Machine Learning: Exercise 0

Amélie Assmayr (12007770) *Konstantinos Damanakis (12106343)*
Teresa Schuch (12007762)

1 Introduction

In this Exercise we have chosen two datasets: Laptop Prices and Road Traffic Accidents. The Laptop Prices dataset will be utilized to develop regression models for predicting laptop prices , while the Road Traffic Accidents dataset will be employed to classify the severity of accidents. We chose these datasets because they enable us to try various techniques as they differ in important aspects. The Laptop Prices dataset has 2275 instances, 15 features and no missing values. In contrast, the Road Traffic Accidents dataset contains 12316 instances, 32 features and missing values in 16 features, making the preprocessing more important and complex.

2 Dataset 1: Laptop Prices

The Laptop Price dataset contains information about various laptops and their price in euros. We chose this dataset because of its importance in real-life situations. Most people today rely on laptops for work, education and entertainment. Predicting these prices can provide retailers with crucial insights into expected revenues before official pricing is announced. Additionally, the variety of features in the dataset, including both categorical and numeric data makes it ideal for experimenting with feature engineering.
The variables, their data types and the number of unique values for nominal and ordinal variables can be seen below.

Variable	Datatype	Unique
Company	nominal	19
Product	nominal	618
Type Name	nominal	6
Inches	ratio quantity	/
Screen Resolution	ordinal	40
CPU Company	nominal	3
CPU Type	nominal	106
CPU Frequency (GHz)	ratio quantity	/

Variable	Datatype	Unique
RAM (GB)	ratio quantity	/
Memory	ordinal	39
GPU Company	nominal	4
GPU Type	nominal	106
Operating System	nominal	9
Weight (kg)	ratio quantity	/
Price (Euro)	ratio quantity	/

