1. CLEANML DATASETS

Citation: This dataset [1] consists of titles of 5,005 publications from Google Scholar and DBLP. Given a publication title, the classification task is to determine whether the paper is related to Computer Science or not. This dataset contains duplicates.

EEG: This is a dataset[4] of 14,980 EEG recordings with 14 EEG attributes. The classification task is to predict whether the eye-state is closed or open. This dataset contains numerical outliers.

Marketing: This dataset [6] consists of 8,993 records about household income from a survey. Each record has 14 demographic attributes including sex, education, etc. The classification task is to predict if the annual household income is less than $25,000. This dataset contains missing values.

Movie: This dataset [10, 5] consists of 9,329 movie reviews, which we obtained by merging data from IMDB and TMDB datasets. Each record has seven attributes including title, language, score, etc. The classification task is to predict the genre of the movie (romance or comedy). It contains duplicates and inconsistent representations of languages.

Company: The original dataset [2] contains over 2.5 million records. We randomly sampled 5% records (128,889 records) from the original dataset. Each record has seven attributes including company name, country, city, etc. The classification task is to predict whether the public sentiment about a company is negative or not. This dataset contains inconsistent company names.

Restaurant: This dataset [7] contains 12,007 records about restaurants, which we obtained by merging data from the Yelp and Yellowpages datasets. Each record has 10 attributes including city, category, rating, etc. The classification task is to predict whether the price range of a restaurant is "$" or not. This dataset contains duplicates and inconsistent restaurant names and categories.

Titanic: This dataset [9] contains 891 records and 11 attributes from the Titanic including name, sex, etc. The task is to determine whether the passenger survived or not. This dataset has a significant number of missing values.

Credit: This dataset [3] consists of 150,000 credit records with 10 attributes including monthly income, age, then number of dependents, etc. The classification task is to predict whether a client will experience financial distress in the next two years. This dataset has a class imbalance problem with only 6.7% records in the minority class. We follow standard procedure to over-sample the minority and down-sample the majority class before training, and we use F1 score as the performance metric for evaluation. This dataset contains missing values and numerical outliers.

Sensor: The original [8] sensor dataset contains 928,991 sensor recordings with eight attributes including temperature, humidity, light, etc. We only used recordings from sensor 1 and sensor 2 and sampled the dataset to include 1 observation per hour for each sensor. The sampled dataset contains 62,076 records. The classification task is to predict whether the readings came from a particular sensor (sensor 1 or sensor 2). This dataset contains outliers.

University: This dataset [11] contains 286 records about universities. Each record has 17 attributes including state, university name, SAT scores, etc. The classification task is to predict whether the expenses are greater than $7,000$ for each university. This dataset contains inconsistent representations for states and locations.

USCensus: This dataset [12] contains 32,561 US Census records for adults. Each record has 14 attributes including age, education, sex, etc. The classification goal is to predict whether the adult earns more than $50,000. This dataset contains missing values.

Airbnb: This is our own dataset with 42,492 records on hotels in the top 10 tourist destinations and major US metropolitan areas, scraped from Airbnb.com. Each record has 40 attributes, including the number of bedrooms, price, location, etc. Demographic and economic attributes were scraped from city-data.com. The classification task is to determine whether the rating of each hotel is 5 or not. This dataset contains missing values, numerical outliers, and duplicates. We will release this dataset in the code repository.

2. REFERENCES

[1] Citation dataset. `https://sites.google.com/site/anhaidgroup/useful-stuff/data`. Accessed: June 19, 2020.

[2] Company dataset. `https://www.kaggle.com/jacksapper/company-sentiment-by-location`. Accessed: June 19, 2020.

[3] Credit dataset. `https://www.kaggle.com/c/GiveMeSomeCredit/data`. Accessed: June 19, 2020.

[4] EEG dataset. `https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State`. Accessed: June 19, 2020.

[5] IMDB movie dataset. `https://data.world/popculture/imdb-5000-movie-dataset`. Accessed: June 19, 2020.

[6] Marketing dataset. `https://sites.google.com/site/anhaidgroup/useful-stuff/data`. Accessed: June 19, 2020.

[7] Restaurant dataset. `https://sites.google.com/site/anhaidgroup/useful-stuff/data`. Accessed: June 19, 2020.

[8] Sensor dataset. `http://db.csail.mit.edu/labdata/labdata.html`. Accessed: June 19, 2020.

[9] Titanic dataset. `https://www.kaggle.com/upendr/titanic-machine-learning-from-disaster/data`. Accessed: June 19, 2020.

[10] TMDB movie dataset. `https://www.kaggle.com/tmdb/tmdb-movie-metadata`. Accessed: June 19, 2020.

[11] University dataset. `https://archive.ics.uci.edu/ml/datasets/University`. Accessed: June 19, 2020.

[12] USCensus dataset. `https://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29`. Accessed: June 19, 2020.