

**TeTe - Breast Cancer Prediction model**



Naam: Amaryllis Lee  
Datum: 29 juni 2020

# Inhoudsopgave

[**Probleembeschrijving**](#_3dy6vkm) **2**

[**Eisen**](#_1t3h5sf) **3**

[**Algoritme**](#_4d34og8) **4**

# **Probleembeschrijving**

Borstkanker heeft in de laatste jaren wereldwijd een grote impact veroorzaakt op zowel traumatisch - als financieel niveau. In Nederland krijgt 1 op de 7 vrouwen borstkanker. In 2019 kregen er 14.940 mensen borstkanker terwijl in 2018 3.088 mensen hiervan stierven. En er wordt steeds gezocht via onderzoeken en AI-technieken hoe ze de diagnose nauwkeurig of exact kan doen zodat de behandeling en de gevolgen hiervan minder ingrijpend zijn.

Het doel van dit algoritme is om een voorspellingsmodel te ontwikkelen voor de diagnose van borstkanker voor medische klinieken.

**Waarom is dit van belang?**In de praktijk worden vrouwen vaak verkeerd gediagnosticeerd. De diagnose kan zijn kwaadaardig terwijl in realiteit het goedaardig is of de diagnose is goedaardig terwijl het kwaadaardig is in realiteit. Beide gevallen hebben grote gevolgen voor de patiënt. Indien de borstkanker kwaadaardig is en in realiteit het goedaardig is, is het gevolg dat de patiënt een lange en ingrijpend behandeling die pijnlijk en tramautisch is, moet volgen. Andersom indien de borstkanker kwaadaardig is en de diagnose zegt goedaardig, zal de patiënt niet een behandeling hiervoor krijgen.

Daarom is het van belang dat de nauwkeurigheid van de diagnose bij borstkanker (bijna) exact is. Voor deze reden wordt gewerkt aan een voorspellingsmodel (applicatie) met een grote nauwkeurigheid voor borstkanker. Voor deze application zal er gebruikt worden van de statistische model Decision Tree .

Er is een naam gegeven voor dit project. Het naam die gegeven wordt is TeTe. Het model is gemaakt op Aruba en daarom wordt een papiamentse naam ervoor gekozen, namelijk TeTe( = borst) .

# **Eisen**

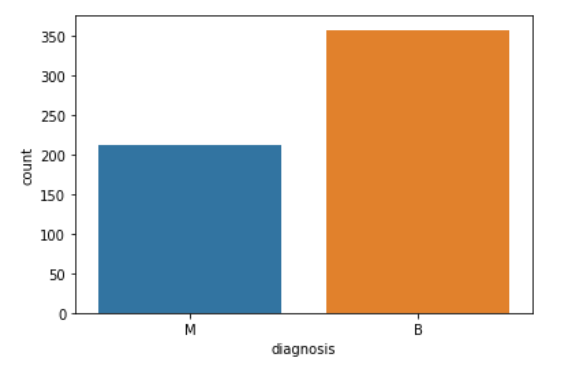
* De nauwkeurigheid van de algoritme moet bijna 100% (of relatief dichtbij) zijn.

Dit is belangrijk om de aantal gevallen van *false negatives* en *false positive*  te verminderen.

* Data moet duidelijk en begrijpelijk gepresenteerd worden
* Hergebruik voor andere toepassingen die kunnen voordoen;
* Applicatie moet snel kunnen draaien.
* Algoritme moet goed gedocumenteerd zijn.

# **Algoritme**

Aan de hand van een dataset met een diagnose van patiënten met borstkanker wordt een voorspellingsmodel voor het detecteren van borstkanker ontwikkeld. De dataset is afkomstig van dr. Wolberg en was opgenomen via de fijne naald aspiratie (FNA)[[1]](#footnote-0). Van de opgenomen monster van een cel via FNA, wordt 10 eigenschappen van de cel bepaald. Van elk van deze eigenschappen wordt *het gemiddelde*, de *standaardfout* en de '*worst' of slechte (gemiddelde van de drie grootste waarden)* berekend resulterend in een totaal van 30 eigenschappen. Deze eigenschappen vormen de basis van de dataset die gebruikt wordt voor het voorspellingsmodel.

Het resultaat van de diagnose in de dataset wordt visueel gepresenteerd. Figuur 1 geeft de presentatie hiervan. Uit deze dataset van 569 patiënten blijken 357 borstkanker goedaardig (benign) te zijn en 212 kwaadaardig (malign). Dat betekent dat 37.3% van de diagnose van de 569 patiënten kwaadaardig zijn. 

Hierna wordt de dataset gesplitst in een training (75%) en testing (25%) dataset. De training dataset heeft een bekende resultaat en wordt gebruikt om de voorspelling uit te voeren. De testing dataset wordt gebruikt om de voorspelling op ons model te testen.

De algoritme wordt gekozen om het beste resultaat te voorspellen. In de dataset hebben we de afhankelijke variabele Y met twee keuzes, namelijk M (Malign=kwaadaardig) of B (Benign=goedaardig). Voor deze reden wordt gekozen voor een classificatie algoritme van supervised learning[[2]](#footnote-1).

