**Project 1: Data Integration, Data Warehousing, Data Pre-processing**

**CS548 Knowledge Discovery and Data Mining – Spring 2016**

**Prof. Carolina Ruiz**

**Students:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |
| --- | --- |
| **Problems** | **Scores** |
| **I. KDD Process (20 points)** | / 20 |
| **II. Data Preprocessing (65 points)** | /65 |
| **III. Feature Selection (60 points)** | /60 |
| **IV. Exploring Real Data (65 points)** | /65 |
| **V. Data Integration, Data Warehousing and OLAP (50 points)** | /50 |
| **TOTAL written report** | /260 = / 100  |

**Problem I. Knowledge Discovery in Databases (20 points) AT MOST 1 PAGE**

1. (5 points) Define knowledge discovery in databases.
2. (10 points) Briefly describe the steps of the knowledge discovery in databases process.
3. (5 points) Define data mining.

**Problem II. Data Preprocessing (65 points) AT MOST 2 PAGES**

1. (5 points) Three approaches for replacing missing values.
2. (5 points) Transforming the OUTLOOK attribute.
3. (5 points) Discretizing TEMPERATURE into 4 equal-width intervals. Show your work.
4. (5 points) Discretizing HUMIDITY into 4 equal-frequency intervals. Show your work.
5. (5 points) Keeping the DATE attribute?
6. (10 points) Discretizing HUMIDITY using new procedure. Show your work.
7. (30 points) Supervised discretization.
	1. (2 points) Results of discretizing TEMPERATURE with Weka.
	2. (20 points) Describe the process followed by the supervised discretization Weka filter code in your own words. Describe the inputs received by the code, the output it produces and the process it follows to do the discretization.
	3. (8 points) Follow the code by hand to show how the TEMPERATURE attribute was discretized.

**Problem III. Feature Selection (60 points) AT MOST 1.5 PAGES**

1. (5 points) Attributes selected by CfsSubsetEval over the weather.arff dataset.
2. Weka code implementing CfsSubsetEval:
	1. (5 points) Initial subset of attributes under consideration. Forward or backward search?
	2. (25 points) Show step by step the process followed by CfsSubsetEval. Use the lattice below to show neatly the order in which it considers the subsets and the goodness value of each of the subsets considered.



* 1. (25 points) Use the CfsSubsetEval formulas to calculate the goodness of the "best" (sub)set of attributes considered. Show your work.

**Problem IV. Exploring Real Data (65 points) AT MOST 3 PAGES**

1. Dataset Exploration
	1. (5 points) Two observations of good things about the dataset and two observations of bad things about the dataset (include visualizations).
	2. Percentiles for the *horsepower* attribute

Histogram of the *horsepower* attribute

* 1. (10 points) Covariance matrix (5 points) and heatmap visualization (5 points)

(10 points) Correlation matrix (5 points) and heatmap visualization (5 points)

(5 points) Two attributes you’d remove, if any, based on these two matrices and why. Explain your answer.

1. (10 points) Dimensionality Reduction.
	1. PCA results in Weka (5 points). How many dimensions are obtained? How much of the variance is explained? First new component/attribute. Interesting observations.
	2. PCA results in Python (5 points) How many dimensions are obtained? How much of the variance is explained? First new component/attribute. Interesting observations.
2. (10 points) Feature Selection.
	1. CFS results in Weka (5 points). Which attributes were selected? Interesting observations.
	2. CFS results in Python (5 points). Which attributes were selected? Interesting observations.
3. (5 points) Attribute Transformation. Modifying car-name.
	1. Difference in PCA results in Weka and Python.
	2. Difference in CFS results in Weka and Python.

**Problem V. Data Integration, Data Warehousing and OLAP (50 points) AT MOST 2.5 PAGES**

1. (10 points) Describe the main differences between the mediation approach and the data warehousing approach for data integration.
2. Data warehouse consisting of the three dimensions *time*, *doctor*, and *patient*, and the two measures *count* and *charge*, where *charge* is the fee that a doctor charges a patient for a visit.
	1. (5 points) Depiction of this data warehouse as a cuboid / multidimensional array.
	2. (5 points) Sequence of specific OLAP operations to go from the base cuboid [day, doctor, patient] to the total fee collected per doctor in 2014. Explain your answer.
3. Data warehouse consisting of the three dimensions MODEL, YEAR, and COLOR and one measure SALE.
	1. (5 points) Depict the data as a multidimensional cuboid.

* 1. (5 points) Depict the result of rolling-up MODEL from individual models to all.
	2. (5 points)Depict the result of drilling-down time from YEAR to month. (Although month data is not provided above, make up a couple of values to illustrate the drill-down operation.)
	3. (5 points)Depict the result of slicing for MODEL=Chevy.
	4. (5 points) Depict the result of dicing for MODEL=Chevy and YEAR=2011.
	5. (5 points) Starting with the basic cuboid model, year, color, sales, what specific OLAP operations should one perform in order to obtain the total number of red cars sold? Make your sequence of operations as efficient as possible.