Exercise 0 - Dataset description

VU Machine Learning - Group 73

Olsacher (01520348) Ponesch (11818774) Winkler (12013078)

October 15, 2025

Dataset 1: UJIIndoorLoc [1]

Dataset Description

The **UJIIndoorLoc** dataset is an indoor localization dataset based on WiFi fingerprinting, collected in 2013 at Universitat Jaume I, Spain. It covers four floors across three buildings and was designed as a reference benchmark for comparing different indoor localization methods based on WiFi signals. The dataset consists of a **training set** with 19,937 samples and a **validation/test set** with 1,111 samples. Each entry contains 520 features, **WAP001**–**WAP520**, representing the measured RSSI (Received Signal Strength Indicator) values ranging from -104 dBm to 0 dBm, or 100 if the device was not in range of this specific access point. These features are **numeric (interval-scaled)**.

As target variables, the dataset provides **Longitude** (numeric, interval), **Latitude** (numeric, interval), **Floor** (ordinal), **Building ID** (nominal), **Space ID** (nominal), and **Relative Position** (nominal). Additional metadata include the **User ID** (nominal), **Phone ID** (nominal), and **Timestamp** (ordinal/temporal).

Two main prediction tasks can be formulated: a regression task (predicting longitude, latitude, and height) or a **classification** task (predicting the combination of Building ID, Floor, and Space ID). Space IDs are not globally unique, but the triplet combination of **(Building ID, Floor, Space ID)** is unique. In total, the dataset contains 735 unique combinations and 123 distinct Space IDs. The dataset has no missing values.

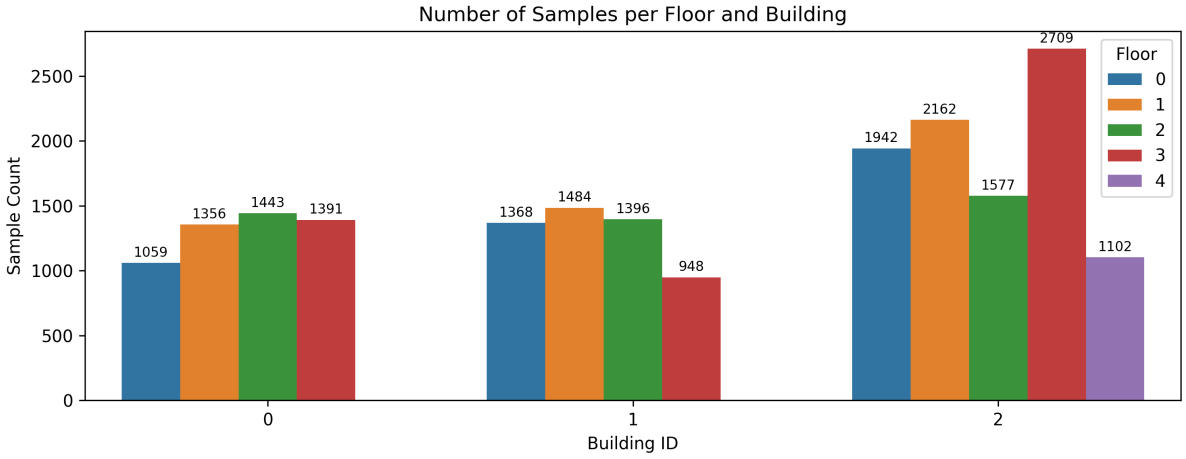

Figure 1: Distribution of the data samples over the buildings and the floors.

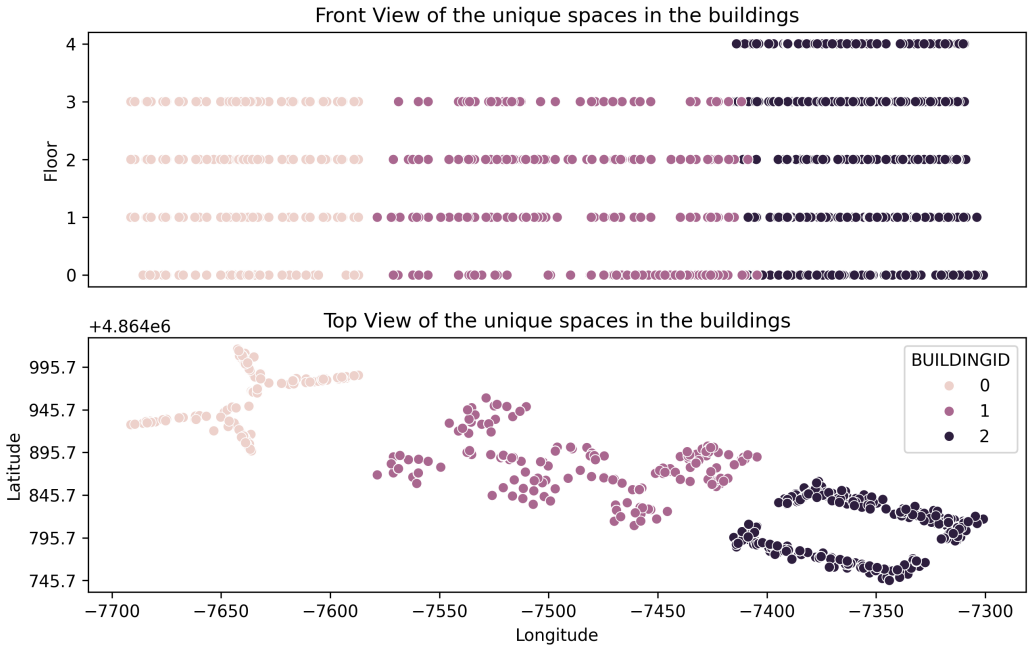

Figure 2: Front View and Top View of the unique spaces in the dataset.