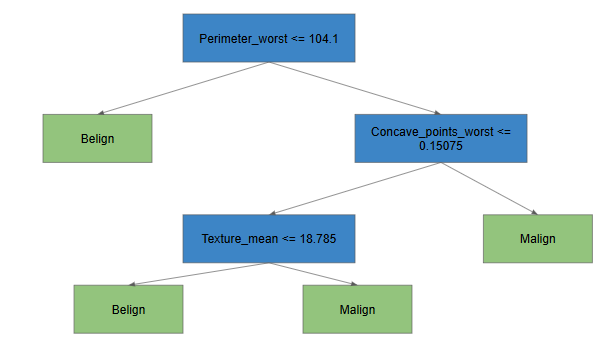
Er zijn verschillende methoden van classificatie algoritmes en ik heb gekozen voor Decision Tree Classification. De reden van mijn keuze is omdat het veel gebruikt wordt in de praktijk voor voorspellingsmodellen.

Voor het maken van de decision tree algoritme worden 3 functies ontwikkeld, namelijk Tree, Classify en Accuracy.

Hieronder wordt elke van de 3 functies van de decision tree algoritme uitgelegd:

1. **Tree**

Deze functie is bedoeld om de Decision tree van de training dataset te bepalen.



*Visualisatie van de Decision Tree*

1. **Classify**

Met behulp van de tree kunnen we een datapoint classificeren. Hierdoor kunnen we de diagnose van de test dataset voorspellen.

['M', 'B', 'M', 'B', 'M', 'B', 'B', 'B', 'B', 'B', 'B', 'M', 'M', 'B', 'M', 'M', 'B', 'M', 'B', 'M', 'B', 'B', 'B', 'B', 'B', 'M', 'B', 'M', 'B', 'M', 'B', 'B', 'B', 'B', 'M', 'M', 'B', 'B', 'M', 'M', 'B', 'M', 'M', 'B', 'B', 'B', 'B', 'B', 'M', 'B', 'B', 'B', 'B', 'M', 'B', 'B', 'M', 'M', 'M', 'B', 'B', 'B', 'B', 'M', 'M', 'M', 'M', 'M', 'B', 'B', 'B', 'M', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'M', 'M', 'M', 'B', 'M', 'B', 'B', 'B', 'M', 'B', 'B', 'B', 'B', 'B', 'M', 'B', 'M', 'M', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'M', 'B', 'M', 'B', 'B', 'B', 'M', 'B', 'B', 'B', 'B', 'B', 'B', 'B', 'M', 'B', 'B', 'B', 'B', 'B', 'B', 'M', 'M', 'M', 'B', 'B', 'B', 'B', 'B', 'B', 'M', 'B', 'M', 'M', 'M', 'M', 'B', 'B']

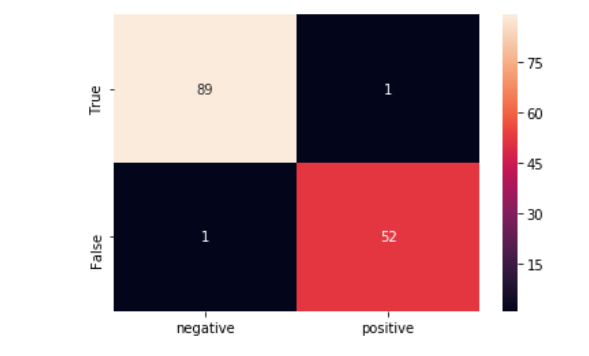
*Weergave van de voorspelde diagnosis van de test dataset*

1. **Accuracy**

Door de voorspelde diagnose te bepalen kunnen we de nauwkeurigheid van de Decision Tree algoritme berekenen. De nauwkeurigheid van de Decision Tree algoritme is 88.73239%.

Model Decision Tree Classifier Testing Accuracy ="88.73239436619718%"

De volgende grafiek presenteert ook het resultaat.



We kunnen concluderen dat ons voorspellingsmodel een goed model is om borstkanker te detecteren.

Bijlage 1. Uitleg van de werking van de Decision Tree algoritme

1. . Fijne Naald Aspiratie (FNA) is een diagnostische procedure die wordt gebruikt om klonters of massa's te onderzoeken. Bij deze techniek wordt een dunne (23-25 ​​gauge) holle naald in de massa gestoken voor het nemen van monsters van cellen die, na te zijn gekleurd, onder een microscoop worden onderzocht (biopsie). Biopsies met fijne naaldaspiratie zijn zeer veilige kleine chirurgische procedures [↑](#footnote-ref-0)
2. Supervised learning is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing. Supervised learning problems can be further grouped into **Regression** and **Classification** problems. A **regression** problem is when the output variable is a real or continuous value, such as “salary” or “weight”. A **classification** problem is when the output variable is a category like filtering emails “spam” or “not spam”. Unsupervised Learning : Unsupervised learning is the algorithm using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. [↑](#footnote-ref-1)