XGBoost uses **weather records** to achieve up to **81%** classification accuracy for **Leptocybe** and **Sirex**

Enabling weather-based decision making for forestry pest and disease management

INTRO

This project analyses weather and pest-prevalence data in South Africa to identify possible relationships between changing climatic conditions and pest populations using machine learning techniques. Three predominant datasets were received from the Forestry and Agricultural Biotechnology Institute (FABI) for contribution to this project, namely:

- Temperature and rainfall records of roughly 6000 weather stations across South Africa.
- Sirex noctilio (Sirex) pest inspections
- Leptocybe invasa (Leptocybe) pest inspections

METHODS

1. Perform feature engineering to link raw weather to pest data (see right side panel).
2. Train XGboost and Support Vector Machine (SVM) models using several perturbations of the parameters generated in 1.
3. Test for stability using 3-fold cross validation.
4. Prepare models for deployment:
   a. Prepare interface to allow an user to provide weather measurements and location.
   b. Pass inputs (4a) through feature engineering (adapted to use the provided weather measurements as “actuals”) to generate model parameters .
   c. Pass parameters (4b) to pre-trained models (2).
   d. Generate results for the user-defined location (4a) and across South Africa.

Gené Fourie, **Connor** McDonald

RESULTS

Table 1 indicates the modelling results. The 3-fold cross validation generates fold accuracies with standard deviations of less than 1.0% which indicates a high degree of model stability. The XGboost model performs 10-15% better than the SVM model when classifying both pests. The models classify Leptocybe more accurately than Sirex.

Model	Leptocybe		Sirex	
	Min. accuracy	*St. dev*	*Min. accuracy*	*St. dev*
XGBoost	81%	1.0%	64%	0.3%
SVM	63%	0.6%	55%	0.4%

Table 1: Modelling results

DISCUSSION

Application of the feature engineering and models to the user-defined parameters allows for pest prediction at the provided location and across South Africa, if the same input weather conditions are experienced. Extrapolation of the prediction enables the user to compare the results for the provided location against those in the vicinity to identify areas more prone to infestation. Figure 1 indicates the results of a Leptocybe prediction example, using XGBoost.

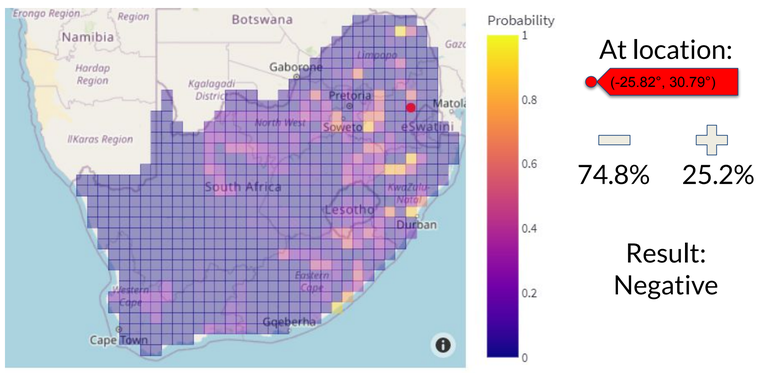

Figure 1: Prediction probability of Leptocybe example

FEATURE ENGINEERING

To link the weather and pest datasets and prepare the data for modelling, we used the following feature engineering algorithm:

1. **Aggregation**: Aggregate measurements per month and per active station between January 1950 and July 2019. A weather station is deemed ‘active’ if a record is provided for every day of a month.
2. **Matching**: Use Voronoi diagrams (Figure 2) to generate convex polygons around the stations per year-month period. Assign weather records of the generating station to pest inspections that fall in a given polygon.
3. **Averages**: Calculate the monthly average weather readings using the aggregate readings (1) from the assigned stations per pest ID (2).
4. **Actuals**: Collect the actual weather readings per month-year period (1) from the assigned station (2) for three years before and during the inspection year. This results in 48 data points per pest inspection.
5. **New features**: Subtract the average monthly readings (3) from the actual monthly readings (4) to determine a change from normal conditions.

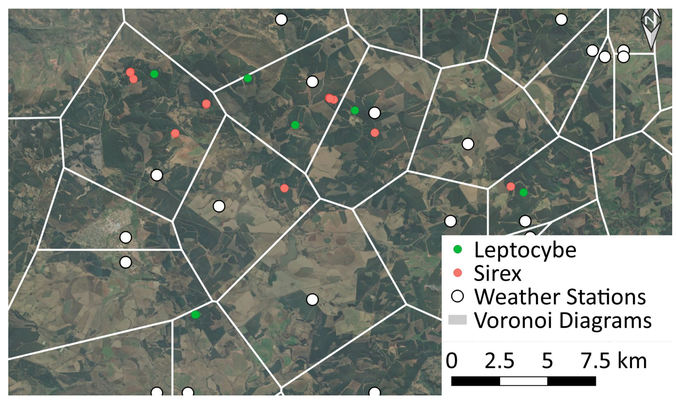

Figure 2: Voronoi Diagrams

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

Faculty of Engineering, Built Environment and Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Capstone Project - MIT 808

Course Coordinators:
Dr. Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Abiodun Modupe (abiodun.modupe@cs.up.ac.za)

Scan me