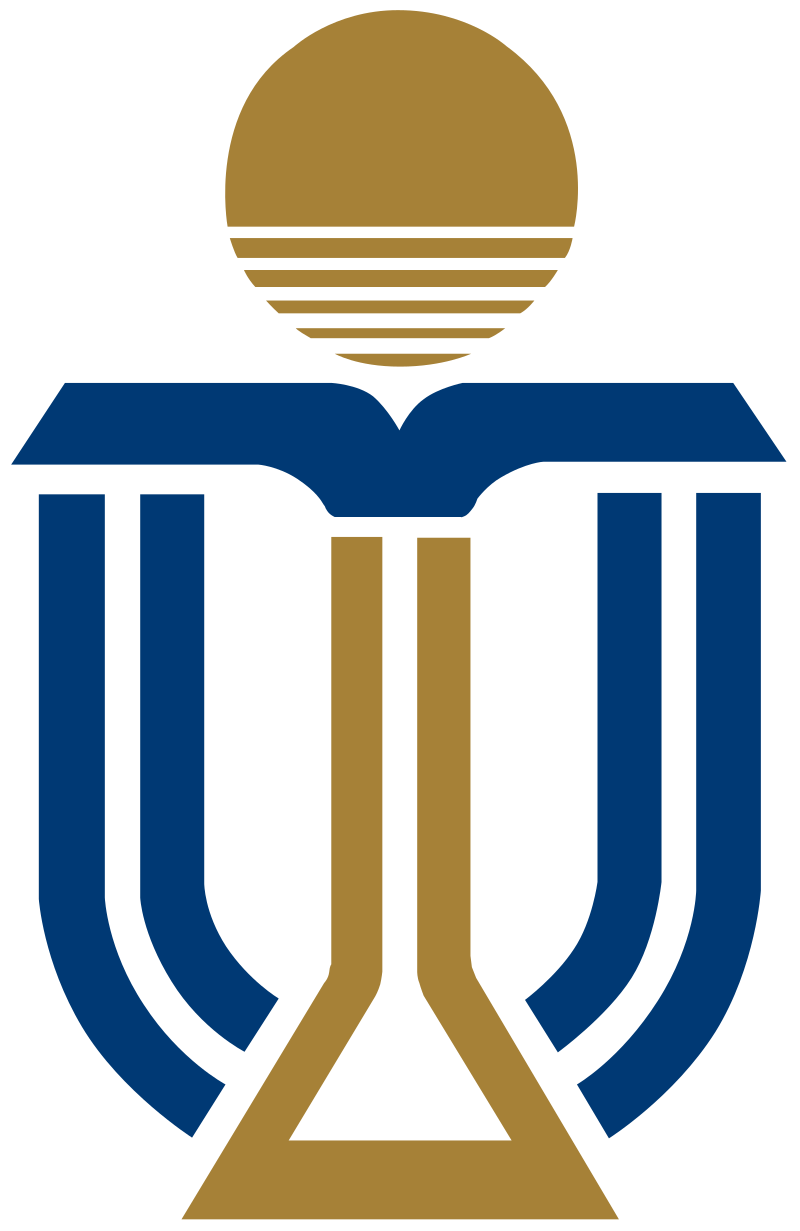
Programming Assignment 1

COMP 4211



Andreas Wold Sløgedal

Student ID: 21196840

Contents

[1 Introduction 3](#_Toc190772446)

[1.1 Description of the dataset 3](#_Toc190772447)

[2 Part 1: Data exploration and Preparation 3](#_Toc190772448)

[2.1 [Q1] Dataset Overview 3](#_Toc190772449)

# Introduction

The purpose of this programming assignment is to use the material from COMP4211 to explore and preprocess a provided dataset (data.csv) and create linear regression and logistic regression models to … . The models will use information about individuals from the dataset and be trained using advanced machine learning techniques. Additionally, the aim is for the models to …

## 1.1 Description of the dataset

Data.csv includes eleven independent variables. c1 describes gender, and is a binary variable….

2 Part 1: Data exploration and Preparation

In order to do sufficient data exploration and preparation I import the Python-library Pandas, which is a powerful tool for data manipulation and analysis. Using Pandas built in function pd.read\_cs()I import data.csv and save dataset in the variable df.

## 2.1 [Q1] Dataset Overview

Using pd.shape I am able to examine the structure of the dataset. The dataframe contains 891 rows and 11 columns, where the number of rows corresponds to the number of individuals, and the number of columns the number of attributes assigned to each individual. Furthermore, df.size identifies the total number of elements in the dataframe. This number is beneficial to examine, as it gives the total count of all data points, including missing values.

Missing values in a dataset are important to identify and process according to their structure. Using df.isna().sum() I sum over the total number of missing values for each attribute. As a result, I identify that c2 is missing 177 values, c6 is missing 687 values, and c7 is missing 2 values. This corresponds to a proportion of 19.87%, 77.10% and 0.22% respectively.

Before training regression and classification models it is necessary to handle NaN (Not a Number) values to ensure data quality and prevent bias. c2 has a moderate proportion of missing values (19.87%), which could distort model training. Hence, it could be reasonable to substitute missing values with mean or median imputation. Furthermore, c6 has a much higher proportion of missing values (77.10%). Using such a large percentage of missing values when training models could impact the model negatively, and it could be beneficial to drop the column depending on the importance of the attribute. Additionally, one could impute using KNN implementation, which considers relationship between other features to predict missing values. Lastly, c3 has a very low proportion of missing values, hence the impact on model performance is relatively low.

## 2.2 [Q2] Feature Distribution

**Q1:**

* The dataset contains 891 rows and 11 columns. The shape of the dataset is 9801, which is the total elements in the dataset.
* There are some missing vales in the dataset.
  + Category c2 is missing 177 values. The proportion is 177/891 = 0.198653
  + Category c6 is missing 687 values. The proportion is 687/891 = 0.771044
  + Category c7 is missing 2 values. The proportion is 2/ 891 = 0.002245
* Category c2 is missing roughly 20% of its values. Since the model lacks values, bias is introduced in the regression- and classification models. Hence, the regression model will have inconsistencies, unless we use imputation techniques for the missing values.
* Category c6 is missing roughly 80% if its values. It will be very difficult for the models to predict consistent values due to the high proportion of missing values. Hence, we should use imputation techniques on this category to eliminate missing values.
* Category c7 is missing a low proportion of values, which will have less impact on model performance. Still, we should use imputation techniques to eliminate the NaN values.

Classification models 🡪 most common measure = cross-entropy loss