CS 5402 – Intro to Data Mining
Fall 2020
HW #1

- This assignment is **due by 11:59 p.m. on Friday, Sep, 11, 2020**.
- You are to work on this assignment by yourself. It's ok to discuss general approaches and help one another with technical questions, but your overall work should be your own.
- This assignment is worth **50 points**.

Project Description

For this assignment you are to **preprocess/clean** a dataset. You are only allowed to use **Python and/or Weka methods**; part of the objective of this assignment is to have you practice those methods (as opposed to using Microsoft Excel, R, C++, etc.).

The dataset (**census.csv**), which is posted on Canvas along with this assignment, contains U.S. census data from 1994. There are **32561 instances** which have the following **16 attributes**:

- **Date**
- **Age** (integer)
- **Workclass** (e.g., Private, Self-emp-not-inc, Federal-gov, etc.)
- **Population-wgt** (integer)
- **Education** (e.g., Bachelors, Some-college, 11th, etc.)
- **Education-num** (integer)
- **Marital-status** (e.g., Divorced, Never-married, etc.)
- **Occupation** (e.g., Tech-support, Sales, etc.)
- **Relationship** (e.g., Wife, Husband, etc.)
- **Race** (e.g., White, Other, etc.)
- **Sex** (e.g., Female, Male)

- **Capital-gain** (integer)
- **Capital-loss** (integer)
- **Hours-per-week** (integer)
- **Native-country** (e.g., United-States, England, etc.)
- **Salary** (e.g., >50K, <=50K); this is the decision attribute

Specifically, here are the **only preprocessing/cleaning tasks** that you are to perform:

1. **Date**: make the dates have a consistent format (e.g., MM/DD/YYYY); also, if any date has a year other than 1994, change the year to 1994
2. **Age**: discretize the values into 10 bins using equal width (note: this now makes Age into a nominal attribute)
3. **Workclass**: replace missing values (represented as ?) with Other
4. **Population-wgt**: normalize the values
5. **Occupation**: replace missing values (represented as ?) with Other
6. **Sex**: fix typos (valid values are Male and Female)
7. **Hours-per-week**: discretize the values into 5 bins using equal frequency (note: this now makes Hours-per-week into a nominal attribute)
8. **Native-country**: replace missing values (represented as ?) with Unspecified
9. Perform a **chi-square test** (using 0.05 for significance) between **each pair** of nominal-valued (non-decision) attributes; identify which attributes are not independent of each other by filling in the entries in the table shown below as I=Independent or N=Not independent:

	age	workclass	education	marital-status	occupation	relationship	race	sex	hours-per-week	native-country
age										
workclass										
education										
marital-status										
occupation										
relationship										
race										
sex										
hours-per-week										
native-country										

10. Perform a **Spearman test** between **each pair** of non-nominal (non-decision) attributes; identify which attributes are not independent of each other by filling in the entries in the table shown below as I=Independent or N=Not independent. For the purposes of this assignment, consider the absolute value of correlation coefficient ≥ 0.8 as being "close to 1."

	date	population-wgt	education-num	capital-gain	capital-loss
date					
population-wgt					
education-num					
capital-gain					
capital-loss					

11. Perform a **Principal Components Analysis (PCA)**. Determine the 9 "most important" non-decision attributes according to the PCA results. Provide results (e.g., a vector display, eigenvector values, etc.) that justify your determination.

What To Submit for Grading

You should submit a **zip** file that contains **only two** items:

(1) A single **pdf file** that that **CLEARLY identifies** how you performed **EACH task** (e.g., Python source code, Weka KnowledgeFlow screenshots). Additionally, **provide answers for what you are being asked for in tasks 9-11**.

(2) A **csv file** containing your transformed data.

If your submission contains more than this, we reserve the right to DEDUCT POINTS from your homework score for wasting the grader's time; he has to grade ~75 of these submissions and doesn't have time to wade through extraneous material!

Grading:

Here's how many points each task is worth:

Task	Points Possible
Date: make format consistent	2
Date: make all years be 1994	2
Age: discretize into 10 bins using equal width	2
Workclass: replace missing values with Other	2
Population-wgt: normalize values	2
Occupation: replace missing values with Other	2
Sex: fix typos	4
Hours-per-week: discretize into 5 bins using equal freq	2
Native-country: replace missing values with Unspecified	2
Chi-square test between nominal attributes	14
Spearman test between non-nominal attributes	8
PCA between non-decision attributes	8
Total	**50**