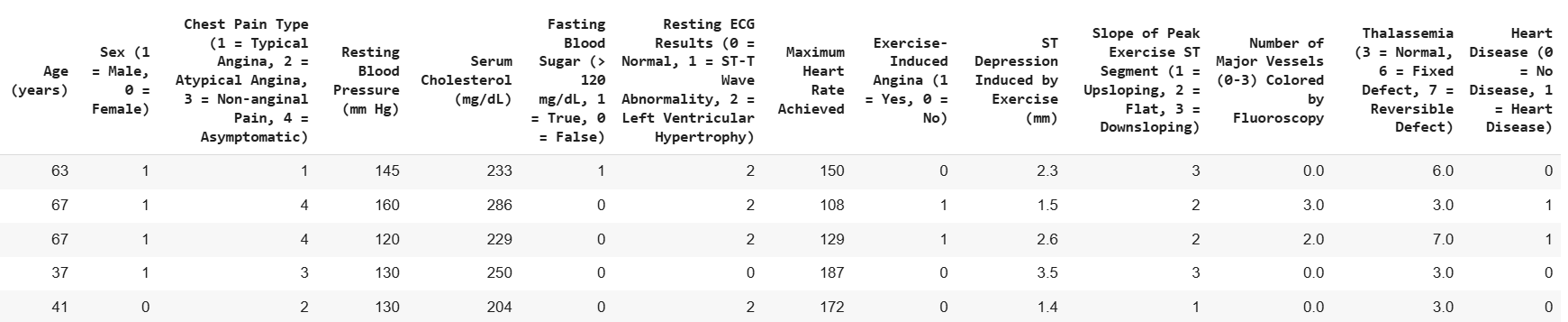
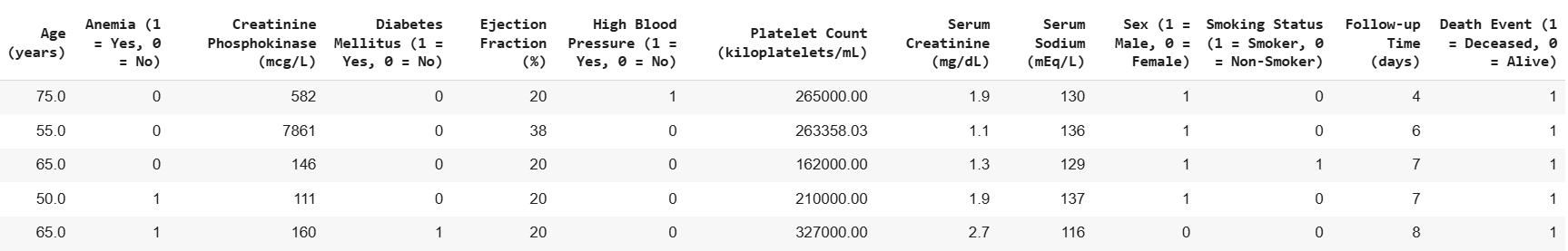
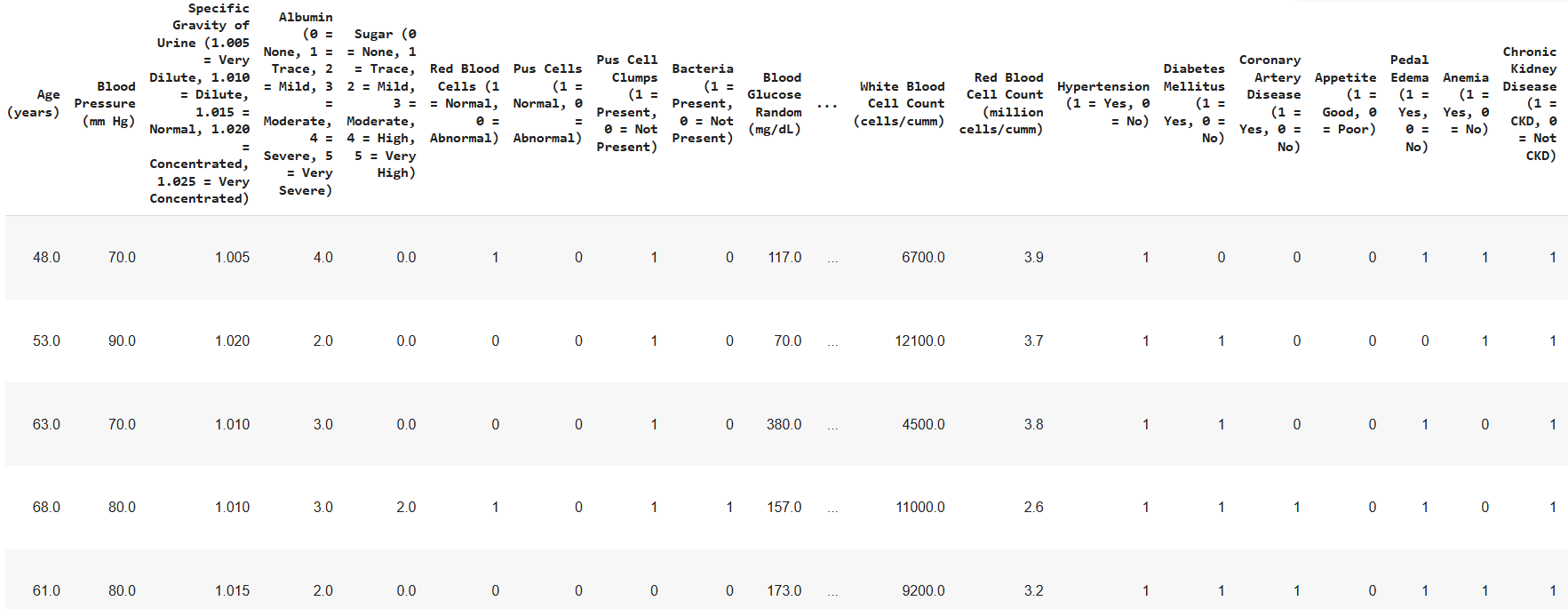
**UCI Heart Disease Dataset**



