SKIH2103 (DATA ANALYTICS) – A232 Group Project

Instructor: Associate Professor Dr. Azizi Ab Aziz
Submission date: 18th July 2024 (before 11.59 pm)
via UUM Learning Portal

"The art of knowing is knowing what to ignore." [Rumi]

The Magic of the Sorting Hat - A Journey from Hogwarts to Data Analytics

Imagine a world where a simple hat can peer into your soul and determine your destiny. Welcome to the enchanting world of Harry Potter, where the Sorting Hat plays a pivotal role in shaping young wizards' futures at Hogwarts School of Witchcraft and Wizardry.

In the Harry Potter universe, the Sorting Hat is more than just an accessory; it's a relic imbued with the wisdom of the four Hogwarts founders.

It assesses students based on their personalities, traits, and potential, making an almost instantaneous decision on their house placement. This ancient, sentient hat assigns new students to one of the four houses: Gryffindor, Hufflepuff, Ravenclaw, or Slytherin, each with its unique traits and values. Each house has distinct characteristics and values:

- Gryffindor: Courage, bravery, and determination.
- Hufflepuff: Hard work, loyalty, and fair play.
- Ravenclaw: Intelligence, knowledge, and wit.
- Slytherin: Ambition, cunning, and resourcefulness.

Objective

We aim to create a data analytics model that emulates the Sorting Hat's sorting process for UUM students. This involves:

- Data Collection: Gathering data on student characteristics.
- Feature Engineering: Identifying and processing relevant features.
- Model Selection: Choosing appropriate machine learning algorithms.
- Training and Testing: Training the model and evaluating its performance.
- Prediction: Using the model to predict house assignments for new students.

Data description:

Variable	Description
Personality Traits	The Big Five personality test is a widely accepted personality test theory, also known as the OCEAN personality test, is based on the Big Five model that defines human personality as the combination of five personality traits or factors: • Openness • Conscientiousness • Agreeableness • Extraversion • Neuroticism The Big Five personality test is based on a continuum wherein individuals are ranked on a scale between two extreme ends. • For data collection purposes, you can use a simple Big Five Personality test or OCEAN test: https://www.idrlabs.com/short-big-five/test.php • For application development purposes, you can use this material as a reference for encoding purposes. o brief-big-five-personality-inventory.pdf o The Big Five Inventory-calculation.pdf
Behavioural Traits	The characteristics that consistently describe a person's behaviour: • Risk-Taking: Assesses the student's willingness to take risks and try new things. • Collaboration: Measures the student's ability to work effectively with others. • Discipline: Evaluates the student's self-control and adherence to rules and schedules. • Independence: Gauges the student's ability to work autonomously and make decisions on their own.
Hobbies	Types of hobbies: • Physical (e.g., active hobbies like dancing, yoga, hiking, sports, gardening, martial arts, singing)

	• Cerebral (e.g., activities like sudoku, reading, and puzzles can help another part of our minds by activating our concentration) • Creative (e.g., activities like writing, painting, singing, or cooking may provide a sense of accomplishment) • Community activities (e.g., volunteering, tutoring, helping people) • Collecting (e.g., coin /stamp collectors) • Making & Tinkering (e.g., self-motivated projects like building new things, self-restoration, and repairing stuff)
Academic Performance	Current GPA (grade point average)
Hometown	The city or town where one was born or grew up; also the place of one's principal residence
INASIS	Residential college
Co-curriculum Activities	Chosen co-curriculum for the academic programme
Leadership	(Yes / No) – either a person who used to hold/ currently holding a main position in a club / Inasis / academic school
Favourite Cuisine	Malay, Chinese, Indian, Western, Japanese, Korean, Thai, Exotic
Family Income	Estimated income/salary (parents/guardian)
Number of Best Friends on Campus	Numbers of your good friends at UUM (best friend forever!)
Academic Programmes at UUM	Your enrolled undergraduate / graduate programme
Hogwarts House	Questionnaires: https://www.theguardian.com/childrens-books-site/quiz/2015/feb/05/harry-potter-night-quiz-sorting-hat-which-house

INSTRUCTIONS:

