Practical Assignment

BIP: Machine Learning for Data Science
Year: 2023/2024
Due Date: 23/01/2024

Please take the following notes in consideration:

- The practical assignment is to be developed in teams of **five** students.
- It is possible to use any framework or programming language you want.
- It is also possible to use other sources of information and "inspiration", such as ChatGPT, GitHub, Kaggle, or many others.
- Each team can submit up to 3 classification results.
- In addition, it is expected that all the teams submmit a report, in the format of a scientific paper, describing the context, frameworks used, methodology, results and discussion.
- Do not forget to **mention all the team members** in the report!

Overview

Water is, probably, one of the most important resources for human survivability. It is used in almost all aspects of human activity, from ingestion, food production, hygiene, industry, leisure, and many others.

As people tend to concentrate in cities, the responsibility of supplying a constant stream of fresh water lies with the municipality which, to rationalize the consumption, records the amount each citizen, family, or factory consume. Moreover, taxing the consumption can also depend on the type of consumer so, municipalities also register this information in the database of consumers.

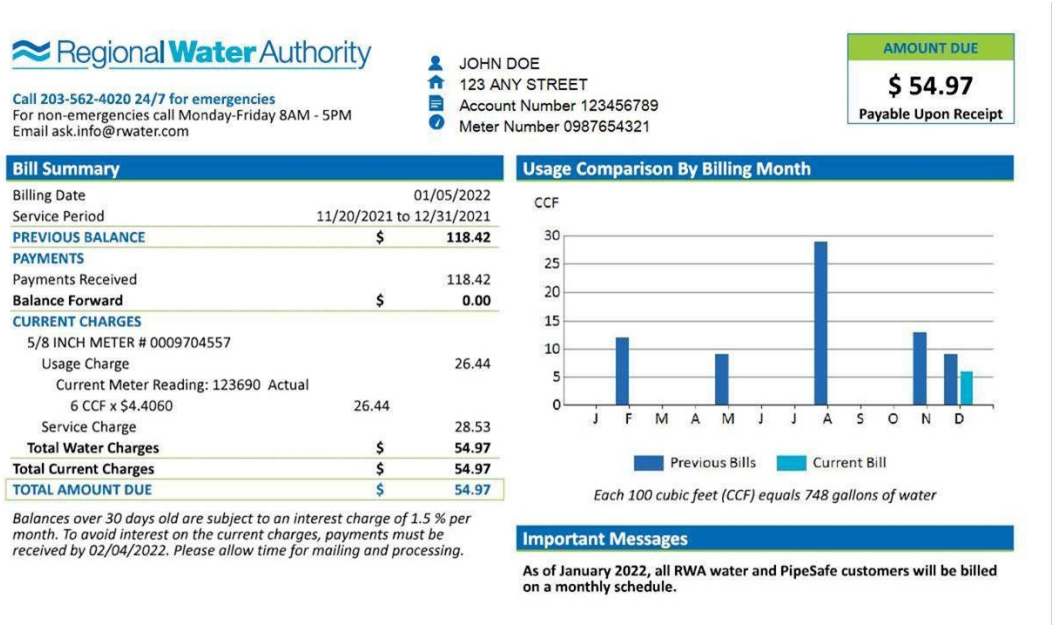
All this information is used to produce the consumption report and the bill for the payment of the consumption (Figure 1). If you pay attention to the header of the water bill, it is possible to find information such as the `consumer number`, `installation zone`, `installation number`, `consumer type`, and others. In the detail of the bill, you can find the consumption, usually taxed according to a set of degrees of consumption.

As you may already have guessed, this is the main topic of the practical assignment for the Machine Learning for Data Science BIP.

Dataset

1. `Details`

   The dataset to be used in this assignment was collected from a water consumption record in a municipality. It comprises a monthly record of water consumption by all the measuring devices installed in the region. Each example correspond to a single measurement, tagged with the year and month when it was collected, and containing the consumer identification, installation local, and consumption. This means that there are several examples for the same consumer...



