Project Description

Your team consists of 4 to 5 students. Each team will submit a project report in the format of **Jupyter Notebook and its PDF version** and will give a **10 minutes powerpoint presentation**. In your project, you will perform the following steps on the selected dataset:

**1. Data Preparation and Pre-prediction Analysis**

a. Explore Data

b. Find statistics

b. Apply cluster analysis/Time series analysis and other methods you learned on data

**2. Predictive Modeling (Classification)**

a. Classification using Decision Tree

b. Classification using Naive Bayes

c. Choose one another classification algorithm of your choice, apply and describe it

d. Compare the results of the 3 techniques

**3. Conclusions and Recommendations**

## Data Preparation and Pre-prediction Analysis

The first and foremost step of data mining process is to understand the data and identify the research question(s). Here are some suggestions to explore and understand datasets:

* Look at the attribute type; e.g., categorical, ordinal or quantitative.
* Find max, min, mean and standard deviation of attributes.
* Determine any outlier values (records) for each of the attributes or attributes under
* consideration (min, max, std. dev, scatter plots, box plots or others can be used).
* Analyze the distribution of numeric attributes (normal or other).
* Plot histograms for attributes of concern and analyze whether they have any influence on the class
* Try to answer these questions by different visualization techniques:
  + Which attributes seem to be most linked to the class attribute?
  + Which attributes seem to be most linked to the class attribute?
  + Which attributes do you think can be eliminated or included in the analysis?
  + Determine whether the dataset has an imbalanced class distribution (same
  + proportion of records of different types or not).

## Predictive Modeling (Classification)

After an overall understanding about the dataset, you can use classification algorithms, Decision Tree and Naïve Bayes. Also, choose a classification algorithm of your own choice, explain it a at a high level and compare your results

* You will predict the class attribute by using each classification algorithm.
* Determine the right strategy for dataset split: simple training or testing, 10-fold cross validation, 3-fold cross validation, etc.
* Repeat the same process for Decision Trees, Naïve Bayes and the third classification algorithm of your choice.
* Determine your performance measures (accuracy, recall, etc.).
* Identify which algorithm performs well and in which settings.

## Conclusions and Recommendations

State your major findings from different sections. State your recommendation to the company that they can put into place to solve their problem.

## Datasets

Select only one of the following datasets for your group and solve the problem for this

company described for a particular dataset below. **The class attribute of each dataset is highlighted.**

**NOTE: You can also select your own dataset by sharing the details with me. The procedure should be same, and you need to solve a classification problem.**

### Churn Dataset

Customer churn or customer attrition means the loss of customers for a company. The

problem for customer churn is that the company would like to know in advance which

customers would churn in near future. You are a member of a team of data scientists and

the task of your team is to help this company in characterizing customer churn through data

analytics methods. This dataset has 21 attributes including a binary class attribute about

churn. The descriptions of the attributes are given below:

1. State: Customer’s state.

2. Account Length: Integer number showing the duration of activity for customer account.

3. Area Code: Area code of customer.

4. Phone Number: Phone number of customer.

5. Inter Plan: Binary indicator showing whether the customer has international calling plan.

6. VoiceMail Plan: Indicator of voice mail plan.

7. No of Vmail Mesgs: The number of voicemail messages.

8. Total Day Min: The number of minutes the customer used the service during day time

(continuous quantitative data type).

9. Total Day Calls: Discrete attribute indicating the total number of calls during day time.

10. Total Day Charge: Charges for using the service during day time (continuous data

type).

11. Total Evening Min: The number of minutes the customer used the service during

evening time.

12. Total Evening Calls: The number of calls during evening time.

13. Total Evening Charge: Charges for using the service during evening time.

14. Total Night Min: Number of minutes the customer used the service during night time.

15. Total Night Calls: The number of calls during night time.

16. Total Night Charge: Charges for using the service during night time.

17. Total Int Min: Number of minutes the customer used the service to make international

calls.

