# Project 2 Applied Programming – Weka Machine Learning

## De opdracht

Het maken van een toepassing die het verschil tussen een clipart en foto kan zien. De toepassing moet dit doen aan de hand vaan een beslissingsboom en opvallende kenmerken van een foto of clipart. Hierbij wordt de “Weke data mining and machine learning” tool gebruikt, die gemaakt is door de Nieuw-Zeelandse University of Waikato.

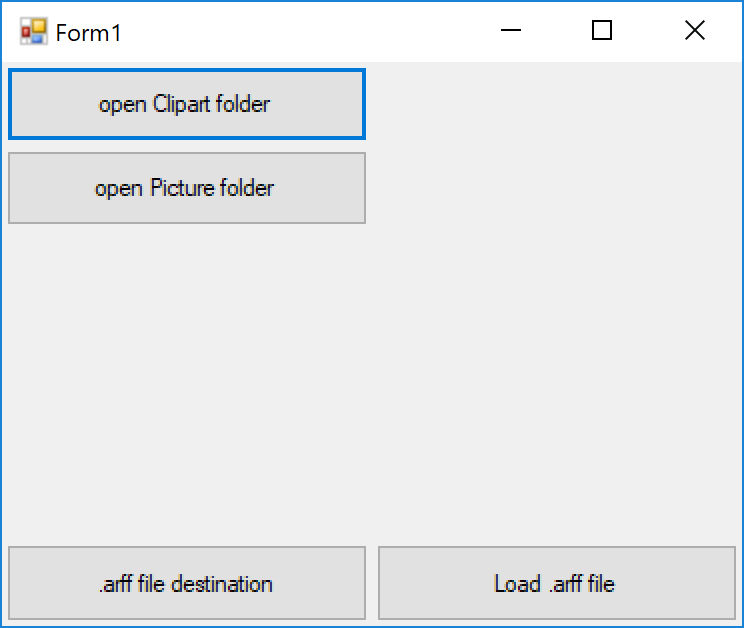
## Het onderzoek

Het onderzoek begint bij het verstaan van de werking van de Weka tool. Hierbij staan online vele lessen gegeven door Professor Ian H. Witten, die uitgebreid uitlegt hoe het programma in zijn werking gaat. De Weka tool is een programma waaraan een databestand gegeven moet worden en aan de hand daarvan een beslissingsboom maakt. Het data bestand is een .arff bestand, daarin zitten alle kenmerken van een afbeelding, waarvan laatste eigenschap aanwijst of het clipart of een foto is. Aan de hand van de inhoud van dit bestand, kan Weka een beslissingsboom vormen. Hoe meer data er in het bestand staat, hoe beter de beslissingsboom. [1]

Na het begrijpen van de werking van Weka, moet men weten welke kenmerken een clipart of foto definieert. Aan de hand van de gevonden kenmerken zal het uiteindelijke .arff bestand gegenereerd worden.

## De praktische uitwerking

Het .arff bestand word door een C#-toepassing gegenereerd. Deze werkt simpel aan de hand van een GUI waar de folder met clipart en de folder met foto’s geselecteerd worden, de locatie waar het .arff bestand gegenereerd moet worden en de knop om het bestand te genereren.