Regional Water Authority

Call 203-562-4020 24/7 for emergencies
For non-emergencies call Monday-Friday 8AM - 5PM
Email ask.info@rwater.com

JOHN DOE
123 ANY STREET
Account Number 123456789
Meter Number 0987654321

AMOUNT DUE
$ 54.97
Payable Upon Receipt

Bill Summary

Billing Date		01/05/2022
Service Period		11/20/2021 to 12/31/2021
PREVIOUS BALANCE	$	118.42
PAYMENTS		
Payments Received		118.42
Balance Forward	$	0.00
CURRENT CHARGES		
5/8 INCH METER # 0009704557		
Usage Charge		26.44
Current Meter Reading: 123690 Actual		
6 CCF x $4.4060	26.44	
Service Charge		28.53
Total Water Charges	$	54.97
Total Current Charges	$	54.97
TOTAL AMOUNT DUE	$	54.97

Balances over 30 days old are subject to an interest charge of 1.5 % per month. To avoid interest on the current charges, payments must be received by 02/04/2022. Please allow time for mailing and processing.

Usage Comparison By Billing Month

Each 100 cubic feet (CCF) equals 748 gallons of water

Important Messages

As of January 2022, all RWA water and PipeSafe customers will be billed on a monthly schedule.

Figure 1: Sample of a water bill document[1]

[1] Source: https://www.sydneywater.com.au/accounts-billing/paying-your-bill/bill-payment/how-to-read-your-bill.html

2. `Features`

   The dataset has 5 (five) features and 1 (one) label (Figure 2):

   Year – the year of the record

   Month – the month of the consumption record

   Consumer_type – this is the **label**, corresponding to the type of the consumer

   Consumption – water consumption in cubic meters

   Consumer_number – the ID of the consumer (text)

   Installation_zone – the area in the region where the water was consumed (text)

Year	Month	Consumer_type	Consumption	Consumer_number	Installation_zone
2013	1	domestic	0	BUZM22700877312041	Installation_zone 1
2013	1	industrial	5	HZDT30465046614741	Installation_zone 2
2013	1	domestic	6	QONO32975566647957	Installation_zone 2
2013	1	domestic	1	NOEA60398551564108	Installation_zone 2
2013	1	domestic	13	URXS49372209503661	Installation_zone 2
2013	1	industrial	27	PYXS50360080481802	Installation_zone 2
2013	1	industrial	5	DDYQ20906278027034	Installation_zone 2
2013	1	domestic	31	IUKV91270931496744	Installation_zone 2
2013	1	industrial	2	AJLS69672632154052	Installation_zone 2

Figure 2: Feature set sample of the water consumption dataset

3. `Annotation`

   Each data entry is annotated with the type of consumer, among 9 different classes (Figure 3):

- domestic - regular house consumers
- rural domestic - small companies or familiar agro-production consumers
- industrial - industries
- rural commercial - mid-sized agro-production consumers
- construction - construction related companies and sites
- low income families - families with social security support
- rural expansion - agro-production companies that don't categorize as one of the previous

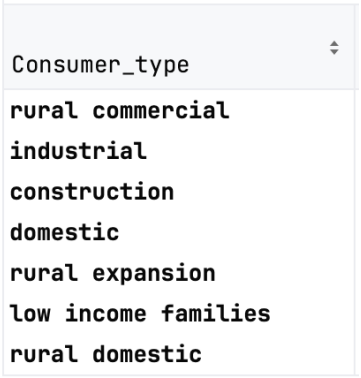


Figure 3: Annotation classes

4. `Time span`

   The records are from 2013 and 2020.

   ⚠Beware that the 2015 year is missing⚠

Task

The assignment has the format of a competition, where students have to find the best results within the goal of this work.

5. `Goal`

   This task aims to classify a set of consumers according to the consumer type (see section 3). In other words, your task is to use multiclass classification to classify a water consumer type, given information about the monthly consumption pattern of the previous 10 years:

   - For that, you should build and evaluate a multi-class classifier for consumer type classification, according to the 9 different classes.
   - After training, you should classify a set of consumers, provided in a CSV file.
   - This result has to be submitted in a CSV format with two columns (Figure 4).

   Each team can submit up to three classification results, usually produced with different models.

Consumer_number	Consumer_type
AABH19026729995402	domestic
AABK96307399687530	domestic
AABP15829373762695	domestic
AABU83206956615238	domestic
AACK80576350114306	domestic
AACM38974282265332	domestic
AADB02534023128621	domestic
AADD80167034367129	domestic
AAET76171415183608	domestic
AAFB08665155868896	rural domestic

Figure 4: Submission format sample

Deliverables

1. All deliverables should be submitted in a single archive
2. The submission should be done in IPB Virtual, in the assignments tool
3. The **archive** should contain:
   - A PDF file with the report
   - One CSV file for each classification result (with a maximum of three)
   - The models and the code produced
4. The Zip file should have the name `team_X.zip`.