COMP 4520 – Undergraduate Honours Project Proposal

Finding a Machine Learning Model that Best Predicts Canola Traits.

Student: Mitul Patel (7851781)

Supervisor: Dr Michael Domaratzki (Mike)

i. Abstract:

This project explores building various Machine Learning Models using a dataset from the Bioinformatics Lab at the University of Manitoba and algorithms from the skikit-learn library to find which Machine Learning model best predicts Canola Crop Traits. These traits include flowering time, maturity time, height, yield, protein content, oil content and glucose content.

ii. Introduction to the topic:

Canola is proudly grown in Canada and is now used as a source of oil. At the Faculty of Agriculture, University of Manitoba, Scientists have planted about 460 Canola seeds each with a different Genetic Sequence. They then recorded the Genetic Sequence of each seed and its corresponding traits. We will use this recorded data to build Machine learning models and find the best Model that will in turn be used to predict Canola crop traits. As a result, researchers can breed new varieties of Canola and easily predict and understand their traits without having to plant them. Furthermore, using these predictions, scientists can quickly find the Genetic Sequence that gives the most favourable traits and thus those seeds can be used by Canola farmers to maximize their revenue.

iii. Your background preparation to do this project:

I have prepared for this project in the following ways

- I learned the programming language Python from the COMP 1012 course.
- I also have a high overview of how some supervised machine learning algorithms work e.g. Random Forest and Linear Regression. I came across them and used them on my research paper (Assignment 4) in COMP 3190.
- I also have a general overview of some other supervised algorithms e.g. Decision Trees and Logistic Regression as it was part of the AWS Machine Learning Certification Course on Linux Academy that I have taken recently.

iv. Related work:

There are similar projects undertaken at the University of Manitoba Bioinformatics Lab, specifically, the published journal *Sparse Bayesian Learning for Predicting Phenotypes and Ranking Influential Markers in Yeast*. It aims at predicting the yield and drought resistance in crops while doing Genomic Selection.

v. Problem Statement:

For all the different Genetic Sequences for Canola Seeds, it's expensive, and time consuming to plant and grow each one in order to know the resulting Crop Trait. An accurate Machine Learning Model would help eliminate seeds with a Genetic Sequence whose predicted Crop Traits are extremely unfavourable thus scientists will not need to plant those and instead plant and test seeds whose predictions give favourable Crop Traits. In addition, different Machine Learning Models work in a different way and must build and test various models to find the best one.

vi. Methodology:

TASKS TO BE COMPLETED	DUE BY
SVM • Process Data • Train • Test • Build a Confusion Matrix, Find F1, Improve and Infer.	June 15th
Decision Trees • Train • Test • Build a Confusion Matrix, Find F1, Improve and Infer.	July 2nd
Random Forest Model • Train • Test • Build a Confusion Matrix, Find F1, Improve and Infer.	July 23rd
Logistic Regression • Train • Test • Build a Confusion Matrix, Find F1, Improve and Infer.	August 13th
Presentation and Report	August 17th - August 21st

vii. Infrastructure and facilities and expert personnel required:

I will be working directly with my supervisor Dr Michael Domaratzki (Mike) as I appreciate his expertise in Bioinformatics and will seek guidance where necessary. We will arrange Skype meetings and communicate remotely.

viii. Outcome and Deliverables:

I am expected to deliver:

- **Dr Michael Domaratzki (Mike):** Weekly meetings, a Final Paper and a Presentation discussing what was accomplished for the Honors Project.

- **Dr Ruppa Thulasiram (Tulsi):** A proposal document at the start of the project, bi-weekly check-ins on the progress of the project, a Final Paper and a Presentation discussing what was accomplished for the Honors Project.

References:

Ayat, Maryam & Domaratzki, Michael. (2018). Sparse Bayesian learning for predicting phenotypes and ranking influential markers in yeast. 10.1101/489245.