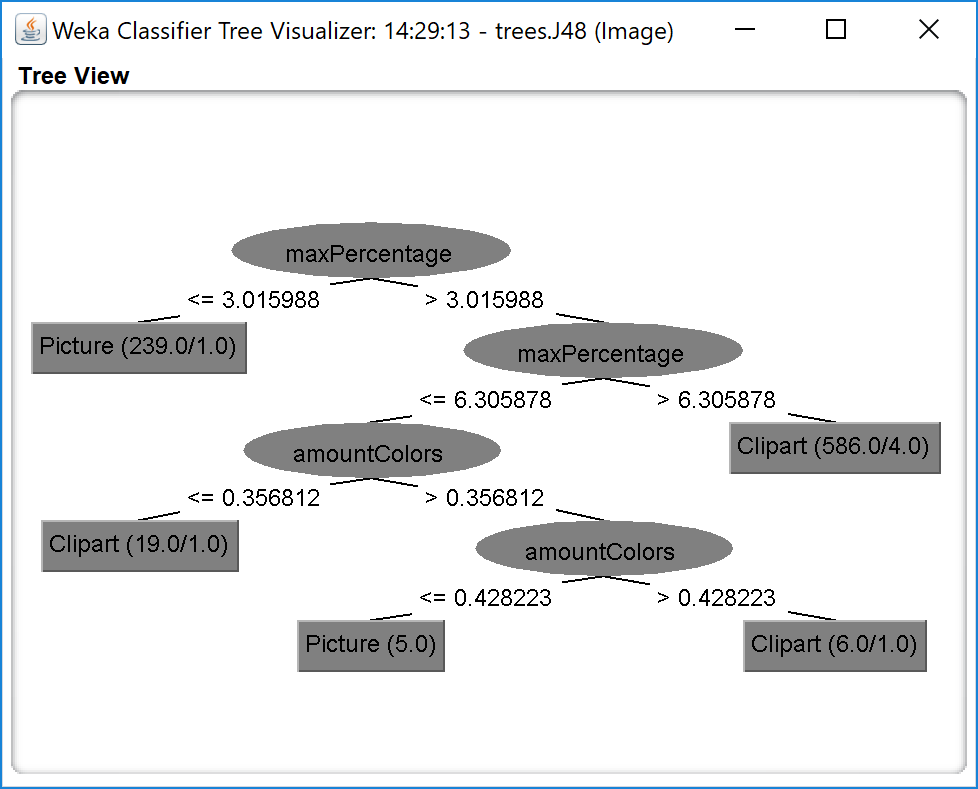


Het .arff bestand gebruikt drie kenmerken:

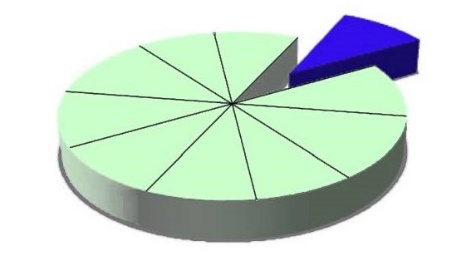
* Het percentage pixels dat een verschillende kleur hebben.
* Het percentage pixels die de meest voorkomende kleur zijn.
* Of er een doorzichtige pixel is

Deze toepassing gaat elke pixel van elke afbeelding af om alle kenmerken van elke afbeelding te kunnen krijgen. Dit kan een tijd duren, aangezien het bij de eerste versie van de toepassing slechts één core gebruikt. Hierbij duurt het ongeveer tien minuten om het .arff bestand te generen. Bij versie twee gebruikt de toepassing parallel loops. Hierbij zal elke core gebruikt worden en kan het .arff bestand in dertig seconden gegenereerd worden.

