AMA1600 Fundamentals of AI and Data Analytics - Project Instructions

Project instructions:

- Each group consists of 4 members.

- The topic should be chosen from one of the topics provided. Each group can submit a preference of up to three topics.

- Project 11-15 are about classification, Project 21-25 are about regression.

- Project 2X is a free topic on investment. The group choosing this topic should specify the theme of study (e.g. a listed company / a currency / an industry).

- The lecturer will allocate the topics according to your preference, on a first-come-first-served basis.

- Students should submit a project report, and the source code with program output of numerical results and figures.

- The project report should be in word document format (.docx).

- The numerical results and figures should be produced by Python on Jupyter Notebook.

- Plagiarism is not allowed.

- Citation should be included in the last part of the report.

Requirement:

- introduction to the background of the project topic

- assumptions, limitation and methodology of the project

- analytics of data with descriptive statistics

- visualization of data

- explanation of the classification model / linear regression model used

- evaluation of accuracy and efficiency of the model

- interpretation of the result and conclusion

Grading criteria (total 25):

- Concise introduction to the problem and clear explanation (5)

- Appropriate use of quantitative methods and models to produce accurate result (10)

- Communication skills in both tidiness, language and presentation of data (5)

- Insightful, meaningful result and interpretation to the data (5)

Project 11: breast cancer

The Wisconsin breast cancer diagnostics dataset contains measurements of breast tumours of more than 500 patients. Those tumours can be classified into malignant or benign. The goal of this project is to study the relationship between these measurements and set up an efficient way to classify the tumour using machine learning.

Description of the dataset *breast_cancer.csv*:

(1) ID number
(2) Diagnosis (M = malignant, B = benign)

There are ten real-valued features computed for each image of a cell nucleus, namely:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

For column (3) to (32), the mean, standard error and worst (largest) of these ten features are displayed.

Project 12: heart disease

The dataset was collected from Hungary, Switzerland and USA in 1988. It contains 14 attributes about test result of patients with chest pain. The patients are classified into two categories: whether they have heart disease or not. The target of this project is to study these test results, and to train a classification model for determining whether a patient has heart disease or not based on his/her test result and health features.

Description of the dataset *heart_disease.csv*:

1. age
2. sex (0 = female, 1 = male)
3. chest pain type (4 values)
   -- Value 1: typical angina
   -- Value 2: atypical angina
   -- Value 3: non-anginal pain
   -- Value 4: asymptomatic
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl (0 = false, 1 = true)
7. resting electrocardiographic results (3 values)
   -- Value 0: normal
   -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. maximum heart rate achieved
9. exercise induced angina (0 = false, 1 = true)
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
    -- Value 1: upsloping
    -- Value 2: flat
    -- Value 3: downsloping
12. number of major vessels (0-3) colored by flourosopy
13. thalassemia (0 = normal; 1 = fixed defect; 2 = reversable defect)
14. target (have heart disease, 0 = false, 1 = true)

Project 13: personal loan

When deciding whether to accept or reject a personal loan, a bank would usually make use the financial status of the applicant compared to the dataset containing information of all the loans in the past. Lending Club (LC) is an online marketplace that matches borrowers with lenders. This dataset contains information about financial status and loan status of 20000 loan borrowers on LC. The target of this project is to study this information and train a model to classify good/bad loans in order to set up an efficient way for decision making.

Description of the dataset *personal_loan.csv*:

1. id: unique ID of the loan application.
2. grade: LC assigned loan grade.
3. annual_inc: the self-reported annual income provided by the borrower during registration.
4. short_emp: 1 when employed for 1 year or less.
5. emp_length_num: employment length in years. Possible values are - between 0 and 10 where 0 means less than one year and 10 means ten or more years.
6. home_ownership: Type of home ownership.
7. dti (Debt-To-Income Ratio): a ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
8. purpose: a category provided by the borrower for the loan request.
9. term: the number of payments on the loan. Values are in months and can be either 36 or 60.
10. last_delinq_none: 1 when the borrower had at least one event of delinquency.
11. last_major_derog_none: 1 when the borrower had at least 90 days of a bad rating.
12. revol_util: revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
13. total_rec_late_fee: late fees received to date.
14. od_ratio: overdraft ratio.
15. bad_loan: 1 when a loan was not paid.

Project 14: diabetes

Diabetes is a group of metabolic disorders characterized by a high blood sugar level. This dataset is contributed by the National Institute of Diabetes and Digestive and Kidney Diseases. The data was collected from women at least 21 years old of the same ethnic group. The objective of this project is to predict whether a patient has diabetes or not based on certain diagnostic measurements.

Description of the dataset *diabetes.csv*:

1. Pregnancies - number of times pregnant
2. Glucose - plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. BloodPressure - diastolic blood pressure (mm Hg)
4. SkinThickness - triceps skin fold thickness (mm)
5. Insulin - 2 hour serum insulin (μU/ml)
6. BMI - body mass index, weight in kg divided by square of height in m
7. DiabetesPedigreeFunction - a function which scores likelihood of diabetes based on family history
8. Age - age in years
9. Outcome - 1 if the patient has diabetes

Project 15: wine quality

The quality of wine depends on a number of factors such as acidity, alcohol level and composition of various chemicals. These can be examined by physiochemical tests. The quality of wine can be scored from 0 to 10 based on sensory test. Those score over 5 is regarded as "good" while those score 5 or below is regarded as "bad". The dataset is based Portuguese red and white wines produced in the same province. The target is to build a machine learning model to classify whether a wine is good or bad based on these variables.

