**Machine Learning**

**Exercise 0:Dataset Description**

Group 36:

Nadine Kitzwögerer 11713117  
Katharina Gillig 12445598  
Mohammadali Dehghani 12432957

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

## 1. Overview of Attribute Types

The Email Spam Filter dataset consists of 5,171 samples and 4 attributes, classified into different types based on their properties. The dataset contains both numerical and categorical attributes, which require appropriate preprocessing before training a machine learning model.

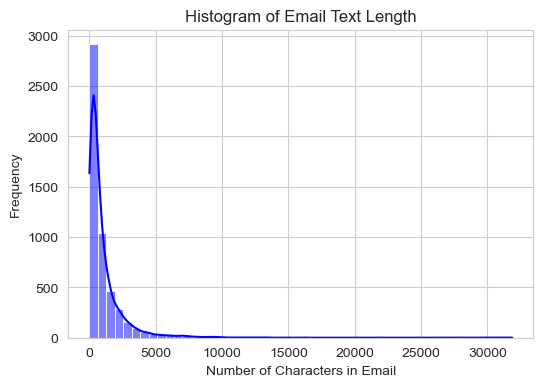
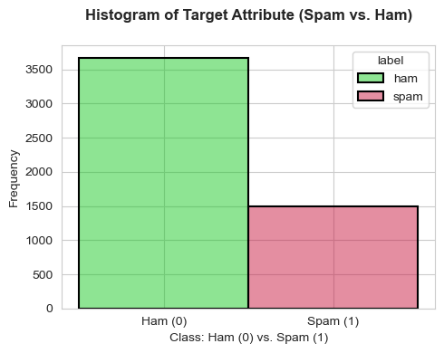
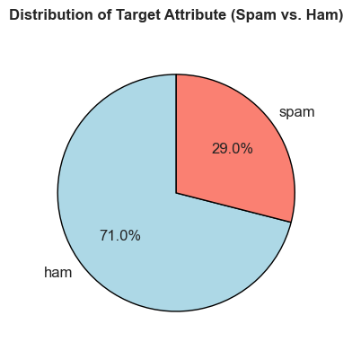
### Types of Attributes

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Type | Description | Preprocessing Considerations |
| label | Nominal | Categorical variable indicating spam or ham | Needs encoding (already mapped to label\_num) |
| text | Nominal | Email content, consists of free-form text | Requires NLP preprocessing (TF-IDF, tokenization, lemmatization, etc.) |
| label\_num | Nominal/Ordinal | Numeric representation of the label (0 = ham, 1 = spam) | No additional encoding needed |

## 2. Distribution and Histograms of Target and Input Attributes

### Target Attribute Distribution (Spam vs. Ham)

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. Solutions include oversampling (SMOTE), undersampling, or class-weighted models to improve performance. A histogram of the target attribute (label\_num) confirms this imbalance, indicating that without resampling, the model may favor predicting "ham." 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. Text preprocessing techniques like TF-IDF and word embeddings help extract meaningful features, improving spam classification accuracy.



## 3. Numeric and Categorical 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', the text content must be transformed into a numerical format using One-Hot Encoding or Label Encoding. Additionally, there are no missing values, simplifying data preparation, and feature engineering opportunities exist, such as extracting word count, punctuation usage, or special character frequency. 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.

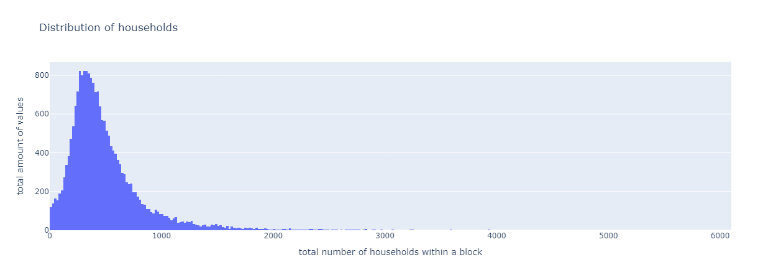
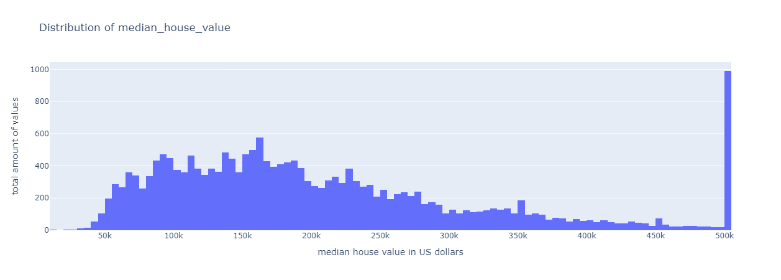
**Regression Dataset: California Housing Prices**

## 1.Overview of the Data

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.  
Types of Attributes

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Type | Description | Preprocessing Considerations |
| Longitude | Numerical (continuous) | Indicating the position of the housing block. | Standard preprocessing steps + check if the position is still in California. |
| Latitude | Numerical (continuous) | Indicating the position of the housing block. | Standard preprocessing steps + check if the position is still in California. |
| Housing\_median\_age | Numerical (discrete) | Median age of the houses in this block. | Standard preprocessing steps + converting the age to a median building year. |
| Total\_rooms | Numerical (discrete) | Total number of rooms in a housing block. | Standard preprocessing steps. |
| Total\_bedrooms | Numerical (discrete) | Total number of bedrooms in the housing block. | 207 missing values need to be addressed |
| Population | Numerical (discrete) | Total number of people living in a housing block. | Standard preprocessing steps. |
| Households | Numerical (discrete) | Total number of households within a housing block. | Standard preprocessing steps. |
| Median\_income | Numerical (continuous) | 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) | Median house value for houses within a block.  Unit: US Dollars | Standard preprocessing steps. |
| Ocean\_proximity | Nominal (categorical) | 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.

2.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.

3.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.