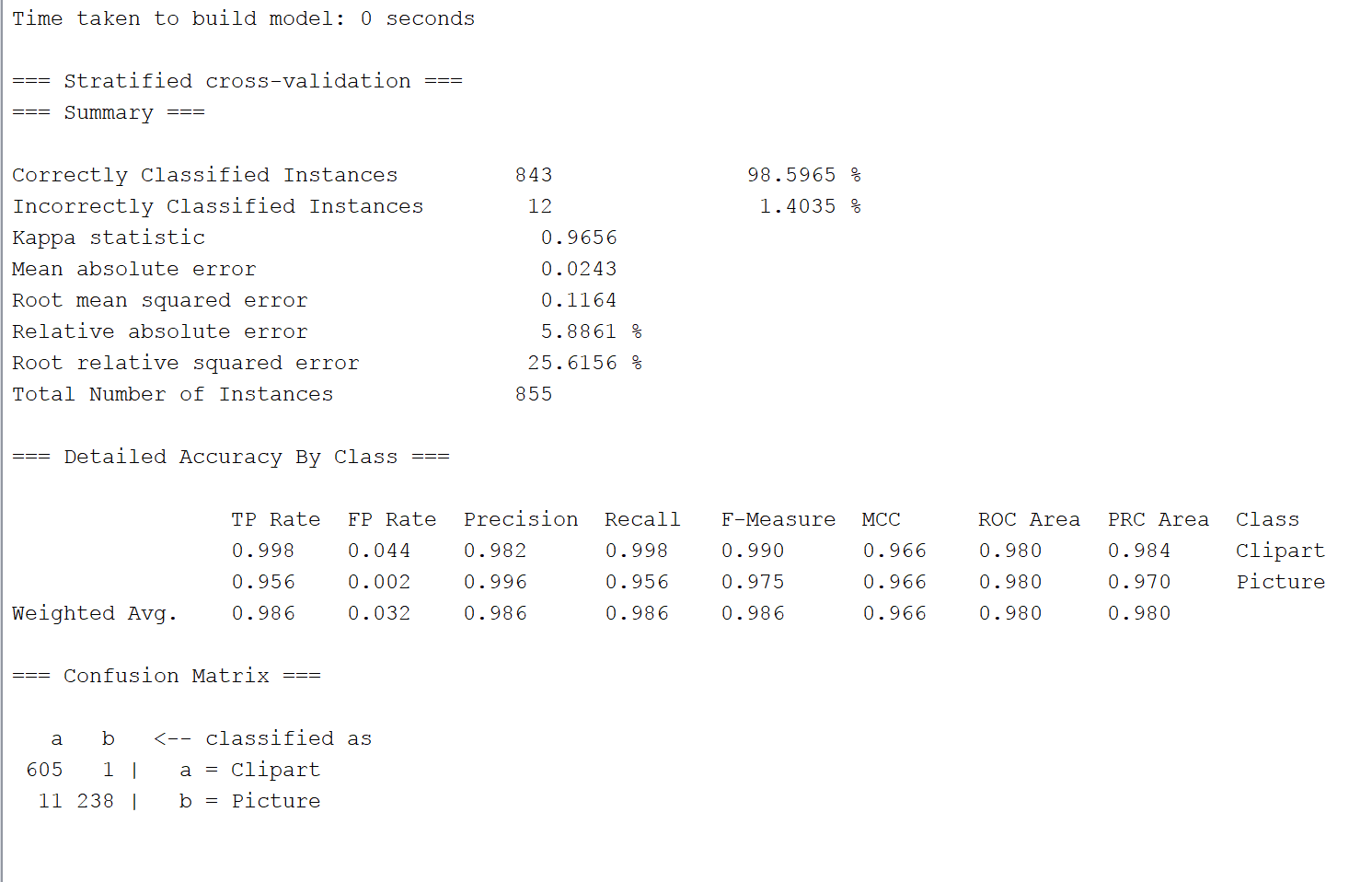
Bij het inladen van het .arff bestand in Weka kan met de j48 boom structuur de volgende beslissingsboom gegenereerd worden.



Hieruit blijkt dat Weka de doorzichtige pixel eigenschap niet gebruikt. Daardoor zal het programma dit in zijn beslissingsboom ook niet gebruiken.

Weka test je data aan de hand van cross-validatie. Hier wordt alle data in tien gelijke stukken opgesplitst. Negen van de tien stukken worden gebruikt om Weka te trainen en de bovenstaande beslissingsboom te genereren. Het overige stuk wordt gebruikt om te testen en te zien hoeveel van de foto’s Weka juist krijgt.