2.1 Target attribute

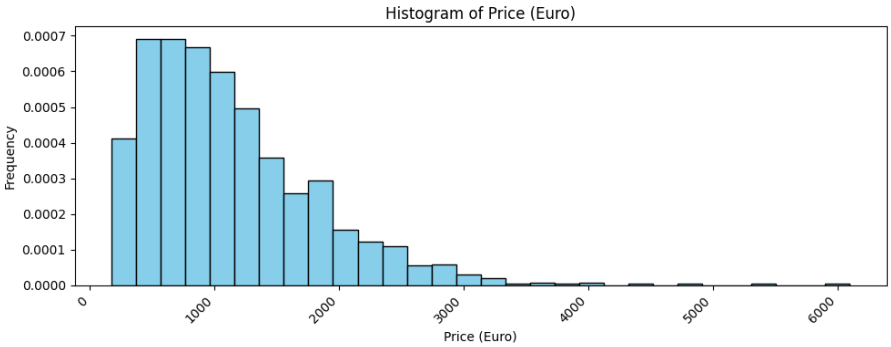


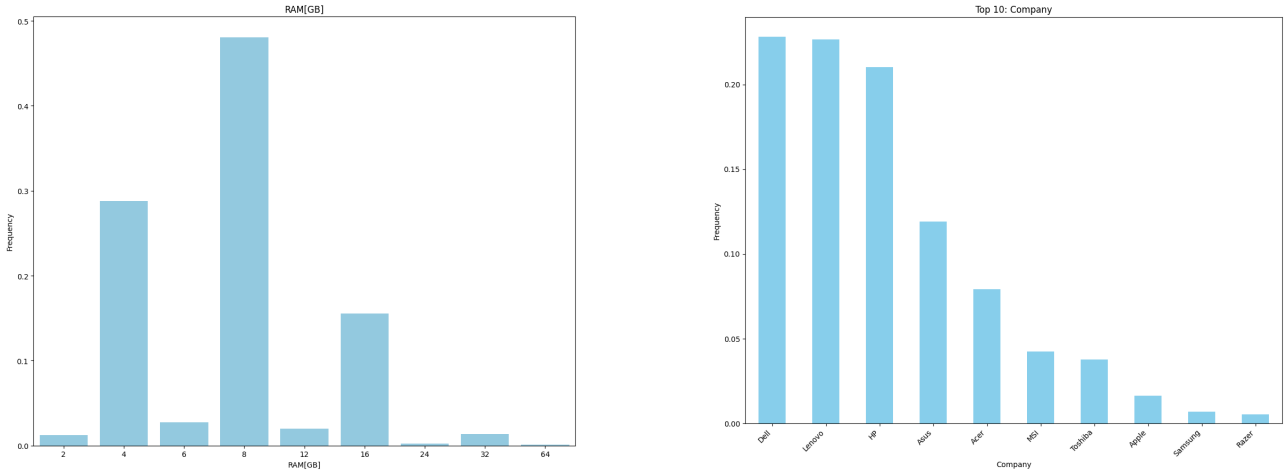
Figure 1: Histogram of Laptop Prices

The target variable is the price of the laptop. The prices range between 174€ and 6099€. The histogram in Figure 1 illustrates the distribution of the prices, revealing that most laptops are priced between 400 and 800€. The mean price is 1135€ and the median price 989 €. There are a few outliers that cost over 4000€, but these outliers are valid values.

2.2 Additional Attribute Insights

The screensize ranges from 10.10 to 18.40 inches, with a mean of 15.02 inches. The weight varies between 0.690 kg and 4.7 kg, with an average weight of 2.041 kg. The CPU frequency spans from 0.9 to 3.6 GHz, with a mean value of 2.5 GHz and RAM takes integer values between 2 and 64 GB, with a median value of 8 GB. The variable product may require preprocessing and summarizing due to its numerous unique values.

Due to limited space not all variable plots could be displayed. In Figure 2 the distribution of RAM and the Top 10 Companies can be seen as an example.



(a) Barplot of RAM

(b) Barplot of the Top 10 Companies

Figure 2: Plots from Laptop dataset

3 Dataset 2: Road Traffic Accidents

The second dataset includes instances of road traffic accidents recorded in the years 2017 to 2020 in Addis Ababa in Ethiopia. This dataset contains information regarding severity of the accident, the number and the characteristics of the involved vehicles and people, area etc.We chose this dataset because it addresses a critical public safety issue. A study of this data could help predict the severity of the accidents which can help identify causes and factors contributing to road safety.
Missing data can be observed in 16 of the 32 features. Overall there are s 20.057 missing values. In the following table all attributes, their datatype(r=ratio quantity, o=ordinal, n=nominal), the number of unique elements for categorical variables and the number of missing values can be seen :

Variable	**Type**	**Unique**	**Missing**	**Variable**	**Type**	**Unique**	**Missing**
Time	r	/	/	*Road Surface Condition*	n	4	172
Day of Week	n	7	/	*Light Condition*	n	4	/
Age of Driver	o	/	/	*Weather Condition*	n	9	/
Sex of Driver	n	3	/	*Collision Type*	n	10	155
Educational Level	n	7	741	*Number of Vehicles*	r	/	/
Vehicle-Driver Relation	n	4	579	*Number of Casualty*	r	/	/
Driving Experience	o	/	829	*Vehicle Movement*	n	13	308
Vehicle Type	n	17	950	*Casualty Class*	n	4	/
Vehicle Owner	n	4	482	*Sex of Casualty*	n	3	/
Vehicle Service Year	r	/	3928	*Age of Casualty*	r	/	/
Defect of Vehicle	n	3	4427	*Casualty Severity*	o	4	/
Accident Area	n	14	239	*Work of Casualty*	n	7	3198
Lanes or Medians	n	7	385	*Fitness of Casualty*	n	5	2635
Road Alignment	n	9	142	*Pedestrian Movement*	n	9	/
Junction Type	n	8	887	*Cause of Accident*	n	20	/
Road Surface Type	n	4	172	*Severity of Accident*	n	3	/

3.1 Target Attribute

The target variable is the severity of the accident. It is classified into three classes: Light Injury, Serious Injury or Fatal Injury. While the severity could be considered an ordinal variable, we will treat it as nominal, framing the task as a multiclass classification problem. The following plot shows the distribution of these classes. It can be seen that the classes are unevenly distributed. Most accidents fall under the "Slight Injury" category, followed by "Serious Injury" and only a few accidents are classified as "Fatal Injury." This imbalance is a critical factor to consider. Methods like bootstrapping might be helpful to employ.

