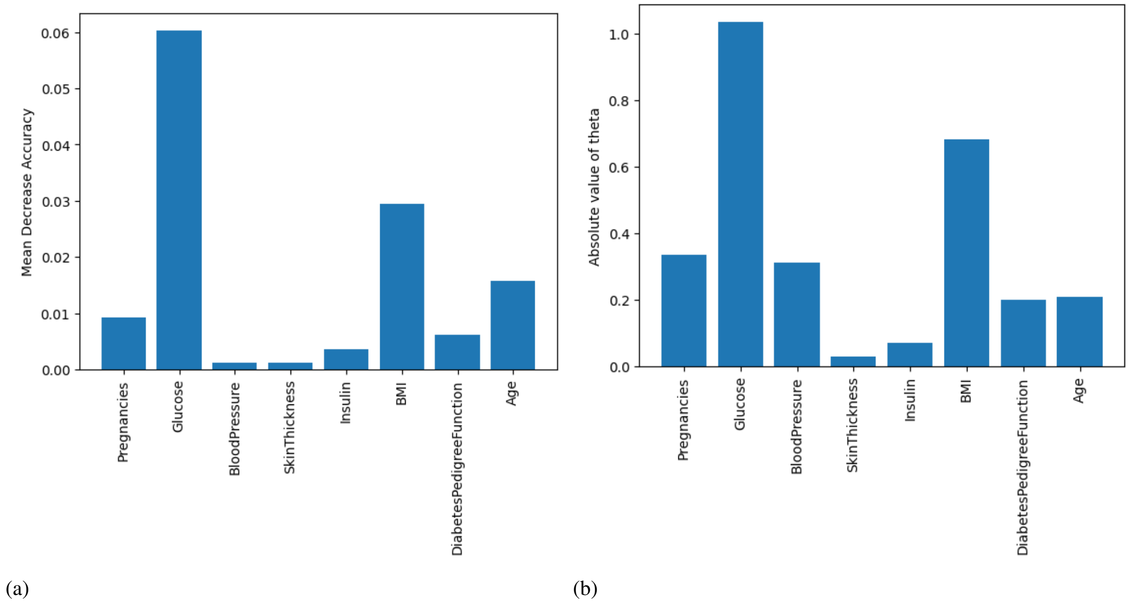


Figure 1: Values of MDA and the absolute value of regression coefficients for all 8 features.

Looking at plot a) that shows that the most important features are glucose, BMI and age. The coefficients would indicate that blood pressure and pregnancies are more important. However, one can only have more pregnancies the older one is and blood pressure is correlated with BMI. So it might be the case that these parameters get a boost since they are correlated to other features that are more critical.
The MDA analysis can therefore more clearly show which features that are more significant.

# Question 4

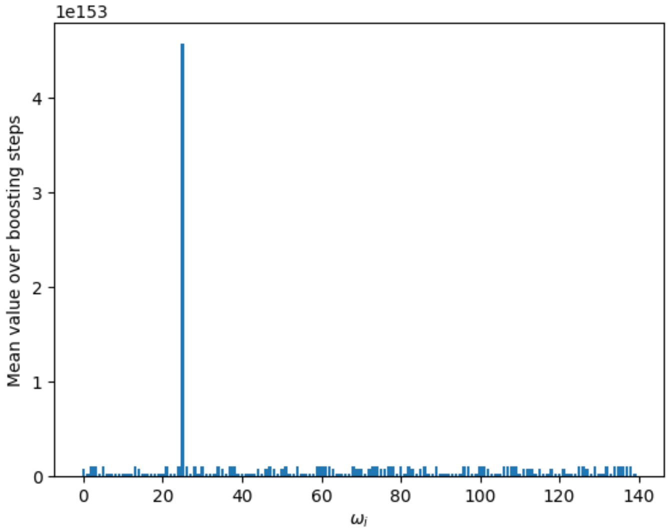


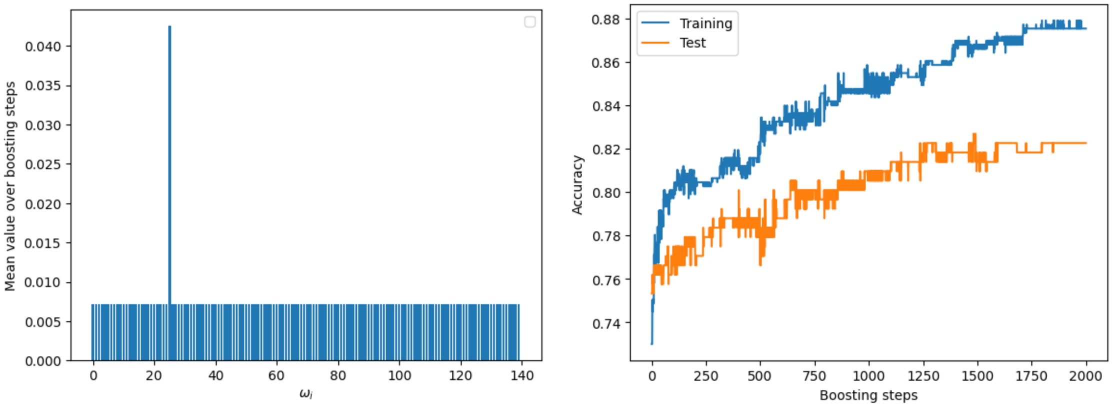
Figure 2: Bar-plot of average value of weights

The observed issue is that almost of all of the weight is concentrated on one weight. What this means is that the data corresponding to this observation is very different to all others. So when the model is set up the classification continuously becomes wrong for this observation, and the value for its weight grows exponentially.

The result may be a model that scarifies precision only to get this observation right.

# Question 5

My proposition is to limit the how big one singular weight can become relative to the others. So a weight is set the normal way $\omega_{i+1} = \omega_i \cdot e^{a_b I(y_i \neq C_b(x_i))}$, But then for all $\omega_i = min[\omega_i,\ 5 \cdot mean(\Omega]$, where $\Omega$ is the vector containing all $\omega$. In other words, no $\omega_i$ is allowed to be more than five times larger than the mean for the other omegas. It limits the concentration of weight while still preserving the same method of calculating $\omega$.



(a)

(b)

Figure 3: (a) shows the $\omega$-distribution with the modified method of calculating $\omega$. (b) shows the result from task 7 using the modified method.

Worth noting in Figure 3 is that one weight doesn't skyrocket to an order of magnitude of $10^{153}$ and the result from Task 7 remain identical to the unmodified version of calculating $\omega$. This presumably because no weight ever becomes more than 5 tome larger that the weight average.

# Question 6

## XGBoost

XGBoost is a gradient-boosted decision tree. It combines aspects from gradient descent and decision trees by first creating a number of decision trees and then formalizing it to create a gradient decent optimization problem. The aim is to minimize errors for subsequent decision-trees and the outcomes are based on the gradient of the error between each iteration. The result is a highly salable and flexible model that has won more or less every big Kaggle competition.

## LightGBM

LightGBM uses tree based algorithms together with gradient boosting, just like XGBoost. It is meant to be really fast and memory efficient, thereof it's name LightGBM. The way in which is faster is grows its trees leaf-wise, which means that is chooses the leaf that it thinks will decrease loss the most. Another central feature is using histograms which bunches data together in order to speed up the training.