Description of the dataset *wine.csv*:
Input variables (based on physicochemical tests):

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol

Class label:

12 - quality ("good" / "bad")

Project 21: CO2 emission by vehicles

Emission of carbon dioxide (CO2) is believed to be a factor of global warming. Vehicle is one of the main sources of CO2 emission. To regulate CO2 emission by vehicles, the Canadian government has collected data from a few thousands of vehicle models. Their features as well as fuel consumption and CO2 emission and shown on the file. Your target is to study the relationship between these features and use a regression model to figure out how CO2 emission level is related to them.

Description of the dataset *co2_emission.csv*:

1. Manufacturer
2. Model
3. Vehicle class
4. Engine size (L)
5. Number of cylinders
6. Transmission type with number of gears:
   A = Automatic
   AM = Automated manual
   AS = Automatic with select shift
   AV = Continuously variable
   M = Manual
   3 - 10 = Number of gears
7. Fuel type:
   X = Regular gasoline
   Z = Premium gasoline
   D = Diesel
   E = Ethanol (E85)
   N = Natural gas
8. Fuel consumption in city roads (L/100 km)
9. Fuel consumption in highways (L/100 km)
10. The combined fuel consumption (55% city, 45% highway) (L/100 km)
11. The combined fuel consumption in both city and highway (mile per gallon)
12. tailpipe CO2 emissions (g/km) for combined city and highway driving

Project 22: concrete strength

Concrete is an essential material in civil engineering. The ratio of its components affect its strength. Over a thousand samples of concrete mixture have been tested for the compressive strength. The target of this project is to use linear regression model that describes the relationship between concrete compressive strength and the amount of different components.

Description of the dataset *concrete_strength.csv*:

- Component 1 - Cement
- Component 2 - Blast Furnace Slag
- Component 3 - Fly Ash
- Component 4 - Water
- Component 5 - Superplasticizer
- Component 6 - Coarse Aggregate
- Component 7 - Fine Aggregate
- Non component - Age (1 to 365 days)
- Output - Concrete compressive strength (MPa)

(Each of the components are measured in kg in a m^3 concrete mixture)

Project 23: health insurance

A personal health insurance covers the medical cost of a person in case of illnesses. Since the risk of illnesses depends on a lot of factors related to health status and living style, the insurance charge varies between individuals. The dataset contains the data of more than one thousand beneficiaries of medical insurance plan. The target of this project is to study the relationship between the health data and the insurance charge of different groups of individuals, and to build a regression model to predict the charge.

Description of the dataset *health_insurance.csv*:

1. age: age of primary beneficiary
2. sex: gender of the insurance contractor
3. bmi: body mass index, weight in kg divided by square of height in m
4. children: number of children / dependents covered by health insurance
5. smoker: 1 for smoking, 0 for non-smoking
6. region: the beneficiary's residential area in the US
7. charges: individual medical costs billed by health insurance per annum

Project 24: marketing promotion

A marketing director is responsible for organizing promotion campaigns and placing advertisement on different media channels. The effectiveness of these activities is reflected on the revenue of the stores. The dataset contains the revenue and amount of investment in marketing of a global corporation with about a thousand stores. The target of this project is to build a regression model to predict the revenue based on the investment on different means of marketing activities.

Description of the dataset *marketing.csv*:

1. id - identity number of the stores
2. reach - number of posts on social media such as twitter
3. local_tv - local television advertisement investment ($)
4. online - online advertisement investment ($)
5. instore - in store merchandizing investment ($)
6. person - number of store sales staff input
7. event - nature of the promotion event

Project 25: yacht hydrodynamics

Residuary resistance of a ship is a useful indicator for evaluating its performance and for estimating the required propulsive power. The Delft Ship Hydromechanics Laboratory has comprised three hundreds full-scale experiments resulting this dataset. The inputs include hull dimensions and boat velocity. The target of this project is to study how the residuary resistance is related to these coefficients and ratios, so that we can set up a regression model for prediction. Notice that all the measurements in the dataset are adimensional, i.e. no physical unit required.

Description of dataset *yacht.csv*:

Input variables:

1. Longitudinal position of the centre of buoyancy
2. Prismatic coefficient
3. Length-displacement ratio
4. Beam-draught ratio
5. Length-beam ratio
6. Froude number

Output variable:

7. Residuary resistance per unit weight of displacement

Project 2X: stock price

The price of a stock listed on the US market is affected by a number of social and economic factors. These factors can be reflected by various stock market indexes such as Dow Jones Industrial Average (DJI), Standard and Poor's 500 (S&P 500), Nasdaq 100 (NDX). Other factors including foreign currency rates, crude oil price, treasury yield bond rate, etc might also affect stock price. This project is an open topic. Choose any listed company you like and use its adjusted close price in recent 5 years as the output variable. Build a regression model with those input variables you think is essential to predict the stock price. You may visit finance.yahoo.com and download the data as a csv file as your data set.