184.702 Machine Learning (VU 3,0) 2024S

Exercise0: Dataset Description

Ouassim Kiassa, 12113875, ouassim.kiassa@student.tuwien.ac.at
Antal Spilyka, 11907555, e11907555@student.tuwien.ac.at
Antoine Origer, 12343636, e12343636@student.tuwien.ac.at

March 23, 2024

Dataset 1 Overview

This dataset consists of housing information extracted from the 1990 California census. It comprises demographic, geographic, and economic features pertaining to housing blocks, including location, age, number of rooms and bedrooms, population, household count, median income, median house value, and proximity to the ocean.

Source: 1990 California census data.

Attributes Description

The dataset comprises a total of 20,640 observations, each described by 10 distinct attributes. Among these attributes, nine are quantitative, embodying numeric data that can be either interval or ratio in nature. There is one categorical attribute, which is nominal, indicating classification without an inherent order. No attributes in the dataset are identified as ordinal, which would signify categorical data with a defined sequence or ranking.

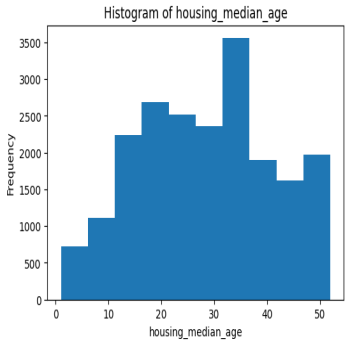

Figure 1: Distribution Housing Median Age

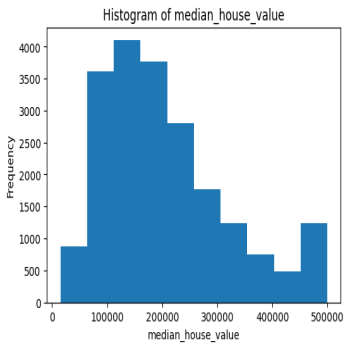

Figure 2: Distribution Media House Value

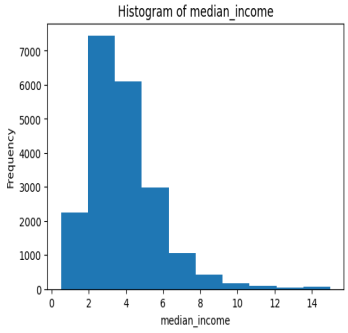

Figure 3: Distribution Median housing Income

Target Attribute

Objective: The dataset's primary application lies in the prediction of median housing values within Californian districts. This prediction is based on the analysis of the dataset's diverse features.

Count	20,640
Mean	206,855
Std	115,395
Min	14,999
25%	119,600
50%	179,700
75%	264,725
Max	500,001

Table 1: Descriptive statistics of the target attribute: Median House Value.

Dataset 2 Overview

The second dataset describes the music genres, which is a classification dataset (in comparison to the first dataset which was regression dataset), and includes the following attributes: artist name, instance id, track name, popularity, acousticness, danceability, duration (ms), energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, obtained date, valence and the target variable "music genre", which is a nominal variable. The second dataset is more complex than the first one by having more attributes, with various attribute types and more data entries.

Source: Music Genres Dataset

Attributes Description

This dataset consists of 50005 entries, out of which we have 5 entries with missing values in each attribute. The majority of the attributes are decimal numbers (ordinal data), for example popularity, acousticness and danceability. The other attributes are text attributes, such as artist name, track name and the key (nominal data). The target variable "music genre" has also the type "text" and is nominal data. Some attributes also have values such as "?" or "-1" (for example for duration). During preprocessing it will be useful to clean the completely blank entries and impute the missing values instead of "?" and "-1". During our analysis we could also identify multiple outliers for attributes (for example for loudness and tempo) throughout every class. Furthermore, we computed the correlation matrix and could identify high correlations between attributes, such as between "energy" and "acousticness" (negative correlation of -0,79), "loudness" and "energy" (positive correlation of 0,84) and "loudness" and "acousticness" (negative correlation of -0,73). Therefore, we also have to consider correlations during out preprocessing stage.

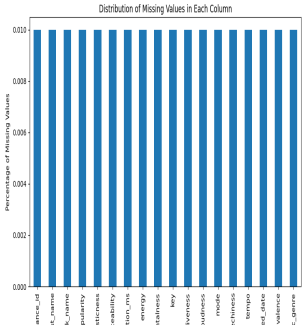

Figure 4: Distribution Missing Values (Music Genres)

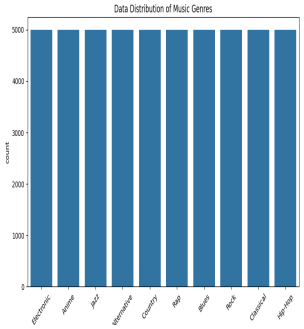

Figure 5: Distribution of Target Attribute "Music Genres"

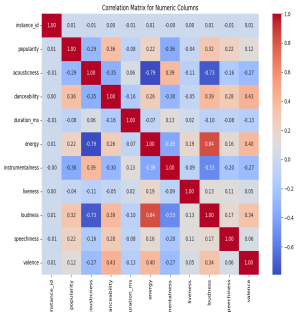

Figure 6: Correlation Matrix for Attributes (Music Genres)

Target Variable

The target variable "music genre" is evenly distributed throughout the classes. There are 10 classes and each class has 5000 entries. The classes are: Electronic, Anime, Jazz, Country, Blues, Classical, Rap, Rock, Hip-Hop and Alternative.