# **Machine Learning: Exercise 0 - Dataset Description**

**Group 36**: Nadine Kitzwögerer 11713117, Katharina Gillig 12445598, Mohammadali Dehghani 12432957

**Classification Dataset: Email-Spam-Filter**

## **Basic Information**

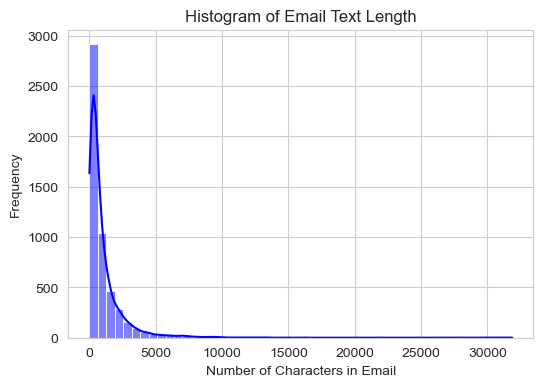
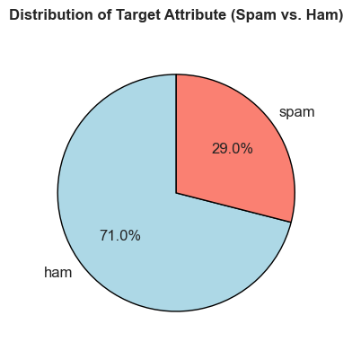
The Email Spam Filter dataset consists of 5,171 samples and 4 attributes, with two being boolean, one being an ID, and one containing free-form text. The dataset contains both numerical and categorical attributes, which require appropriate preprocessing before training a machine learning model.

### **Attribute Overview**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attribute | Type | Missing Values | Description | Preprocessing Considerations |
| Unnamed: 0 | ID | No | Distinct IDs for each row | Remove from dataset as ID should not have any significance Classification |
| label | Nominal | No | Categorical boolean variable indicating spam or ham | Needs encoding/deletion (already mapped to label\_num) |
| label\_num | Nominal | No | Numeric boolean representation of the label (0 = ham, 1 = spam) | No additional encoding needed |
| text | Nominal | No | Email content, consists of free-form text including subject | Requires NLP preprocessing (TF-IDF, tokenization, lemmatization, etc.) |

## **Distribution of Target and Input Attributes**

The dataset is imbalanced, with 71% ham (3,672 emails) and 29% spam (1,499 emails), meaning spam emails are significantly fewer. This imbalance can bias the model, causing poor spam detection.  
A closer look to the target attribute *label\_num* confirms this imbalance. Additionally, email length distribution varies, with some emails being much longer, often indicative of advertisements or phishing attempts. Histogram analysis shows that most emails contain a moderate number of characters, but spam emails tend to be longer.



**Data Considerations, Categorical Data and Other Important Aspects**

The *label\_num* column represents the target attribute for spam classification, where 0 = Ham (Non-Spam) and 1 = Spam, making it a binary classification problem with no need for scaling. However, class imbalance could be an issue if non-spam emails significantly outnumber spam, potentially biasing the model. To address this, techniques like SMOTE (oversampling), undersampling, or class weighting should be applied. The dataset also contains categorical data, including the *label* and *text* columns, which are nominal and require preprocessing such as TF-IDF, tokenization, or word embeddings to be useful in machine learning models. While the *label* is already encoded as *label\_num* and can therefore be ignored, the text content must be transformed into a numerical format using One-Hot Encoding or Label Encoding.   
The attribute *Unnamed: 0* enumerates the data from 0 to 5170 and shoud be removed/ignored in preprocessing, as it should be of no importance for the classification.

There are no missing values for *text*, although some entries only contain the subject or nothing at all, with the only content being: “Subject:”. However those entries are not missing values, due to the emails having been sent that way. This simplifies data preparation and feature engineering opportunities, such as extracting word count, punctuation usage, or special character frequency. The subject in *text* could be derived as an additional feature. Since spam detection involves class imbalance, standard accuracy is insufficient; Precision, Recall, and F1-score should be used for evaluation to ensure a reliable spam classification model.

**Explanation of Choice:** We chose the Email Spam Filter dataset because it is well-suited for binary classification tasks and presents real-world challenges such as class imbalance, text preprocessing, and feature extraction. The dataset offers a clear labeling system (label\_num: 0 = ham, 1 = spam), and while it includes free-form text requiring natural language processing, it contains no missing values, making it ideal for training and evaluating spam detection models. Additionally, the presence of both categorical and numerical data allows for a rich exploration of preprocessing techniques and evaluation metrics beyond accuracy, such as Precision, Recall, and F1-score.

