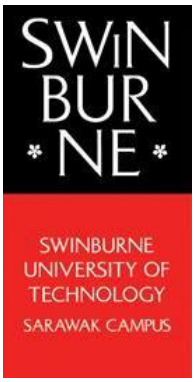

Swinburne University of Technology Hawthorn Campus
Dept. of Computing Technologies

COS10022 Data Science Principles

Assignment 1 - *Semester 1, 2024*

Assessment Title: Predictive Model Creation and Evaluation

Assessment Weighting: 20%

Due Date: Sunday, 24th March 2024 at 11.59 pm (AEDT)

Assessable Item:

- One (1) piece of a written report no more than 10-page long with the signed Assignment Cover Sheet.
- The submitted report must be checked by Turnitin, and the similarity from **not the template part** should be less than 12%.

The submitted report should answer all questions listed in the assignment task section in sequence.

You must include a digitally signed Assignment Cover Sheet with your submission.

Purpose of Assignment

This assignment aims to evaluate students' achievement of the following unit learning outcomes:

1. **Explain the key concepts, techniques, and tools for handling the data and creating prediction models.**
2. **Work on feature and model selection and implementation in a data science project.**

This is an individual assignment that requires peer review and communication with colleagues. Refer to the Unit Outline for the late submission penalty policy. You can ignore the high similarity on the cover page and the template wording, but not in your report content. You must ensure your submitted report has a similarity lower than 12% in total and less than 6% from a single source. Otherwise, your report will not be marked.

Key Lessons:

You are asked to divide the dataset and then utilise the linear and logistic regressions to build two models in the KNIME analytic platform.

Introduction

The dataset contains 150 tuples of 7 commonly seen fish species in the market. There are 6 attributes included in the source data. We have two goals in this assignment: the first goal is building a linear regression model for predicting the weight of the fish, e.g., the value in the "Weight_of_Fish_in_Gram" attribute; the second one is building a logistic regression model for predicting the species of the fish. You are expected to follow the instructions for building your predictive model and answer questions.

Assignment Goal

This assignment aims to build experiences for students to select independent attributes, split the data into training and test sets, train a usable predictive model, and explain the outputs. A small part of the discovery and research component is included in the assignment to expand the students' skill set.

Assignment Task

The dataset has been cleaned and organised with no missing data. Your tasks are to select the proper attributes and to create the predictive models according to the instructions for answering the questions listed below. The source file is "**Fish_Species_2024.csv**". The report should be prepared with the template and answer the questions, followed by finding the required information, splitting the dataset, model training and testing. A table of Content is not required.

Data Preparation (10%)

You must follow the instructions to split the given data set into training and test sets. Remember, a well-split dataset is the foundation of support for the model training and test. You are required to use a **Shuffle node** with **9214** as the seed value to shuffle the input data. Moreover, you need to partition 80% of the input data in the training set by the "draw randomly" method with **9214** as the seed value.

Linear Regression (40%)

The data source contains many details of the record. Our goal is to build a predictive model for predicting the weight of the fish. The weight of the fish is recorded in the attribute "Weight_of_Fish_in_Gram" in the given file. Your mission is to create a linear regression model in KNIME and visualise the prediction result.

Logistic Regression (40%)

Using the same source file, we aim to build a predictive model for classifying the input fish into the corresponding species.

Performance Improvement (10%) – This part can be optional

Having a linear regression model created is simple. How to improve the accuracy of the prediction result requires a bit more effort. Let's focus on a single species of fish - Perch. If you are limited to selecting three (3) attributes as the input for your linear regression model only, find a way to decide which attributes should be included. Note that when building the linear regression model, you must ensure that tuples in the new training and test sets are fully the subset of the original training and test sets.

Important Note

You must use the seed value specified in the instructions. Otherwise, you will get different results than the correct answer in almost all questions.

There are 100 marks on this assignment. Your proposal must address the following tasks.

1. Follow the instructions above to split the source data into training and test sets. Answer the following questions after splitting the data. **[10 marks in total]**
    1) Submit the workflow of Assignment 1 via Assignment 1.1. **[2.5 marks]**
    2) How many tuples are included in the training set? **[2.5 marks]**
    3) How many species are included in the test set? **[2.5 marks]**
    4) Do species "Whitefish" and "Smelt" have the same number of tuples included in the test set? **[2.5 marks]**

2. Build a Linear Regression Model using **all** available attributes to predict the value of the "Weight_of_Fish_in_Gram". Answer the following questions after completing the model training and test. **[40 marks in total]**
    1) What is the R^2 value of your test result? **[5 marks]**
    2) Give the screenshot of the scatter plot result of your test output using "Weight_of_Fish_in_Gram" on the x-axis and the prediction value on the y-axis. Assign different colours to the data points based on the "species." **[15 marks]**
    3) Which species has the heaviest predicted weight in your test result? **[5 marks]**
    4) How many prediction results are infeasible in your test result? **[5 marks]**
    5) Looking at your source data before splitting them, which two species can be easily separated from others if looking at the "Height_in_cm" and "Diagonal_Width_in_cm" attributes? Post your visualisation result on data observation in the report. **[5 marks]**
    6) Draw a doughnut chart of the original input data with 0.55 as the doughnut hole ratio before splitting it into training and test sets. Use different colours for each species and show the percentage of data in the pie chart. **[5 marks]**
3. Build a Logistic Regression Model with **all** attributes and use "Smelt" as the reference category. The maximal number of epochs and epsilon should be set to **10,000** and **0.00001**, respectively. Use "LineSearch" as the learning rate strategy. Use **9214** as the seed in the logistic regression node. Answer the following questions after completing the model training and test. **[40 marks in total]**
    1) Which species have/has no "True Positive (TP)" case in the prediction result? **[5 marks]**
    2) For the species with no TP case, which species will be misplaced? **[5 marks]**
    3) What is the overall accuracy of the prediction result? **[5 marks]**
    4) List all species names with 100% correctly classified test results. **[15 marks]**
    5) Which species has a 33.33% chance of being misplaced into another species in the test result? **[5 marks]**
    6) In the test result, how much percentage of the species "Perch" is misplaced into others? **[5 marks]**
4. Build a new linear regression model different from the one built when answering question 2. This time, let's focus on the species "Perch" only. You are limited to using three attributes in the input to predict the "Weight_of_Fish_in_Gram." Use a "Scatter Matrix (local)" node to observe your data and decide the suitable attributes to be included. The linear regression model should be the same as the one used in question 2 except for the input attributes. Build, train, and test the model and then answer the questions below. **[10 marks in total]**
    1) Give the reasons for each eliminated attribute and why they are not selected as the input. **[5 marks]**
    2) List the R^2 of your test result and compare it with the one in question 2. Reveal both R^2 values obtained in question 2 and in question 4. If you can improve the model, you get the mark. **[5 marks]**

Submission Requirement

To fulfil the requirement of this assignment, the submission should be prepared in MS Word or PDF format, named ***COS10022_[Student_ID]_Assignment_1*** and submitted. Replace the ***[Student_ID]*** with your student ID number.

Failure to adhere to the submission requirements will immediately result in losing marks for this assignment.

---------- End of Assignment ----------