THE UNIVERSITY OF SYDNEY

Project Report for ENGG2112

Amazing Title of project (which you should keep to within two lines)

Bob Smith, 510670000, Electrical Engineering
Yuan Lin, 480421111, Biomedical Engineering
David Mahendran, 49052222, Software Engineering

Faculty of Engineering

October 17, 2022

Executive Summary

Put your executive summary here. Its purpose is to summarize the objectives, methodology, main findings and conclusions of the project. Its intended audience is a reader who does not have enough time to read the main body of the report carefully, usually a busy executive, hence the name.

Contents

1 Background and Motivation

This project is focused on the use of machine learning to predict indigestion among university students, using a feature set consisting of (among others) gender, age, course enrolled in, blood pressure and number of hours spent on study per semester week. It attempts to explain the curious and widely observed phenomenon of students reporting indigestion during the exam fortnight, and particularly during the very days of their exams. Our results indicate that the problem is not easily solved and requires further research. It is a sufficiently interesting problem that our project team has managed to secure a promise of $50,000 from the Deputy Vice Chancellor Education office to study it more deeply. This is an indication of the substantial impact the successful continuation of our project will have on the university sector in Australia at large.

In our project, our data set was obtained from DataSets-R-Us.com, a reputable open source data repository. The project team discussed at length the choice of data to work on in this project, and settled on one that we knew would resonate with all of us. The data was collected from the University of South Antarctica, through a poll of its 200 students across all academic departments, and had a total of 25 features or independent variables.

2 Objectives and Problem Statement

The main objective was to use some or all of the available features to find the probability of a student having indigestion sometime during the examination period, as well as on the day of any scheduled examination. Some of the features were clearly relevant to the classification problem, but others were less so and therefore a secondary objective was to extract the minimal set of features that would yield satisfactory classification accuracy.

Denoting the features as x_i, $i = 1, \ldots, 25$, and the target variables y and z as indicator functions of whether a student has indigestion sometime in the exam period or on the day of their scheduled exams, the objective is to find the functions $f_y(x_1, \ldots, x_{25})$ and $f_z(x_1, \ldots, x_{25})$ that represent the probability of the two respective events of interest for the set of observed features. Implicit within these function definitions is that not all of the input variables will need to be used, i.e. the function may ignore certain input variables, depending on how relevant they are to the classification problem.

By comparing f_y and f_z to a user-defined threshold, we will then be able to make binary decisions on whether indigestion will occur.

3 Methodology

3.1 Data Pre-Processing

As some of the independent variables are categorical in nature, we first need to use one-hot encoding to convert them to (binary) numerical variables. There are also a number of missing data values, which we replace with zero where appropriate[1] and with the average value of that variable calculated using the non-empty values otherwise. These were implemented using the sklearn module's functions, as follows:

1. onehotencode() for one-hot encoding;
2. dataclean() for zeroing missing data;
3. average() for averaging.

3.2 Feature Extraction

For feature extraction, we used several of the sklearn module's functions designed for this purpose, including feature1() and feature2(). These work on the principle of correlating each feature or set of features against the target variable(s), with different ways of expressing the statistical correlation. We chose to the parameters of these modules as follows: (a) randomly, (b) linearly increasing for 0 to 100, (c) linearly decreasing from 20 to 0.

3.3 Classification

As the problem is one of binary classification, we tried the Thing One and Thing Two methods, as covered in class. These methods required careful tuning of their hyper-parameters for optimal performance, which we did systematically. As is widely known, a machine learning classifier's performance can vary extremely widely over a range of hyper-parameter values. We encountered this phenomenon first hand, as seen for instance in Figure 1.

The Thing Three method was also considered, but initial simulations showed that it was non-trivial to adapt it to our problem within the time we had, and therefore it was abandoned.

3.4 Simulation Environment

The simulations made use of a 80-20 training-testing split, and were run on Google Colab through Python notebooks. Random cross-validation of ten instances was performed, with the final model chosen through averaging of the three most accurate models. Accuracy and area under the curve (AUC) were the main performance measures employed.

[1] These are those variables which are blank because their true values are zero.

