ASDS 5305 Final Project Datasets

2/7/2025

Group Name: sp25_BerKyd

Group Members:

- Henry Berrios #1001392315
- LeMaur Kydd #1001767382

Link to Google Colab
Link to GitHub Repository

Dataset 1

Title: Drug SMILES Strings and Classifications

Source: Research Paper Supporting Information SMILES LINK

Machine Learning Problem Description: SMILES strings are a text encoding of a molecules chemical structure. The abilities and characteristics (stereochemistry, etc.) of each molecule are preserved within the SMILES string, therefore it is useful for Natural Language Processing (NLP) based tasks for drug discovery, classification, etc. The task is to train a model on the tensorized versions of SMILES strings and their drug classifications (antiinfective, antineoplastic, etc.) to reliably generalize and classify unseen SMILES strings into their proper drug classifications.

Dataset Criteria Satisfaction: This dataset satisfies all the dataset requirements. There are 6935 data points spanning across 12 drug classifications; however, the 8 smallest classes will be combined into one class called 'other' to preserve and improve the balance of the target variable, thus satisfying the "**sufficient number of samples**" and "**balanced**" criteria. This dataset essentially no pre-processing and will be simple to tokenize and tensorize using existing packages from TensorFlow and PyTorch, thus satisfying the "**pre-processing**" and "**tensor-ready formats**".

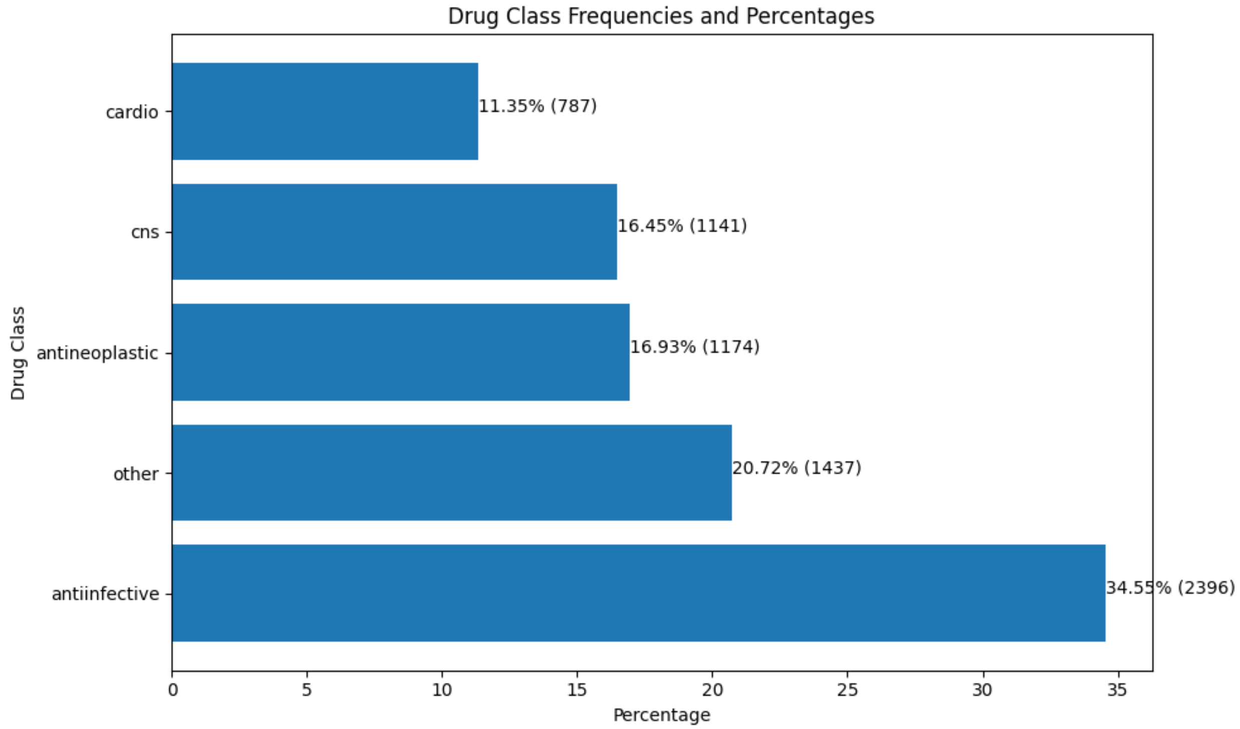
Domain: Cheminformatics

Modality: Text

Goal: Classification of Drug Type

Problem Type: Classification

Target Labels: {antiinfective, antineoplastic, cns, cardio, other}



Dataset Citation: Meyer, Jesse G.; Liu, Shengchao; Miller, Ian J.; Coon, Joshua J.; Gitter, Anthony (2019). Learning Drug Functions from Chemical Structures with Convolutional Neural Networks and Random Forests. ACS Publications. Collection. LINK

Dataset Licensing: Free & Open Usage

Dataset 2

Title: 911 Recordings: The First 6 Seconds

Source: Kaggle LINK

Machine Learning Problem Description: For this dataset the task would be to train a model on the tensorized audio .wav file and their respective potential_death target variable which outlines if the 911 call described a situation in which people's lives are in danger. This combination of training data would hopefully result in the model being able to make a proper fit on the data and be able to generalize on new .wav files and make reliable predictions.