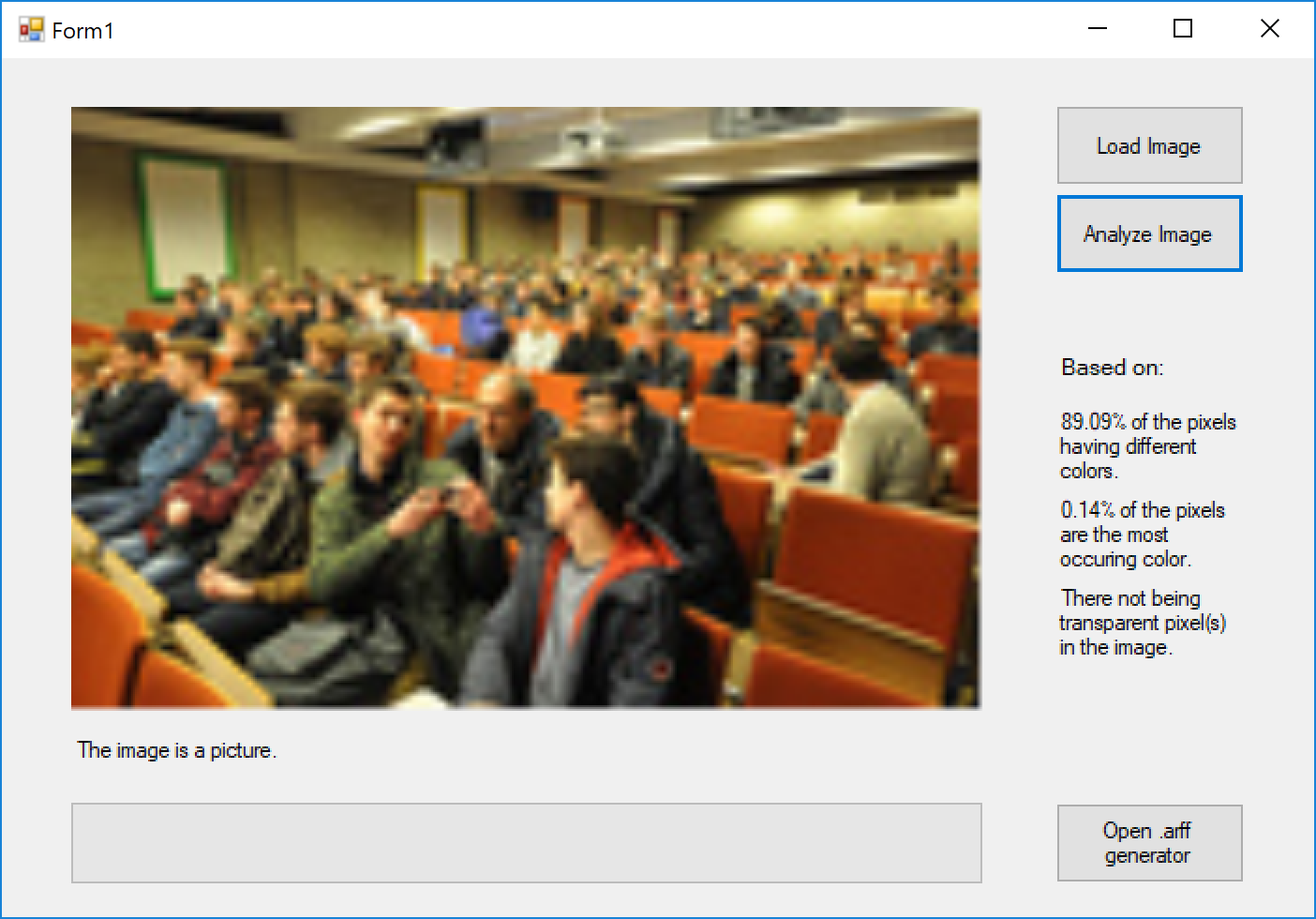
Hieronder zien we de uitkomst van zo een cross-validatie test. We kunnen hieruit afleiden dat de beslissingsboom een faalpercentage heeft van ongeveer 1.4%.



