Pathogen outbreaks represents a serious threat to the forestry industry in South Africa and is an **active area of research** for the Forestry and Agricultural Biotechnology Institute (FABI) at the University of Pretoria. The focus of this project was to develop a **machine learning framework** based on information derived from **plantation samples** collected over a **period of 24 years**, that would be able to accurately **predict specific pathogen type outbreaks** across the country. This objective was achieved and paves the way for a proactive approach to addressing the pathogen outbreaks in the South African forestry industry going forward.

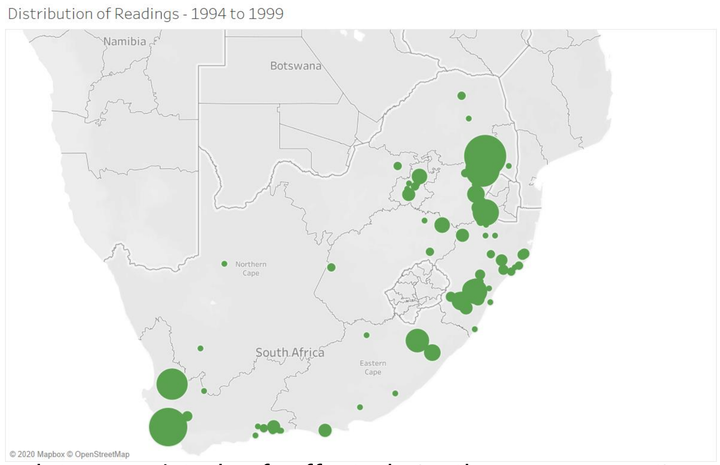
A Machine Learning Framework for Predicting Pathogen Outbreaks in South African Forestry Plantations

Introduction

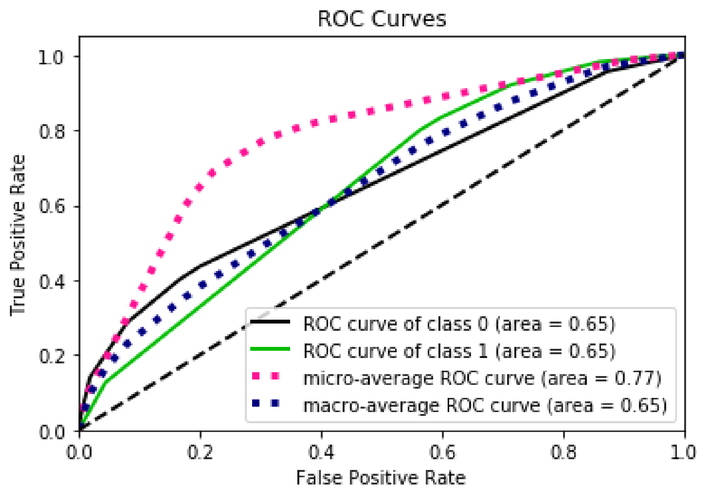
- The FABI Tree Protection Co-operative Program was established with the intention of monitoring and addressing pest and diseases which threaten the national, as well as global forestry sectors.
- FABI has identified the use of advanced machine learning models in the fight against pathogen outbreaks as an opportunity for it to proactively identify such outbreaks and in doing so, minimize the extent / impact of these on the forestry industry.
- This in turn should result in significant cost savings and ensure the sustainability of the industry.
- Tree and pathogen samples have been collected by FABI and its various industry partners for over 20 years.
- The aim of this project was to develop a machine learning algorithm using the information derived from these historic samples, that will provide FABI researchers with a tool that can be used to screen new samples collected in the future to determine (a) whether a pathogen is present in the plantation based on the appearance / symptoms of the sample, and (b) what specific type of pathogen is likely to be present.

Methods

- In total, 17 898 plantation / tree specimens which were collected between 1994 and 2018, were used to develop a machine learning algorithm able to accurately predict specific pathogen type outbreaks.
- These specimens were collected from various plantations across South Africa (below).

