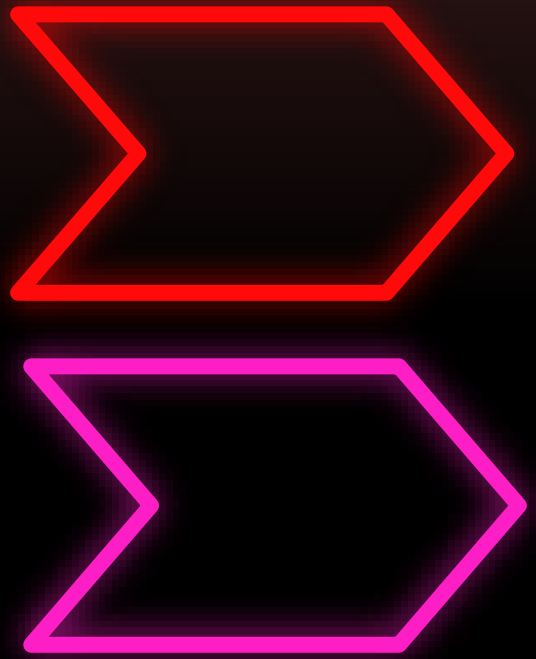
# Interactive web-application for:
# -Visualizing,
# -Analysing,
# -Predicting,
# vineyard data.



Computer Science Honours Project
Simeon Boshoff
Supervised by Dr Trienko Grobler and Dr Tara Southey

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

# Background

## Researcher data transformation

Researchers across various cutting-edge fields collect immensely insightful data. These datasets could provide even more utility as either a product or for further research if interactively visualized in the form of a web-application. This provides easy access to the data, anywhere, anytime.

# Background

## The vineyard dataset by Dr Tara Southey

Dr Tara Southey is a researcher in the field of viticulture and primarily studies the effects of climate change on the wine industry in the Western Cape.

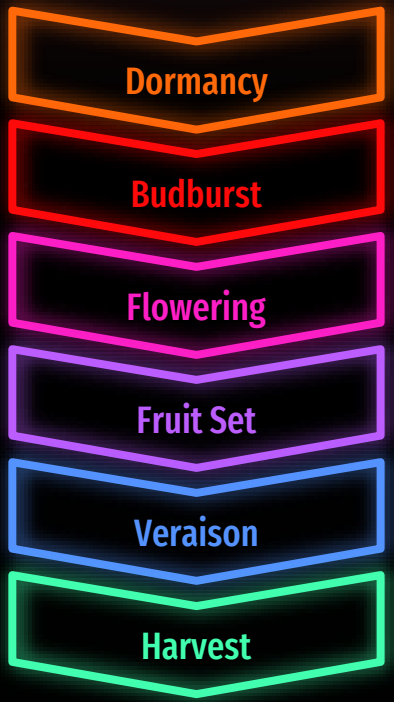
The dataset provided contains climate and vineyard behaviour variables from 4 wine farms over 4 years (2012-2015). This data is then analysed to find meaningful relationships between the above-mentioned variables.

Climate change can hence be measured by observing the behaviour of the grapevines.

# Background

## The vineyard dataset by Dr Tara Southey

### Vineyard Behaviour




Phenology can be defined as the study of periodic events in biological life cycles and how these are influenced by seasonal and inter-annual variations in climate, as well as habitat factors.

Every season, the grapevine undergoes a series of phenological events. The dates of these events are captured in the dataset.

Harvest date greatly affects the sugar-content of the grapes and as a result, the taste of wine.

# Background

## The vineyard dataset by Dr Tara Southey

Climate Data

### Temperature

The number of hours the vineyard was exposed to certain temperatures.

### Humidity

The number of hours the vineyard was exposed to certain humidity levels.

### Wind speed

The number of hours the vineyard was exposed to certain wind speeds.

### Rainfall

The total rainfall for the season.

# Tech Stack

**MySQL**

MySQL was used as the database solution, as it provides all required functionality while maintaining ease of use. MySQL Workbench also used.

**Plotly Dash**

Plotly Dash is a all-in-one web-application framework based on Python. Dash uses Flask for the back-end, and extended Bootstrap for the front-end.