In general, you are required to:

- Form a group of 2-3 people (Note: You can do this alone too).
- Collect your data based on the described attributes/list of questionaries.
- Create a database/spreadsheet that contains related attributes and a target.
- Describe your data (attributes/features) through suitable visualisation approaches.
- Identify possible missing values (if any, you need to explain how to overcome this issue)
- You must analyse your data using at least **THREE (3)** data analytics methods.
- Evaluate your results based on appropriate methods (e.g., Confusion Matrix /ROC-AUC)
- Perform some experiments to obtain the best classification results (at least 80 per cent)
- Your solution should have a working prototype with adequate graphical user interfaces.
- The new classification result should be chosen based on majority voting strategy.
- Conclude your findings based on your experimental results.
- Submission materials:
    a. Report
    b. Datasets
    c. A working source code (without any errors) / execution file (if any)

IMPORTANT QUESTIONS:

These questions give you a critical checklist to ensure the correctness of your deployed solution.

Before Starting the Data Analytics Project Checklist

- What question are you asking/answering, and for whom?
- What data are you using?
- What techniques are you going to try?
- How will you evaluate your methods and results?
- What do you expect the result to be?

What Techniques Are You Going to Try?

- What methods/techniques should I use?
- Why do I think these are the correct methods/techniques for this problem and data set?
- Are there similar projects/references/papers that have already done this that I can learn from before I get started?
- Are these techniques that I would want to use/do in a predictive analytic job?

How will you evaluate your methods and results?

- How will I know I did the analysis and project correctly?
- What are critical parts of the project that will tell me that I am doing things incorrectly?
- What numbers/results/insights will I sense check?

What do you expect the result to be?

- What do I expect the result to be?
- Why do I expect the result to be this?
- Does this result match the results/experiences of other people with similar methods and techniques on similar data?

REPORT:

- **Introduction**
  - Why is your task important? Why should one care? What task are you attempting to cover? How are you covering them? Is there a particular technical challenge/problem you attempted to solve?
- **Background**
  - What methods or ideas have you built on? Any background on the domain topic that one might need to know to understand the application?
- **Data**
  - What did you use? How much was there? Were their labels? Include descriptive statistics where helpful. Anything else we should know?
- **Methods**
  - How did you do it? Which methods? Which setting / hyper-parameters? Be sure to make clear how the frameworks and concepts were used.
- **Evaluation/Results**
  - What were your results? How accurate were they? What insights were derived? Did you analyse what sort of mistakes it made? Examples of output? Anything to demonstrate unique aspects of the approach? Be sure to think of tangible ways to present the results. Each figure should have a point it is trying to convey. Your figures/tables should tell a logical story from first to last.
- **Conclusion**
  - Summarise the takeaways.
- **Reflection (1 paragraph)**
  - What did you learn from this project? What do you wish you had known before you started? What would you do differently? What advice would you offer to future students embarking on this project?
- **References**
  - Be sure to cite and add references (at least THREE (3) references to others' work) for any ideas, data, or tools you are using or building from. **All figures, quotes, or rephrases from articles, websites, and research papers (anyone else) should be cited, or you may receive a 0 for plagiarism.**
- **Format**: refer to the Springer-Nature article format (format_report.zip).

GRADING RUBRICS:

Your final project will be graded as the following:

- **Task definition**: is the task precisely defined, and does the formulation make sense?
- **Approach**: was a baseline, an oracle, and an advanced method described clearly, well justified, and tested?
- **Data and experiments**: have you explained the data clearly, performed systematic experiments, and reported concrete results?
- **Analysis**: did you interpret the results and explain why things worked (or did not work) the way they did? Do you show concrete examples?
- **Extra credit**: does the project present interesting, complicated datasets, programming, and novel ideas (i.e., would this be publishable at a good conference)?

Policy: *All grading of deliverables will be based on standards indicated for each deliverable. Deliverables may not be turned in late and no cheating! For the purposes of this programme cheating will include: plagiarism (using the writings of another without proper citation), copying of another (either current or past student's work), working with another on individually assigned work, or in any other way presenting as one's work that which is not entirely one's own work.*