18. Total Int Calls: The number of international calls.

19. Total Int Charge: Charges for international calls.

20. No of Calls Customer Service: The number of calls to customer support service.

**21. Churn: Class attribute with binary values (True for churn and False for not churn).**

### Bank Marketing Dataset

This dataset refers to the problem of telemarketing for a bank. The dataset is collected from

a Portuguese bank and the bank wants to have an effective telemarketing strategy to sell

long-term deposit accounts (e.g., bonds, saving accounts, etc.). These marketing

campaigns were based on phone calls and multiple contacts were often needed to

determine whether a customer would subscribe to a long-term deposit account. Your team of

data scientists will help this bank in determining such customers and devising an effective

telemarketing strategy by applying data analytics method on the given dataset.

1 – Age: Age of the customer (numeric).

2 - Job: Type of job (qualitative).

3 - Marital: Marital status (qualitative).

4 – Education: Education of the customer (qualitative).

5 - Default: Shows whether the customer has credit in default or not (qualitative).

6 - Balance: Average yearly balance in Euros (numeric).

7 - Housing: Shows whether the customer has housing loan or not (qualitative).

8 - Loan: Shows whether the customer has personal loan or not (qualitative/categorical).

9 - Contact: Shows how the last contact for marketing campaign has been made

(qualitative)

10 - Day: Shows on which day of the month last time customer was contacted (numeric).

11 - Month: Shows on which month of the year last time customer was contacted

(qualitative).

12 - Duration: Shows the last contact duration in seconds (numeric).

13 - Campaign: Number of contacts performed during the marketing campaign and for this

customer (numeric).

14 - Pdays: Number of days that passed by after the client was last contacted from a

previous campaign (numeric, -1 means client was not previously contacted).

15 - Previous: Number of contacts performed before this campaign and for this client

(numeric).

16 – Poutcome: Outcome of the previous marketing campaign (qualitative).

**17 - Y – Class attribute showing whether the client has subscribed a term deposit or not (binary: "yes","no")**

### Credit Card Dataset

In order to provide loans to customers, a bank needs to make right decision in determining who

should get the approval and who should not. This dataset is the German Credit Data that contains 20 attributes and the class attribute showing a good or a bad credit risk. Your team of data scientists will need to develop a data analytics based strategy for the bank managers that can help them in making a decision about loan approval for the prospective applicants.

**1. Creditability: The class attribute (qualitative)showing whether the credit rating is good or**

**bad.**

2. Account Balance: Checking account status (1: < 0 DM, 2: 0<=...<200 DM, 2 > 200 DM, 4: No

checking account), where DM= Deutsche Mark (qualitative attribute).

3. Duration of Credit (month): Duration of credit in months (numerical)

4. Payment Status of Previous Credit: Credit history (qualitative) 0: no credits taken, 1: all

credits at this bank paid back duly, 2: existing credits paid back duly till now, 3: delay in

paying off in the past, 4: critical account.

5. Purpose: Qualitative attribute showing the purpose of the loan (0: New car, 1: Used car , 2:

Furniture/Equipment, 3: Radio/Television, 4: Domestic Appliances , 5: Repairs ,6: Education ,7:

Vacation, 8: Retraining ,9: Business, 10: Others)

6. Credit Amount: Numerical value showing the credit amount

7. Value Savings/Stocks: Qualitative attribute showing average balance in savings and stocks (1 : <100 DM, 2: 100<= ... < 500 DM, 3 : 500<= ... < 1000 DM, 4 : =>1000 DM, 5: unknown/ no savings account)

8. Length of current employment: Qualitative attribute showing length of employment (1 :

unemployed, 2: < 1 year, 3: 1<=...<4 years, 4: 4<=...<7 years, 5:>=7years).

