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

**CS548 Knowledge Discovery and Data Mining - Spring 2015**

**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:** | **Weka**  /10 | **Python**  /10 | | **Weka**  /10 | | **Python**  /10 |
| **Experiments:**   * Guiding Questions | /10 | | | /10 | | |
| * Sufficient & coherent set of experiments | /10 | /10 | | /10 | | /10 |
| * Objectives, Parameters, Additional Pre/Post-processing | /10 | /10 | | /10 | | /10 |
| * Presentation of results | /10 | /10 | | /10 | | /10 |
| * Analysis of individual experiments’ results | /10 | /10 | | /10 | | /10 |
| Summary of Results, Analysis, Discussion, and Visualizations | /20 | | | /20 | | |
| Advanced Topic | /30 | | | | | |
| Total Written Report | /310 | | | | | |
| Total Written Report per student | /90 | | /90 | | /90 | |
| Class Presentation per student | /07 | | /07 | | /07 | |
| Class Participation per student | /03 | | /03 | | /03 | |
| Total Project 2 per student | /100 | | /100 | | /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, necessary dimensionality reduction, …)**

**Weka Code Description: Inputs, output, and process followed by Weka’s code to construct the trees (at most 2/3 page)**

**[10 points] J4.8 Code Description:**

**[10 points] M5P Code Description:**

**[20 points] Python Packages and Functions used (decision trees, linear regression, model/regression trees). Describe inputs & outputs (at most 1/3 page)**

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

1. **…**
2. **…**
3. **…**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[40 points] Summary of Classification Experiments in Weka. Use 10-fold cross-validation** *At most 2/3 page.* | | | | | | | | | | |
| **Tech.** | **Guiding**  **questions** | **Pre-process** | **Parameters** | **Post-process &**  **Pruning** | **Accuracy,** **Precision, Recall, ROC Area** | **Time to build model** | **Size of model** | **Interesting patterns in the model** | **Analysis & observations about experiment** | **You can add other columns** |
| ZeroR?  OneR?  J4.8?  …? | 1? 2? 3? |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |  |  |

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

**[20 points] Summary of Weka and Python Classification Results, Analysis, Discussion, and Visualizations (at most 1/3 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 (U.S. population census), and discuss the answers that the experiments provided to your guiding questions. 3. Include (a part of) the best classification model obtained.

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

1. **…**
2. **…**
3. **…**

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

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

**[20 points] Summary of Weka and Python Regression Results, Analysis, Discussion, and Visualizations (at most 1/3 page)** 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the 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: <include name of the topic here>**

**[7 points] List of sources/books/papers used for this topic (include URLs if available):**

* …
* …
* …

...

**[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.**

**[3 points] How does this topic relate 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)