Text-mining Assignment Submission Cover Sheet

This Assessment Cover Sheet **must** be included on all Assessment submissions.

|  |  |
| --- | --- |
| Assignment Title |  |
| Module |  |
| Student Name  (same as Student Card) |  |
| Student Number |  |
| Programme |  |
| Part-Time/Full-Time |  |
| Year of Study  (First Year, Second Year, etc) |  |

Late Submissions: Assessment submitted after the deadline will have a late penalty applied.

**Academic Integrity for assessment in TU Dublin Programmes**

Each student is responsible for knowing and abiding by TU Dublin Academic Regulations and Policies. Any student in breach of these regulation/policies will be subject to action in accordance with the University’s procedures for breaches of assessment regulations. Please refer to the General Assessment Regulations at

<https://tudublin.libguides.com/c.php?g=674049&p=4794713>

<https://www.tudublinsu.ie/advice/exams/breachesofregulations/>

All students are expected to complete their courses/programmes in compliance with University regulations. No student shall engage in any activity that involves attempting to receive a grade by means other than honest effort, for example:

1. No student shall complete, in part or in total, any examination or assessment for another person.
2. No student shall knowingly allow any examination or assessment to be completed, in part or in total, for themselves by another person.
3. No student shall plagiarise or copy the work of another and submit it as their own work.
4. No student shall falsify any data. Falsification is the invention of data, its alteration, its copying from any other source, or otherwise obtaining it by unfair means, or inventing quotations and/or references.
5. No student shall use aids or devices excluded by the lecturer in undertaking course work or assessments/ examinations.
6. No student shall knowingly procure, provide, or accept any materials that contain questions or answers to any examination or assessment to be given at a subsequent time.
7. No student shall provide their assignments, in part or in total, to any other student in current or future classes of this module/ programme unless authorised to do so by the lecturer.
8. No student shall submit substantially the same material in more than one module/programme without prior authorization.
9. No student shall alter graded assignments or examinations and then resubmit them for regrading, unless specifically authorised to do so by the lecturer.
10. All programming code and documentation, unless correctly referenced, submitted for assessment or existing in the student’s computer accounts must be the students’ original work or material specifically authorized by the lecturer.
11. Collaborating with other students to develop, complete or correct course work is limited to activities explicitly authorized by the lecturer.
12. For all group assignments, each member of the group is responsible for the academic integrity of the entire submission. Consequently, all group members must satisfy themselves that all elements of their submission adhere to the academic integrity statement points above.

By submitting coursework, either physically or electronically, you are confirming that it is your own work (or, in the case of a group submission, that it is the result of joint work undertaken by members of the group that you represent) and that you have read and understand the University’s Regulations and Policies covering Academic Integrity (see General Assessment Regulations)*.*

Coursework may be submitted to an electronic detection system in order to help ascertain if any plagiarised material is present. If you have queries about what constitutes plagiarism, please speak to your lecturer.

|  |  |
| --- | --- |
| Student Signature |  |
| Date |  |

IMPORTANT:

* Sections 1-8 should be no longer than 15 pages (minimum 10 pages), including diagrams, images, screen captures, tables, etc. Careful selection of these is needed.
  + Code does not count to this total. Code should be added to the relevant section.
  + Detailed discussion is expected. Marks are awarded based on depth of information given.
  + You can add as many references as needed. They are not included in the limit of pages.

Student Details for Group Submission:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Student Name  (as per Student Card) | Student Number | Prog Code | Stream  (DS, ASD, etc) | Year  (1, 2, …) |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

For group submission, all students must insure they are compliant Plagiarism policy on cover page.

Cover page to be completed by one student. For all students in group, Submission of an Assessment, either physically or electronically, with or without this cover sheet, acknowledges your compliance with the TU Dublin Academic Regulations and Policies.

1. **Definition of Problem**

Producing new drugs is expensive, a study in 2020 estimated that the median cost of getting a new drug into the market was $985 million (source <https://en.wikipedia.org/wiki/Cost_of_drug_development>).

Because of this, while in development and test, it may be relevant to consider not only the effectiveness of the drug but also have an idea of how well rated this drug would be by the people that will use it. Here we will look at many attributes of the Drug Review Dataset (source <https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29#>), and try to understand better what makes a drug be well rated by its users.

1. **Data Exploration & Descriptive Analytics**

Our target variable here will be rating, as we can see on Plot 1, this is a left skewed distribution with values between 1 and 10. Although it looks more like a left skewed distribution, It has a somewhat large concentration of rating 1 showing that it’s more common to either give a rating or 7 or above or a 1 for a medication.

Chart, histogram

Description automatically generated

Apart from rating, for our analysis we will consider the following variables:

- effectiveness: categorical variable with 5 levels that goes from Highly Effective to Ineffective.

- sideEffects: categorical variable with 5 levels that goes from Extremely Severe to No Side Effects.

- benefitsReview: a free text field where patients put a review of the medication benefits. We will apply a sentiment analysis to understand if this is positive or negative review and create a new field called benefitsReviewSentiment for our analysis.

- sideEffectsReview: a free text field where patients put a review of the medication side effectgs. We will apply a sentiment analysis to understand if this is positive or negative review and create a new field called sideEffectsReviewSentiment for our analysis.

- commentsReview: a free text field where patients put a review of the medication in general. We will apply a sentiment analysis to understand if this is positive or negative review and create a new field called commentsReviewSentiment for our analysis.

As we will look at overall rating of medications and not each individual drug, we will not consider on this analysis the urlDrugName, condition will also not be considered as it is a categorical variable with nearly as many different levels as urlDrugName and, unless somewhat grouped, wouldn’t help on this specific analysis. Also, we will not consider the first unnamed column since it doesn’t have a description of what it is on the datasource website.

1. **Data Preparation**

Include details of any data cleaning, transformations, data enrichment, feature engineering, feature reduction, etc

We’ve transformed effectiveness and sideEffects into categorical variables.

We’ve also checked for NAs, which revealed a very small quantity of NA values for sideEffectsReview 0.06% of its rows and commentsReview 0.26% of its rows. As we will not use this variables directly but do a sentiment analysis on them, and the number is small, we will just consider those missing values as neutral.

.

Sentiment analysis TODO

1. **Details of Algorithms & Configurations**
2. **Model Performance Metrics & Evaluation of Results**
3. **Identification of the most important variables**
4. **Comparison with other Research & Reflections**

Compare your results to at least three other researchers (maximum of five) who used the same data set. What lessons did you learning from doing this? How can your work be improved? Did you include any improvements in your work and what impact did it have?

1. **Action plan to improve equality and other social issues based on your data**
2. **References**

Use one of the commonly used References and Citation formats.