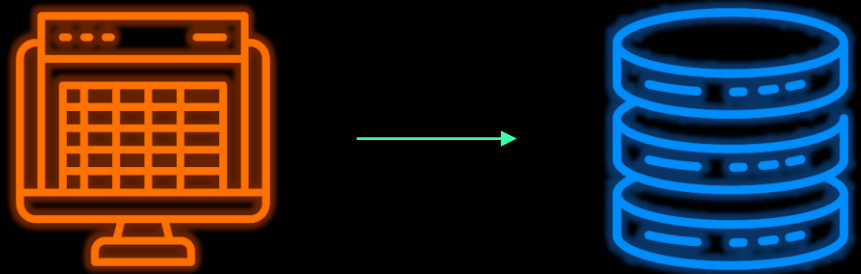
**Sci-kit Learn**

Sci-kit Learn was used for machine learning. Both discriminative and generative models were tested in the prediction of the harvest date.

# Database

-Data provided in the form of an Excel spreadsheet.
-45 rows with 580 columns.
-One row would represent a season, containing all data in every category.
-This is inefficient for computation.

**Solution:**
-Split data into different categories. (Vineyard info, Climate, Phenology, etc.)
-A Python script was developed that creates and populates tables in the database.
-Column reduction was favoured.

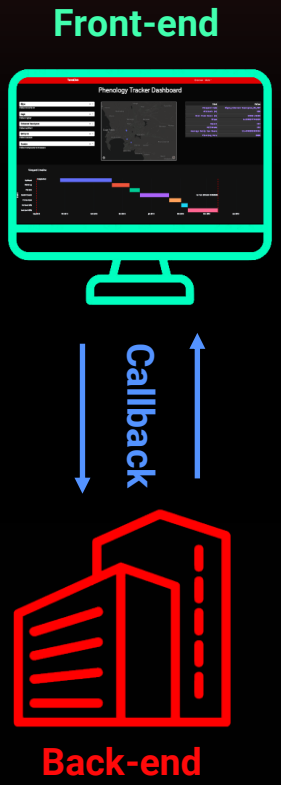


# Dash web-application

Plotly is a well-known Python library for creating charts and plots. Dash is the web-framework which is intended for creating interactive dashboards with Plotly objects.

Code for the front-end and the back-end are contained in the same Python file. The “callback” operator is used to facilitate interaction.

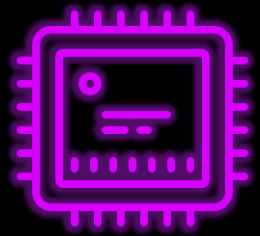
The web-app contains 4 pages:
-Phenology Dashboard (Visualizing all data)
-Analysis Engine (Provides tools to find patterns)
-Harvest Prediction (Visually shows prediction results on a timeline)
-About (Provides context on the subject)



# Machine Learning

-The dataset used for machine learning only contains 32 datapoints. (Cabernet Sauvignon)
-This makes it difficult to train a model properly.
-However, accuracy of the models should improve drastically when dataset is expanded.

Sci-kit Learn
-Gaussian Naive Bayes classifier is used for predicting if the harvest is early or late.
-Linear Regression is used for predicting the continuous variable iPcy (which is the measure for how late or how early the harvest will be).
-The datapoints have been separated manually to ensure independence between the training and test set.