Dataset 2: Myocardial infarction complications [2]

Dataset description

The **Myocardial Infarction (MI) Complications Dataset** is designed to predict potential complications following a myocardial infarction based on patient information collected (i) at the time of hospital admission and (ii) on the third day of hospitalization. The dataset comprises a total of 111 feature variables of various data types, as illustrated in Figure 3. These features include key patient characteristics such as age, sex, and relevant pre-existing conditions. Among the feature variables, 109 contain missing values, while 2 are complete. In total, the dataset consists of 1,700 patient instances.

In addition to the feature variables, the dataset provides 12 target variables. Eleven of these targets are binary, whereas one is categorical with eight distinct classes. Figure 4 presents the binary target variables, while Figure 5 shows the distribution of the categorical variable.

Furthermore, the dataset supports complication prediction at four different time points: at admission, after 24 hours, after 48 hours, and after 72 hours. Depending on the prediction time, specific subsets of the input features are available for model training.

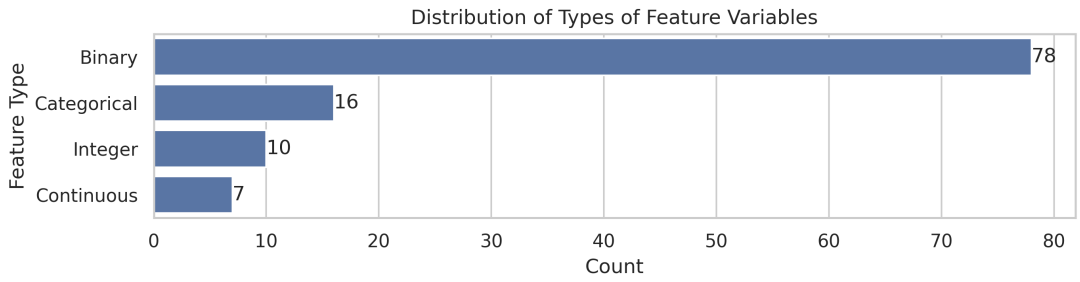

Figure 3: Feature Variable Type

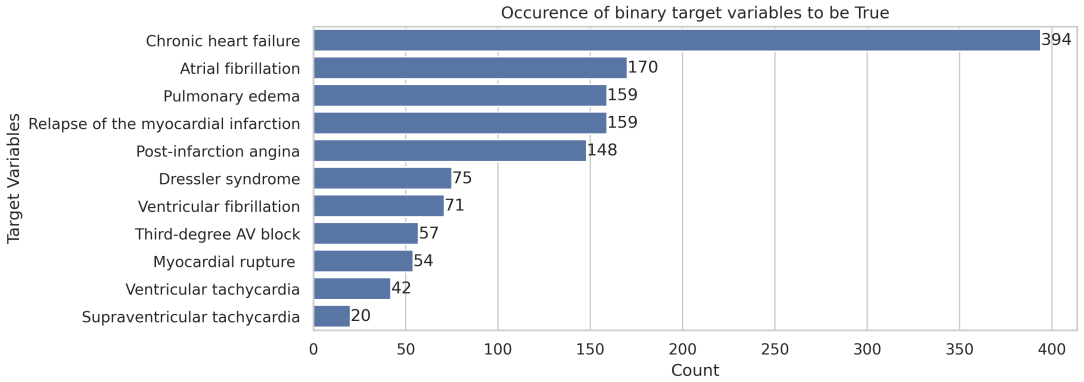

Figure 4: Target Variable Type

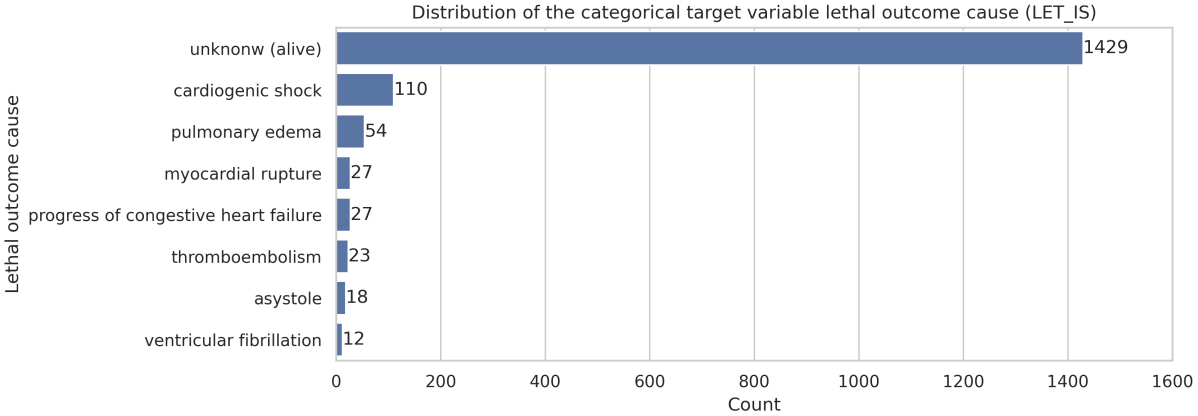

Figure 5: Leathal outcome cause

References

[1] Montoliu Raul Martnez-Us Adolfo Arnau Tomar Torres-Sospedra, Joaqun and Joan Avariento. UJIIndoorLoc. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5MS59.

[2] Shulman V.A. Rossiev D.A.-Shesternya P.A. Nikulina S.Yu. Orlova Yu.V. Golovenkin, S.E. and V.F. Voino-Yasenetsky. Myocardial infarction complications. UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C53P5M.