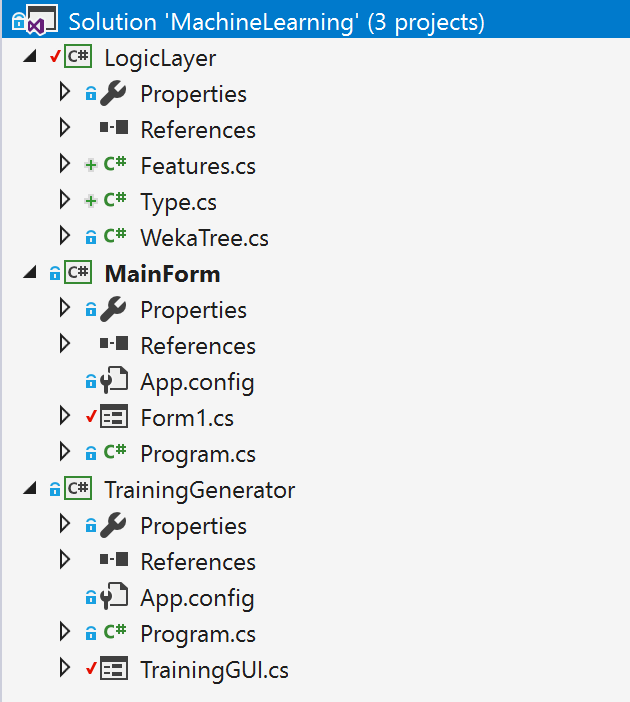
De meeste fouten gebeuren bij foto’s. Dit zijn meestal afbeeldingen zonder veel kleurverschil zoals zwart-wit foto’s, of foto’s met/van sneeuw. Dat toont ook direct het nadeel van testen op het voorkomen van kleuren, ook al zijn dit maar een klein percentage van de afbeeldingen.



Het uiteindelijke programma is een GUI waar één afbeelding wordt ingelezen. Als de gebruiker dan op “Analyze Image” duwt wordt de uitkomst getoond. Rechts vanonder kan de .arff generator geopend worden.



### Code structuur

Het programma bevat drie projecten; twee grafische lagen en één logische laag. De logische laag bevat een enumeratie met de twee types (Clipart en Picture), een klasse ‘’Features’’ waar al de eigenschappen van een afbeelding worden berekend en een klasse ‘’WekaTree” waar de beslissingsboom in zit verwerkt.

De GUI TrainingGenerator is de GUI om het .arff bestand te genereren, deze maakt gebruik van de WekaTree klasse en de Type enumeratie.

De tweede GUI MainForm is het hoofdprogramma waar de ene afbeelding in wordt gelezen en aan de hand van de beslissingsboom word geanalyseerd.

## Besluit

Deze opdracht is voor studenten goede manier om de basis van machine learning te begrijpen. Door de vele lessen online te volgen en dan te werken met Weka, worden de basis concepten van machine learning duidelijker. Weka is een makkelijk te gebruiken en krachtige toepassing om met data om te gaan en kan voor veel meer gebruikt worden dan de toepassingen in dit project.

De manier van kleuren analyse die in het programma gebruikt worden zorgt voor een 98.5% succesrate. De grootste zwakte zijn zwart-wit foto’s, sneeuw-foto’s en de processorkracht nodig om elke afbeelding af te gaan voor het .arff bestand te genereren. Een andere manier is om de randen van een afbeelding te analyseren. Hierbij is minder rekenkracht nodig. Ook dit geeft een redelijk hoge succesrate.

# References

|  |  |
| --- | --- |
| [1] | I. H. Witten, "Data Mining With Weka," August 2013. [Online]. Available: https://www.youtube.com/watch?v=LcHw2ph6bss&list=PLm4W7\_iX\_v4NqPUjceOGd-OKNVO4c\_cPD. |