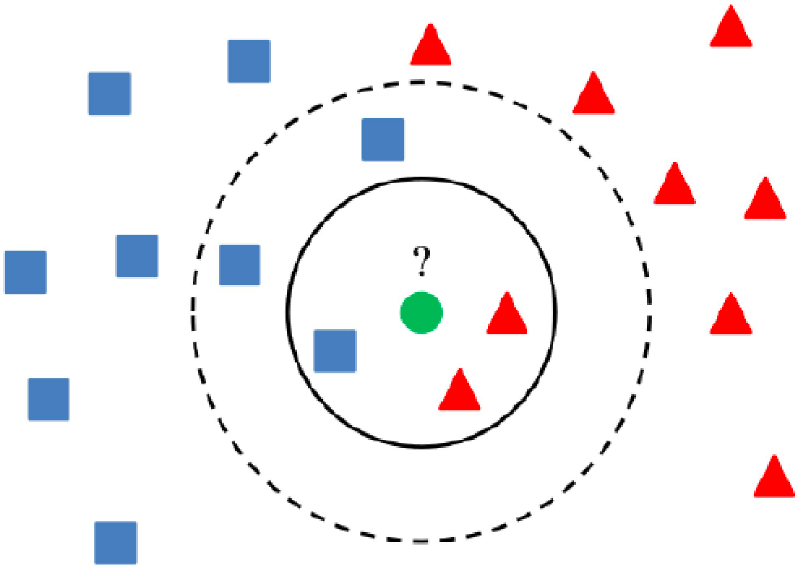

Figure 1: Example of a figure embedded in the text.

4 Simulation Results

4.1 Key Findings and Significance

In this section, the key simulation results and findings are presented concisely. A literature review of related work is also provided, with our results compared against some of this work. It is important to note that our project team only had six weeks to complete the project, and improvements to the methods used could surely have been found had there been more time.

In [1], Thing Two was employed to solve a similar problem, i.e. finding the probability of developing sore eyes after 10 hours of gaming, given a set of attributes of the subject consisting mainly of their vital statistics and health condition. In [2], Thing One and Thing Two were used in a novel combination to solve another biomedical problem, diagnosing the presence of cancer cells in a magnetic resonance imaging (MRI) picture after extracting certain image features. Our novel use of Thing One on its own, with parameters tuned using Method A, proved to be highly competitive against these methods. This is demonstrated in the set of performance curves shown in Figure 2.

4.2 Issues Faced

The coding of the simulations was not entirely problem-free. We faced the following issues in chronological order, and dealt with them as described.

- Problem One:
- Problem Two:

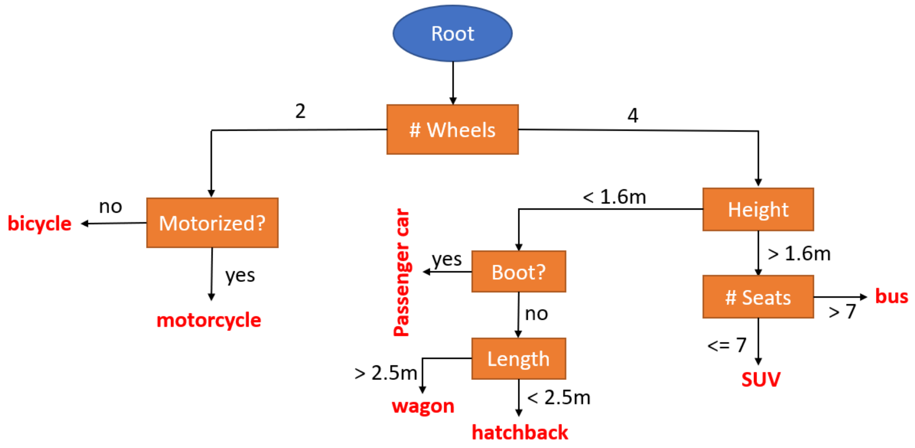

Figure 2: Another example of a figure embedded in the text.

5 Potential for Wider Adoption

It is encouraging that this short project has produced results that appear to be competitive against the methods proposed by highly regarded research groups. We envisage that future work might include:

- Improvement One
- Improvement Two
- Adjustment Three

The interest from industry in this technology is strong, as evidenced by a recent market study conducted by Deloitte [3]. Currently, commercial software that is suitable for solving the problem tackled in this project is not available, to our knowledge (after a thorough Internet search). We believe that, after overcoming the issues raised in an earlier section using the methods discussed above, we can build a prototype that can be demonstrated to potential investors. These would include government departments of education, colleges and universities.

6 Conclusions

The project was completed with some degree of success. We managed to do this, that and the other, though not everything in the proposal workplan was completed. The work had to be modified due to the various issues mentioned in Section 4.2. The team worked well together, meeting for at least 2 hours per week, and every team member contributed their fair share of time and effort. The main findings were as follows:

1. Finding One
2. Finding Two
3. Finding Three

The project can be expanded by capturing more data, and developing more sophisticated methods of classification that could surpass the best result obtained in this project, an accuracy of 85 percent. With an accuracy that approached 95 percent, one could think of commercializing the results either through starting a company or licensing the technology to a company.

References

[1] A. Einstein, "Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]," *Annalen der Physik*, vol. 322, no. 10, pp. 891–921, 1905.

[2] D. Knuth, "Knuth: Computers and typesetting."

[3] M. Goossens, F. Mittelbach, and A. Samarin, *The LaTeX Companion.* Reading, Massachusetts: Addison-Wesley, 1993.