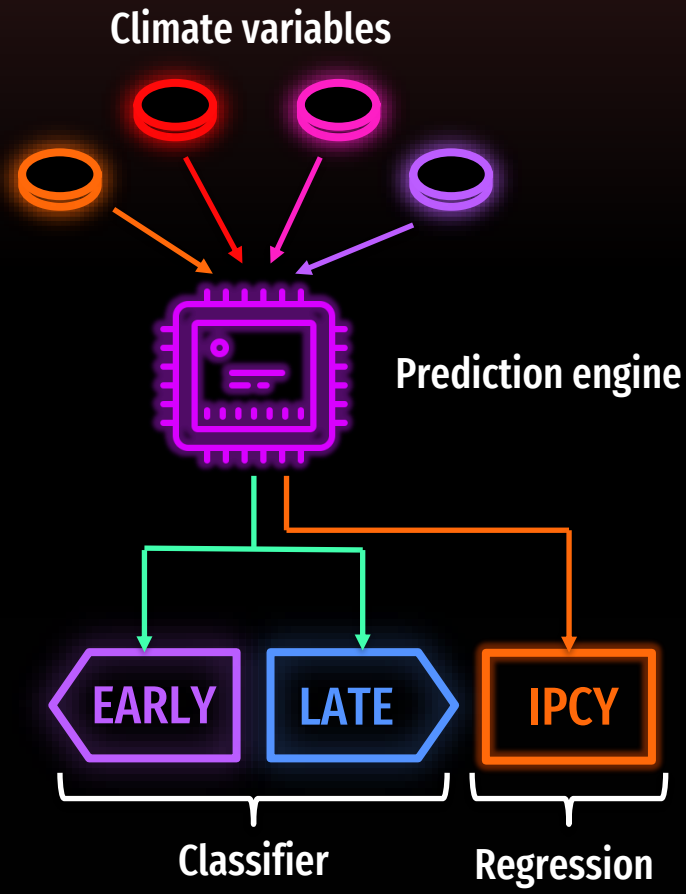


# Machine Learning Results

An input vector is comprised out the temperature hour values between 25° and 55° Celsius for October to December of the previous year, and January to March of the current year.

The classification model obtained an accuracy of 69% in predicting if the harvest will be early or late.

The regression model for predicting the iPcy values resulted in poor performance, most likely due to the limited training data.



# Testing

## Back-end and Database

Pytest was used to create unit tests that test various components of both the back-end and database.

Callback functions were tested by providing all possible input, and ensuring that no errors occur and that a Plotly object is successfully returned.

The first and last rows of each table in the database was tested to ensure integrity of the data.

## Front-end

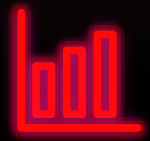
Gremlin.js was used for monkey testing the front-end.

Monkey testing involves providing a web-app with thousands of random inputs, attempting to trigger vulnerabilities or to break components. Results were optimal.

# Functional Requirements

## DATA CONVERSION

Dataset must be transformed to a multi-table relational database.

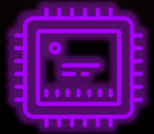


## DATA VISUALIZATION

All data present in the dataset must be visualized intuitively.

## ANALYTICS TOOLS

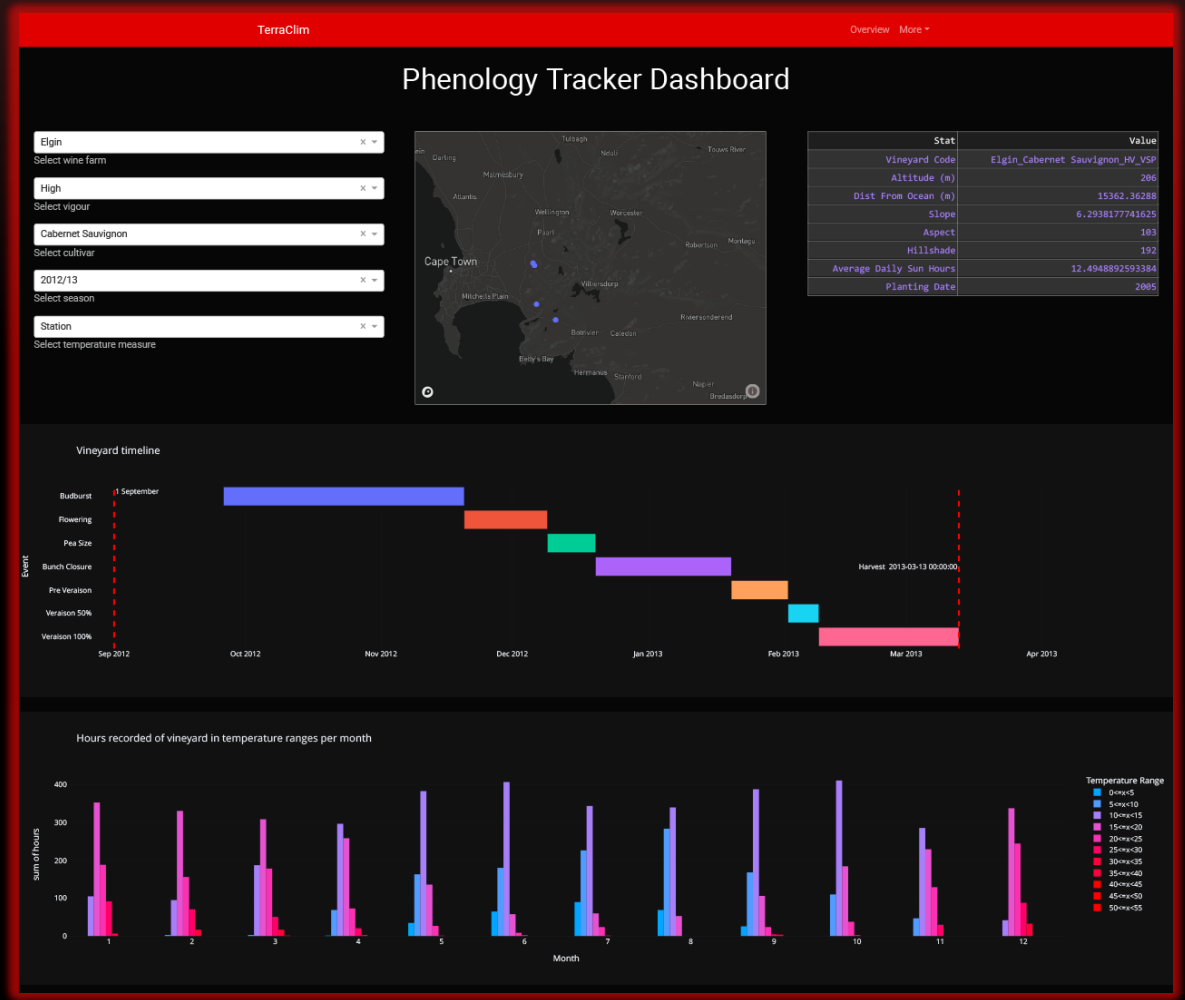
Analytics tools must be provided to further compare data.