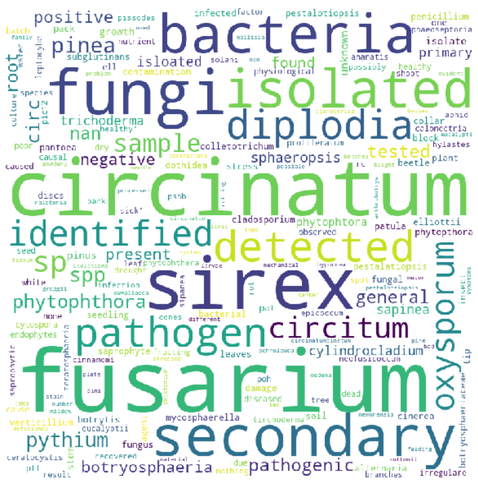


- The data consisted of affected (pathogen present) and clean / unaffected samples
- *Fusarium Circinatum* was the most common pathogen detected in the affected samples.
- Missing values were prevalent in the data which affected the overall modelling approach.
- Free text fields were widely used to explain / describe the symptoms / overall appearance observed in the respective samples (see word cloud example on the right).
- Natural Language Processing (NLP) combined with a clustering algorithm (Gaussian Mixture Model) was used to assign observations to clusters based on the text information contained in the captured free text fields (i.e. “Symptoms” and “Overall Appearance”)
- A two-phased *k*-nearest neighbor model was developed to firstly predict the likelihood that a sample is infected with a pathogen, and secondly to predict which type of pathogen it is likely to be if a pathogen is deemed to be present.
- Results were found to be promising, with the model being able to effectively distinguish between the pathogen and non-pathogen specimens.

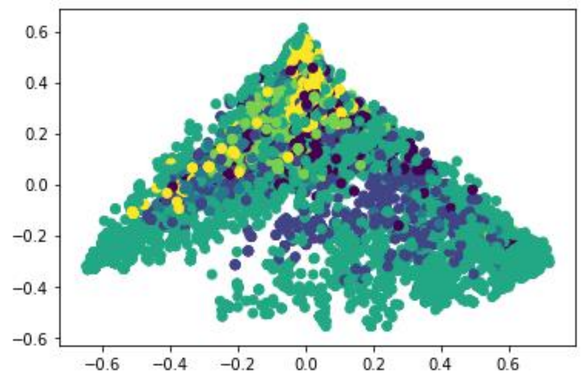


Discussion

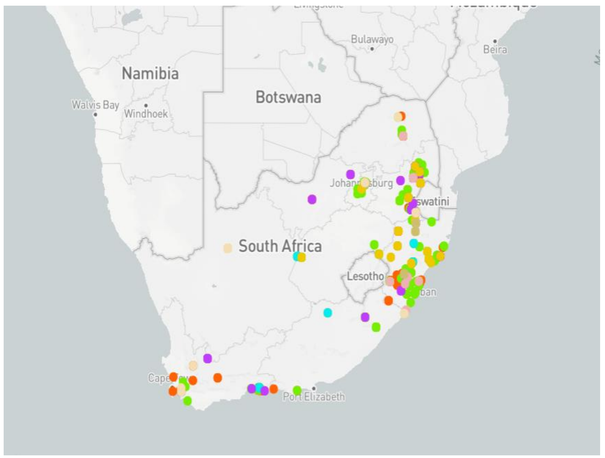
- The results obtained from this study have been very promising.
- NLP was successfully incorporated into the machine learning framework.
- The final algorithm has been able to meet the objectives set out in the beginning of this project, namely, to accurately predict pathogen outbreaks in advance and identify the specific pathogen.
- Proposed next steps will be to expand on the existing machine learning model and consider more variations of machine learning frameworks.



- Word-clouds were extensively used during the exploratory data analysis phase to interrogate the free text fields in an unstructured manner.



- The NLP and clustering combination proved invaluable in extracting sensible and homogenous clusters from the free text fields which in turn proved to be very predictive. The figure above contains a two-dimensional, principal component representation of the NLP features extracted from the free text fields, with an overlay of the derived clusters.



- The final model output consists of a visual, geolocation representation of the predicted outbreaks, together with the specific (predicted) pathogen.

A. Lal (u19204214)
R. Smit (u21150177)

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

FABI