Dataset Criteria Satisfaction: This dataset satisfies the dataset requirements with one criteria that will be a challenge to work with. The dataset has 710 .wav files which is a good sample size for this type of task, considering we are not attempting to generate audio or any other data-costly task, therefore meeting the “**sufficient number of samples”** criteria. The challenging criteria to meet is having “**balanced**” data. The binary target variables distribution is ~85% (1.0 – Serious incident) and ~14% (Not Serious Incident). We will have to implement a form of either under or oversampling techniques or other techniques common with audio data to ensure this criterion is met. This dataset will not require any major data preprocessing as the audio files are already edited down to a standard duration, satisfying the “**pre-processing**” requirement. The .wav files have a standard tensorizing pipeline that we have sourced. The .wav files once loaded are already in vector format when using the scipy.io.wavefile library function, so the “**tensor-ready formats**” requirement is satisfied.

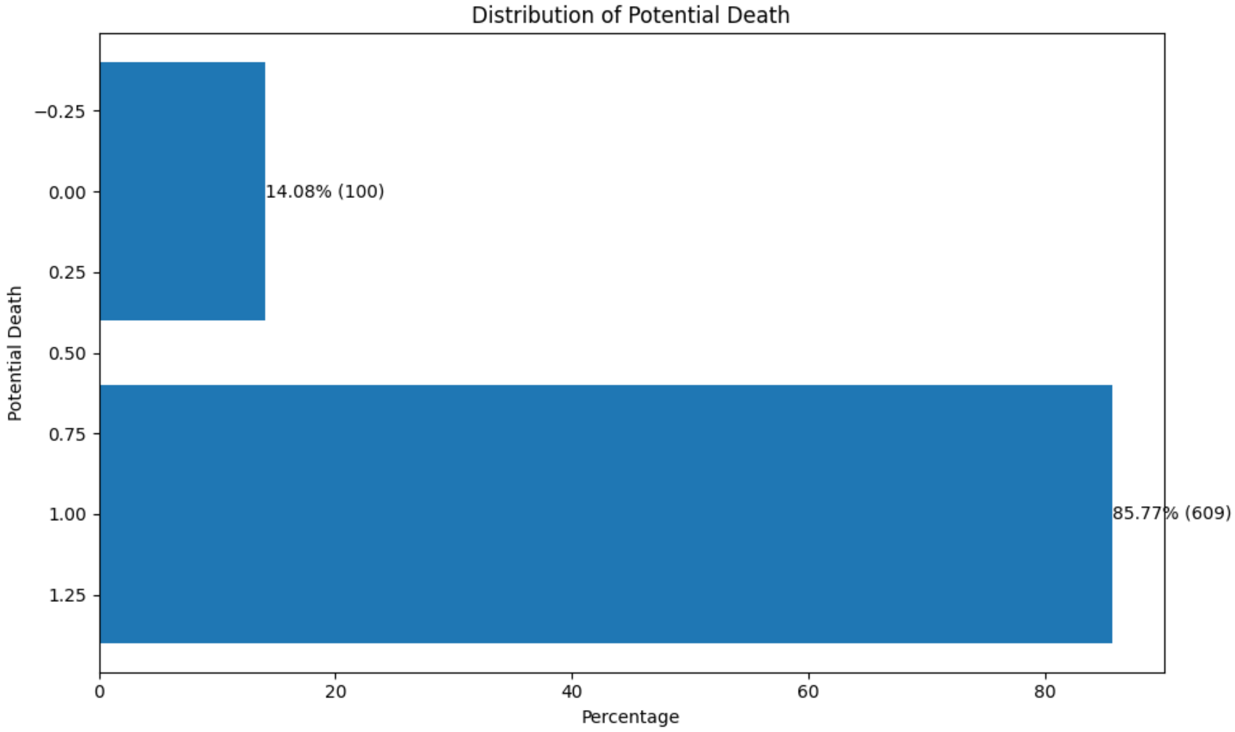
Domain: Police & Emergency Services

Modality: Audio

Goal: Binary Classification of Call Severity

Problem Type: Binary Classification

Target Labels: {1.0 – Serious incident/lives in danger, 0.0 – Not Serious/No lives in danger}



Dataset Citation: LINK

Dataset Licensing: CC0: Public Domain

Dataset 3

Title: TBI-NDSC

Source: Traumatic Brain Injury Model Systems National Data and Statistical Center LINK