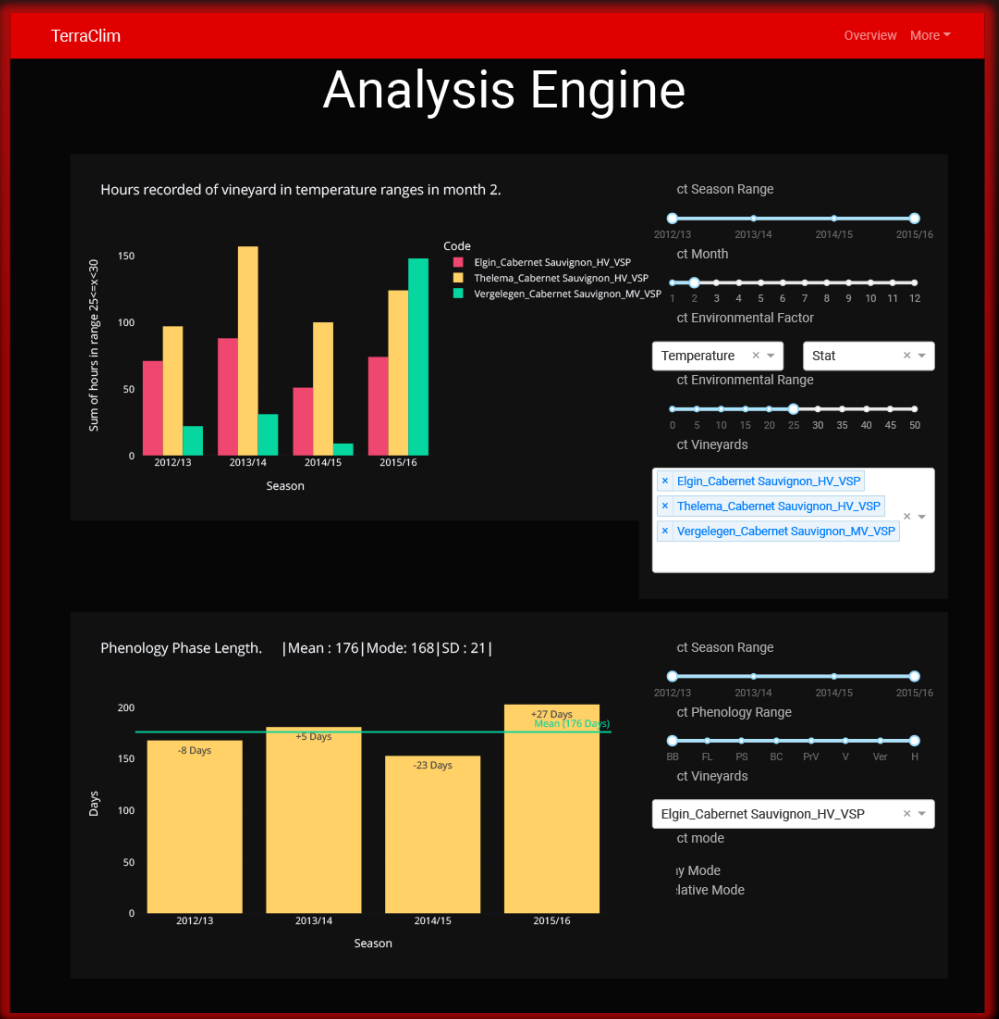
## HARVEST PREDICTION

Harvest prediction must be attempted, with a variety of models.