**Regression Dataset: California Housing Prices**

## **Basic Information**

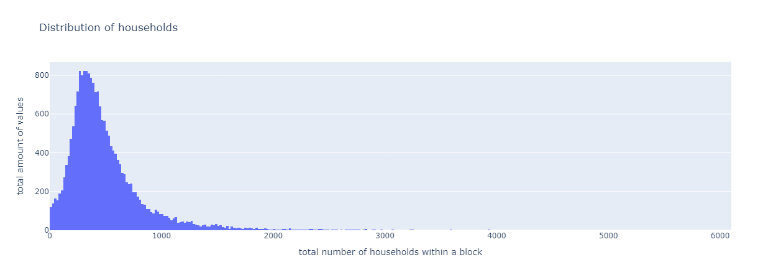
The dataset consists of 20 640 samples and 10 attributes, with 9 numerical attributes and 1 categorical attribute (*ocean\_proximity*). The target attribute is *median\_house\_value*, and the input attributes are the remaining 9 attributes listed below. Each instance in the dataset represents a housing block in California.

**Attribute Overview:**

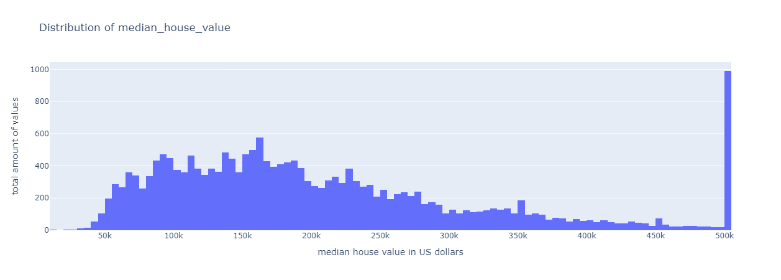
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute | Type | Value Range | Missing Values | Description | Preprocessing Considerations |
| Longitude | numerical,  continuous, ratio scale | -124.35 to -114.31 | No | Indicating the position of the housing block. | Standard preprocessing steps + check if the position is still in California. |
| Latitude | numerical,  continuous,  ratio scale | -32.54 to 41.95 | No | Indicating the position of the housing block. | Standard preprocessing steps + check if the position is still in California. |
| Housing\_median\_age | numerical,  discrete,  ratio scale | 1.0 to 52.0 | No | Median age of the houses in this block. | Standard preprocessing steps + converting the age to a median building year. |
| Total\_rooms | numerical,  discrete,  ratio scale | 2.0 to 39320.0 | No | Total number of rooms in a housing block. | Standard preprocessing steps. |
| Total\_bedrooms | numerical,  discrete,  ratio scale | 1.0 to 6445.0 | 207 | Total number of bedrooms in the housing block. | missing values need to be addressed |
| Population | numerical,  discrete,  ratio scale | 3.0 to 35682.0 | No | Total number of people living in a housing block. | Standard preprocessing steps. |
| Households | numerical,  discrete,  ratio scale | 1.0 to 6082.0 | No | Total number of households within a housing block. | Standard preprocessing steps. |
| Median\_income | numerical,  continuous,  ratio scale | 0.4999 to 15.0001 | No | Median income of a household within a housing block.  Unit: tens of thousands of US Dollars | Standard preprocessing steps + convert the values to US Dollars |
| Median\_house\_value | numerical,  continuous,  ratio scale | 14999.0 to 500001.0 | No | Median house value for houses within a block.  Unit: US Dollars | Standard preprocessing steps. |
| Ocean\_proximity | nominal,  categorical,  ordinal scale | ISLAND, NEAR BAY, NEAR OCEAN, <1H OCEAN, INLAND | No | Indicating proximity of the house to the ocean. | Check for misspellings or inconsistencies in strings, then apply one-hot encoding to convert to numerical values. |

\*Standard preprocessing steps include: Check if the values are in a reasonable range, handle outliers by considering removing or downsampling extreme values. And scaling & normalization, which is particularly important for algorithms sensitive to feature scales.

**Distribution and Histograms of Target and Input Attributes**  
For *housing\_median\_age*, we observe a clear overrepresentation of houses older than 50 years (at the time of data extraction, 1990). This might lead to bias in predictions if not addressed. A potential solution could be downsampling the older instances to balance the distribution.



The distributions of *total\_rooms*, *total\_bedrooms*, *population*, and *households* are heavily right-skewed. This suggests that the dataset contains mostly smaller housing blocks but a few large apartment complexes.  
The distribution of the target attribute *median\_house\_value* shows a peak at values around 500k or above, which could also lead to bias in predictions due to the overrepresentation of high-value housing blocks. Addressing this with techniques such as downsampling might be beneficial to prevent bias.



**Further important aspects:**

Since the value ranges of both the input and target attributes vary greatly, normalization will be crucial to avoid models being biased toward features with larger values. Moreover, some outliers (e.g., in *housing\_median\_age* and *median\_house\_value*) need to be handled appropriately, possibly with downsampling to improve model generalization. For instances with missing values in the *total\_bedrooms* attribute, it may be best to drop these rows, given the large size of the dataset. Imputing missing values could introduce inaccurate information for these housing blocks. Furthermore, attributes such as *total\_rooms* and *total\_bedrooms*, which are highly correlated and essentially represent similar information, could be candidates for removal to reduce redundancy. Additionally, new features can be created, such as the average number of people per household, which could offer valuable insights into household density within a given block and may serve as a strong indicator of housing prices, as it could reflect the socioeconomic characteristics and living conditions of the area.

**Explanation of Choice**