Machine Learning Problem Description: The TBI-NDSC dataset provides clinical and demographic information about patients who have suffered Traumatic Brain Injuries (TBI). This dataset captures a mixture of patient data, including initial assessments at admission, follow-up outcomes at multiple time points, and various clinical factors related to recovery. Instead of using longitudinal data, we are focusing on predicting the Disability Rating Scale (DRS) score at discharge (DRSd) using only the initial admission dataset ('TBIMSForm1_Public_20240405.csv'). The goal is to develop a deep learning regression model that can accurately predict functional recovery at discharge based on a patient's demographic, clinical and injury-related variables. Overall, this model will help in early prognosis estimation for TBI patients, assisting clinicians in treatment planning and rehabilitation resource allocation.

Dataset Criteria Satisfaction: The dataset meets the requirements for deep learning applications. Initially it contained approximately ~19,000 patients but after filtering admission data and handling missing or unavailable values, the final dataset includes ~8,000 patients, satisfying the **"sufficient number of samples**" requirement. Since the Disability Rating Scale at discharge (DRSd) is a continuous variable, the problem type is framed as a *regression task*. Missing values and placeholder codes in the dataset have been addressed. Categorical variables were already stored in numerical format, eliminating the need for any additional encoding steps. The dataset finally satisfies the "**pre-processing"** requirement and is in a "**tensor-ready format"**.

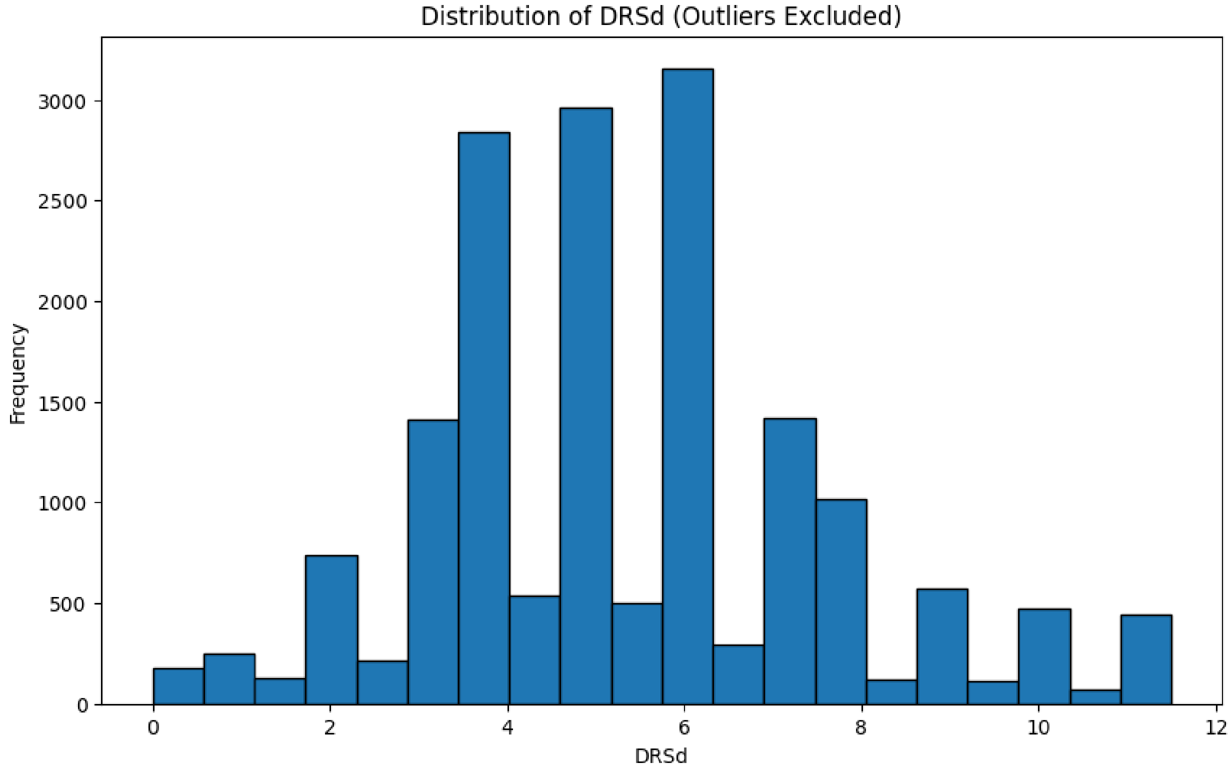
Domain: Health Informatics

Modality: Tabular

Goal: Regression of Disability Rating Scale (DRSd)

Problem Type: Regression

Target Labels: Disability Rating Scale (DRS) {2.0, 4.5, 30.0, 1.0}



Dataset Citation: 10.17605/OSF.IO/A4XZB

Dataset Licensing: Public Use By Data Request