**Project 2 – Decision Trees, Linear Regression, Model Trees, Regression Trees**

**CS548 / BCB503 / CS583 Knowledge Discovery and Data Mining - Fall 2019**

**Prof. Carolina Ruiz**

**Students:** <replace this with your names in alphabetical order by last name>

|  |  |  |
| --- | --- | --- |
|  | **Classification** | **Regression** |
| **Dataset :**   * Dataset Description * Data Exploration * Initial Data Preprocessing (if any) | /05  /10  /05 | |
| **Code Description:** Python libraries and functions, and/or your own code | /10 | /20 |
| **Experiments:**   * Guiding Questions | /10 | /10 |
| * Sufficient & coherent set of experiments | /10 | /10 |
| * Objectives, Parameters, Additional Pre/Post-processing | /10 | /10 |
| * Presentation of results | /10 | /10 |
| * Analysis of individual experiments’ results | /10 | /10 |
| Summary of Results, Analysis, Discussion, and Visualizations | /20 | /20 |
| Advanced Topic | /30 | |
| Total Written Report | /220 = /100 | |

**Dataset Description, Exploration, and Initial Preprocessing: (at most 1 page)**

**[05 points] Dataset Description: (e.g., dataset domain, number of instances, number of attributes, distribution of target attribute, % missing values, …)**

**[10 points] Data Exploration: (e.g., comments on interesting or salient aspects of the dataset, visualizations, correlation, issues with the data, …)**

**[05 points] Initial data preprocessing, if any, based on data exploration findings: (e.g., removing IDs, strings, …)**

Specify whether the preprocessing was used for the classification task, the regression task or both.

*Note: Keep initial data preprocessing to a minimum – just the necessary – and instead experiment with data preprocessing in the following sections.*

**Code Description: Python Libraries and Functions you used and with what parameters. If you wrote your own Python code, describe it here. (at most 2 pages)**

**[10 points] Classification Techniques:**

**Preprocessing Techniques for Classification:**

**“Zero-R”:**

**“One-R”:**

**Decision Trees:**

**Random Forests:**

**[20 points] Regression Techniques:**

**Preprocessing Techniques for Regression:**

**Linear Regression:**

**Regression Trees:**

**Model Trees [10 points]:** Describe the code you implemented to construct model trees.

**Guiding Questions**

**[10 points] Three Guiding Questions for the Classification Experiments: (at most 1/3 page)**

1. **…**
2. **…**
3. **…**

**[10 points] Three Guiding Questions for the Regression Experiments: (at most 1/3 page)**

1. **…**
2. **…**
3. **…**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[40 points] Summary of Classification Experiments Use 10-fold cross-validation** *At most 1 page.* | | | | | | | | | |
| **Tech.** | **Guiding**  **questions** | **Pre-process** | **Parameters** | **Post-process &**  **Pruning** | **Accuracy,** **Precision, Recall, ROC Area** | **Time to build model** | **Size of model** | **Analysis & observations about experiment, and interesting patterns in the trees** | **You can add other columns**  **if you wish** |
| ZeroR?  OneR?  DTs?  RFs? | 1? 2? 3? |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |

**[20 points] Summary of Python Classification Results, Analysis, Discussion, and Visualizations (at most 1 page)** 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the dataset domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include (a part of) the best classification model obtained.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[40 points] Summary of Regression Experiments. Use 10-fold cross-validation.** *At most 1 page.* | | | | | | | | | |
| **Tech.** | **Guiding**  **questions** | **Pre-process** | **Parameters** | **Post-**  **process**  **& Pruning** | **Correlation**  **Coefficient**  **and Error Metric(s)** | **Time to build model** | **Size of model** | **Salient observations about experiment; and Interesting patterns in the model** | **You can add other columns** |
| ZeroR?  LR?  RT?  MTs? | 1? 2? 3? |  |  |  | Specify  what  metric(s)  you use here |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |

**[20 points] Summary of Python Regression Results, Analysis, Discussion, and Visualizations (at most 1 page)** 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the dataset domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include (a part of) the best regression model obtained.

**Advanced Topic (AT MOST 1 PAGE): <include name of the topic here>**

**[7 points] List of sources/books/papers used for this topic (include URLs if available).** *Provide full references (authors, title, where published, year, …).*

* …
* …
* …

...

**[20 points] In your own words, provide an in-depth, yet concise, description of your chosen topic. Make sure to cover all relevant data mining aspects of your topic.** *Your description here should be comprehensive and in-depth (it should reflect work at the graduate level).*

**[3 points] Describe how this topic relates to trees and the material covered in this course.**

**Authorship:** Although each student on the team is expected to be involved in every aspect of the project, describe in detail here the main contributions that each of the team members made to this project. This authorship description must accurately reflect the work done by each team member, and must be approved by all of the members of the team (at most 1/3 page)