9. Instalment percent: Installment rate in percentage of disposable income (numerical)

10. Sex & Marital Status: Qualitative attribute showing gender and marital status (1: male :

divorced/separated, 2: female : divorced/separated/married, 3 : male: single, 4: male :

married/widowed, 5 : female : single)

11. Guarantors: (Qualitative) Guarantors and co-applicants: (1 : none, 2 : co-applicant, 3 :

guarantor)

12. Duration in Current address: Qualitative value showing the duration in current address (1: <= 1 year, 1<...<=2 years, 2<...<=3 years, 3:>4years)

13. Most valuable available asset: Qualitative attribute showing valuable assets ( 1 : real estate

2 : savings agreement/ life insurance, 3 : car or other, 4 : unknown / no property)

14. Age (years): Numerical value showing age in years.

15. Concurrent Credits: Installment plans ( 1 : bank, 2 : stores, 3 : none )

16. Type of apartment: Type of housing ( 1 : rent, 2 : own, 3 : for free)

17. No of Credits at this Bank: Numerical value showing number of existing credits at the bank

18. Occupation: Job (Qualitative) (1 : unemployed/ unskilled - non-resident, 2 : unskilled - resident,

3 : skilled employee / official, 4 : management/ self-employed/highly qualified employee/ officer)

19. No of dependents: Numerical value showing number of dependents

20. Telephone: Qualitative attribute for telephone number (1: yes, 2: No)

21. Foreign Worker: Qualitative attribute showing whether the person is the foreign worker or not (1: yes , 2: no)

### How to Compare Your Classification Models (Model Selection)

In order to evaluate and compare machine learning models with different features, a known

approach is to create a baseline model first. This can be done by training one model (e.g.,

decision tree) on the training set using the entire feature set (all attributes) and evaluating its

performance using the selected metric (such as accuracy, true positive rate, false positive rate,

etc.) on the validation set (or test set). Optimization of the model parameters or selection of

features could be done by changing one variable (e.g., one parameter or one feature) at a time

and re-training the model on the same training set. Finally, one needs to compare the

performance of different models built using different features on the same validation set (or

test set). This would give an indication whether that variable (feature or parameter) has

increased or decreased the performance.

1. Using 3-ways data splitting (Here, you will have to divide a dataset yourself or find a library/function in Python):

• Training (e.g., 60%): for training the model

• Validation (e.g., 20%): for evaluating and comparing the performance after varying

parameters, features, etc. In addition, this is used for selecting the “best” model.

• Testing (e.g., 20%): only used at the end to evaluate the final performance and report

the results of the selected models (best performing models from the above step)

• This video describes this approach: <https://www.youtube.com/watch?v=4wGquWG-vGw>

2. Using 10-fold cross validation (10-FCV) on the entire dataset:

• Create the base model by train and evaluate its (average) performance using 10-FCV.

• Change model parameters or features, retrain and re-evaluate the (average) model

performance using the 10-FCV strategy.

• Report the results of best performing models (using the best parameters and features

select from the above step).

## Prediction ofinjuries in Fire Services Incident Data (Advanced)

You can find many datasets in open data catalogue :

https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/

For this problem You can search Fire

https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#e3d443bb-2593-2615-4972-20e24c0ab876

1. Clean the dataset, and merge them to one clean dataset as a CSV file
2. Follow  **Data Preparation and Pre-prediction Analysis** steps. Similar to the Titanic example, try to figure out the relationships among variables. Give us reasons if you drop a column. Show us which attributes are more important in predicting the injuries. (Hint: You need to convert injuries column to a categorical column)
3. Use Time-series analysis (by plotting different time series plots) for incidents during the months and show the incident trends for different years and months, Do you see any patterns or periods? What about trends?
4. **(Bonus marks)** Try to predict if the incident leads to an injury or not.

**Dataset**: [Fire Services Incident Data](https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#e3d443bb-2593-2615-4972-20e24c0ab876)

https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/

**Variable of interest:** Incidents Injuries (FF\_INJURIES or CIVILIAN\_FIRE\_INJURY or CIVILIAN\_FIRE\_FATALITY , ...). (Any of these could be considered as an injury)

**Some Variables of Interest:**

|  |
| --- |
| RESCUED\_CHILDREN |
| RESCUED\_ADULTS |
| RESCUED\_SENIORS |
| PHYSICAL\_CONDITION\_1 |
| PHYSICAL\_CONDITION\_2 |
| PHYSICAL\_CONDITION\_3 |
| CIV\_FIRE\_CONTROL |
| CIV\_EVACUATION |
| CIV\_EVACUATION\_REASON\_1 |
| CIV\_EVACUATION\_REASON\_2 |