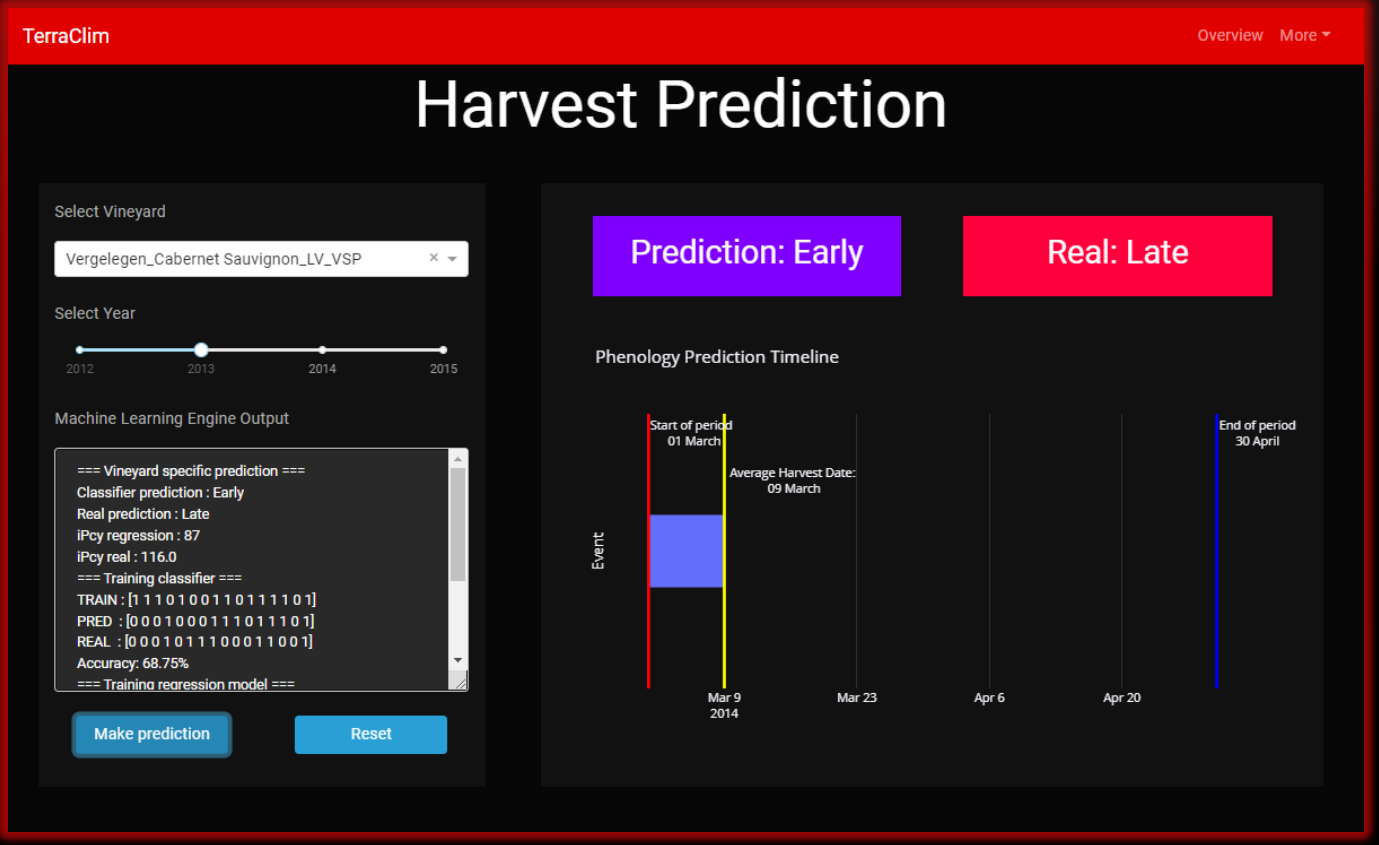
# Phenology Dashboard:



# Analysis Engine:



# Harvest Prediction:



# Future work

**Updating the dataset to contain data up to the present (2022).**

**Live data updates.**

**Train the machine learning algorithms with a completed dataset. Experiment with different models as well as neural networks.**

# THANK YOU

## HOODIES COMING FRIDAY