CITS5504 Data Warehousing

Project Report

Varun Jain 21963986 Akhil Naseem x0x0x0

Semester 1 2021

**Project 2**

**Pattern Discovery and Building Predictive Models**

# Introduction

Mobile Price Classification Dataset, provided by the UWA CITS5508 Team, is set to be analysed for Pattern Discovery and Predictive Modelling. The aim of this project is to apply several different machine learning algorithms to predict weather the price of a mobile phone is high or low. The mobile\_price.csv file will undergo vigorous data pre-processing, and the original dataset may be split into separate data marts dependent on the machine learning technique applied. …

# Data Pre-processing

## Quick Overview of Data

For this scenario, let’s assume the mobile\_price.csv is the Datawarehouse. A quick observation made indicates the data-warehouse contains 22 individual attributes with 2000 instances recorded. Each attribute except the initial column represents a component of a mobile phone. The component refers to a certain feature or the size of the feature. For instance, the following attributes “blue”, “wifi, “ram”, “sc\_h” and “sc\_w” indicates whether the phone has Bluetooth and wife, the ram in megabytes and the height and width of the mobile phone in centimetres, retrospectively.

## Attribute Type

The type of each attribute in the data warehouse is defined in the table below:

May need to change:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Attribute | Type | Attribute | Type | Attribute | Type | Attribute | Type |
| Id | Nominal | **four\_g** | Binary | **px\_height** | Numeric | **three\_g** | Binary |
| Battery\_power | Numeric | **int\_memory** | Numeric | **px\_width** | Numeric | **touch\_screen** | Binary |
| blue | Binary | **m\_dep** | Numeric | **ram** | Numeric | **wifi** | Binary |
| clock\_speed | Numeric | **mobile\_wt** | Numeric | **sc\_h** | Numeric | **price\_category** | Binary |
| dual\_sim | Binary | **n\_cores** | Ordinal | **sc\_w** | Numeric |  |  |
| fc | Numeric | **pc** | Numeric | **talk\_time** | Numeric |  |  |

## Relevant Features

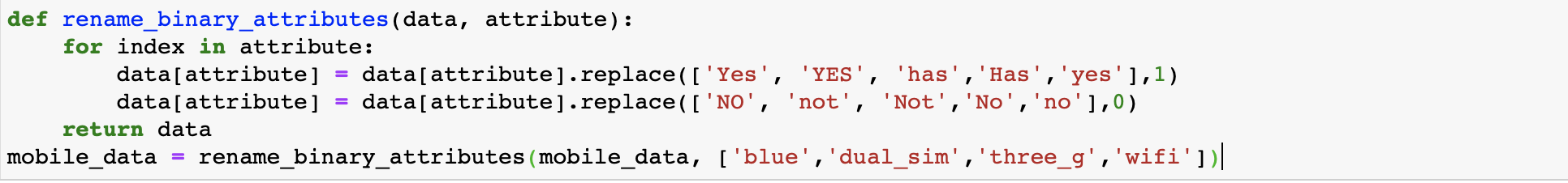
Not sure yet…

## Data Cleaning

Mobile Price Classification Dataset is a relatively clean dataset. From my analysis, I noticed that dataset contains no null values in other words, there are no missing values within the entire dataset. Also the attribute **id** has been removed from the dataset as there is no need for a unique identifier for out classifications tasks ahead.

### Naming Conventions

The attributes **blue**, **dual\_sim**, **three\_g**, and **wifi** map to a binary output, yes or no. Within these attributes, I noticed strange naming conventions, and incorrect capitalisation so in other words, the attribute values within those attributes contain inconsistent character-based data types. For instance, **blue** was encoded with the following values: ['Yes', 'YES', 'has','Has','yes'] for the binary output yes, and ['NO', 'not', 'Not','No','no'] for the binary output no. To fix this issue, I replaced all the values related to ‘yes’ to 1, and ‘no’ to 0. This method was applied to all the attribute values in the columns mentions above. This can be seen in the code below.



### Data Clean for Aprior Mining

1. Select the features that we want (maybe test it after you train the whole dataset)
2. Discretise the numerical data, the bin it with certain values.
3. Expand the data to 1’s and 0’s.
4. Apply the association mining rules.