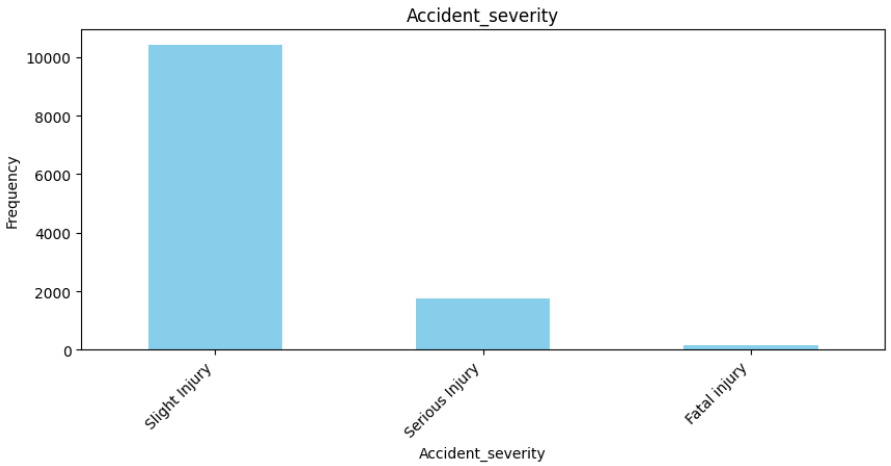
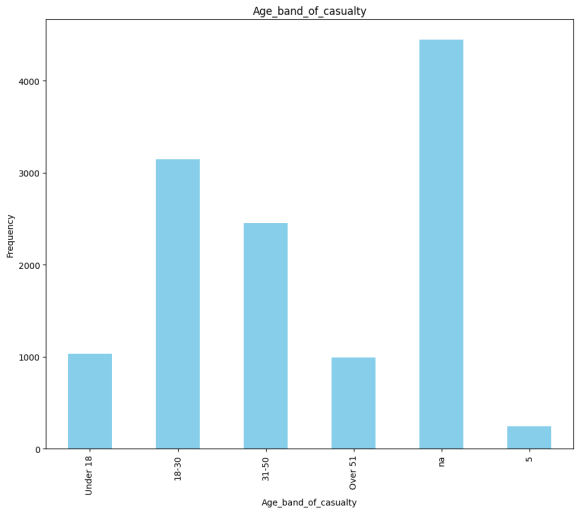


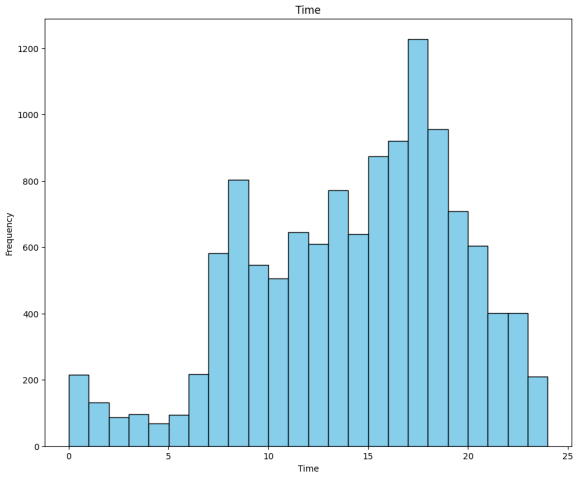
Figure 3: Barplot of Accident Severity

3.2 additional attribute insights

Most of the numeric variables in the dataset are grouped into intervals, effectively converting them into ordinal data. Besides regular missing values, the dataset also has other entries like 'Unknown' or 'na' that should be handled during preprocessing. As an example the histogram of the time of the accident and the barplot of the age classses of the driver are shown below in Figure 4:



(a) Barplot of the age classes



(b) Histogram of the Time

Figure 4: Plots from Road Traffic Accident dataset