00_Introduction

June 8, 2025

1 Supervised Modeling Pipeline for ICU Length of Stay Prediction

Authors: *Giuseppe Pitruzzella, Radvilė Rušaitė, Karlota Bochanaitė*

This project is situated within the field of supervised learning, aiming to predict a continuous clinical variable—**Length of Stay (LOS)** in the Intensive Care Unit—using regression models based on neural networks. The approach aligns with fundamental machine learning paradigms studied during the course, including probabilistic optimization techniques such as Maximum A Posteriori (MAP) estimation.

The analysis uses real-world data from **MIMIC-III** (Medical Information Mart for Intensive Care), a publicly available clinical database developed by the MIT Lab for Computational Physiology in collaboration with Beth Israel Deaconess Medical Center (Boston). The database contains de-identified health-related data associated with over 60,000 ICU admissions between 2001 and 2012, and has become a globally recognized benchmark for research in medical data science and critical care.

Dataset Setup To facilitate reproducibility, all necessary .csv files from the MIMIC-III clinical dataset have been kindly provided by the course instructor in compressed format (.csv.gz). The files have been made available via an educational mirror for use in this project.

Environment Setup Before proceeding with the analysis, make sure all required Python libraries are available. You can automatically install missing dependencies listed in `requirements.txt` by executing the following cell.

[ ]:
```
import subprocess
import sys
import importlib.util

# Path to the requirements.txt file
req_file = "../requirements.txt"  # modify if necessary

def read_requirements(file_path):
    with open(file_path, "r") as f:
        return [line.strip().split("==")[0] for line in f if line.strip() and␣
 ↪not line.startswith("#")]

def is_installed(package_name):
```

```
    return importlib.util.find_spec(package_name) is not None

def pip_install(package_name):
    print(f"Installing: {package_name}")
    subprocess.check_call([sys.executable, "-m", "pip", "install",␣
 ↪package_name])

# Map exceptions between package name and importable module
module_map = {
    "scikit-learn": "sklearn",
    "ipython": "IPython"
}

# Load packages from requirements.txt
required_packages = read_requirements(req_file)

# Install only missing packages
for pkg in required_packages:
    module_name = module_map.get(pkg, pkg)
    if not is_installed(module_name):
        pip_install(pkg)
    else:
        print(f"Already installed: {pkg}")
```

```
Already installed: torch
Already installed: pandas
Already installed: matplotlib
Already installed: seaborn
Already installed: numpy
Already installed: scikit-learn
Already installed: xgboost
Already installed: ipython
```

1.1 Initial Exploration

We begin by exploring the main tables in the MIMIC-III dataset. The goal of this first phase is to understand the structure of the database and identify relevant variables that may influence ICU length of stay. Key reference tables include:

- `D_ICD_DIAGNOSES.csv` – diagnosis codes (ICD9)
- `ICUSTAYS.csv` – ICU stays metadata
- `D_ITEMS.csv` – item IDs and descriptions for time-series events

[ ]:
```
# Install needed packages
!apt-get install texlive texlive-xetex texlive-latex-extra pandoc &> /dev/null
!pip install pypandoc &> /dev/null

# Mount your google drive to get access to your ipynb files
```

```
from google.colab import drive
drive.mount('/content/drive')

# and copy your notebook to this colab machine. Note that I am using *MY*␣
 ↪notebook filename

!cp "/content/drive/MyDrive/Colab Notebooks/00_Introduction.ipynb" ./ &> /dev/
 ↪null

# Then you can run the converter.

!jupyter nbconvert --to PDF "00_Introduction.ipynb" &> /dev/null
```