**UCI Heart Failure Clinical Records Dataset**

****

**UCI Chronic Kidney Disease Dataset**

****

**Dataset Description**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Train (Healthy Class)** | **Train (Positive Class)** | **Test** |
| UCI HD | 128 | 118 | 51 |
| UCI HFCR | 162 | 87 | 50 |
| UCI CKD | 92 | 34 | 32 |

**Tabular to Clinical Text for UCI Heart Disease Dataset**

Samples for Heart Disease Data:

Row 1:

Patient is a male aged 54 years, presenting with chest pain type 4, resting blood pressure 124 mm Hg, serum cholesterol 266 mg/dL, normal fasting blood sugar, resting ECG showing result Left Ventricular Hypertrophy, maximum heart rate achieved 109, with exercise-induced angina, ST depression of 2.2 mm, Slope of Peak Exercise ST Segment: Flat, 1.0 major vessels affected, thalassemia type Reversible Defect.

--------------------------------------------------------------------------------

Row 2:

Patient is a male aged 58 years, presenting with chest pain type 4, resting blood pressure 150 mm Hg, serum cholesterol 270 mg/dL, normal fasting blood sugar, resting ECG showing result Left Ventricular Hypertrophy, maximum heart rate achieved 111, with exercise-induced angina, ST depression of 0.8 mm, Slope of Peak Exercise ST Segment: Upsloping, 0.0 major vessels affected, thalassemia type Reversible Defect.

--------------------------------------------------------------------------------

Row 3:

Patient is a male aged 56 years, presenting with chest pain type 4, resting blood pressure 125 mm Hg, serum cholesterol 249 mg/dL, elevated fasting blood sugar, resting ECG showing result Left Ventricular Hypertrophy, maximum heart rate achieved 144, with exercise-induced angina, ST depression of 1.2 mm, Slope of Peak Exercise ST Segment: Flat, 1.0 major vessels affected, thalassemia type Normal.

--------------------------------------------------------------------------------

Row 4:

Patient is a male aged 50 years, presenting with chest pain type 4, resting blood pressure 150 mm Hg, serum cholesterol 243 mg/dL, normal fasting blood sugar, resting ECG showing result Left Ventricular Hypertrophy, maximum heart rate achieved 128, without exercise-induced angina, ST depression of 2.6 mm, Slope of Peak Exercise ST Segment: Flat, 0.0 major vessels affected, thalassemia type Reversible Defect.

**Synthetic Data for Heart Disease by Paraphrasing (UCI Heart Disease Dataset)**

*(GPT-4)*

*Prompt:*

“Paraphrase the provided clinical text to create diverse synthetic samples while preserving medical relevance. Do not alter any values or their associated units. Rearrange the order of the information in each new version to increase variety. Keep all paraphrased texts concise, medically accurate, and easy to understand.”

Synthetic Samples for above rows:

Row 1:

A 54-year-old male with chest pain type 4 shows ST depression of 2.2 mm and a flat slope on peak exercise. He has a resting blood pressure of 124 mm Hg and serum cholesterol at 266 mg/dL. Fasting blood sugar is normal. Resting ECG indicates Left Ventricular Hypertrophy. Maximum heart rate achieved is 109, and he experiences exercise-induced angina. One major vessel is affected, and thalassemia is of the Reversible Defect type.

--------------------------------------------------------------------------------

Row 2:

Resting ECG shows Left Ventricular Hypertrophy in this 58-year-old male with chest pain type 4. Blood pressure is 150 mm Hg, and cholesterol level is 270 mg/dL. Fasting blood sugar is normal. The patient reached a maximum heart rate of 111 and had exercise-induced angina. ST depression is 0.8 mm with an upsloping segment. No major vessels are affected, and thalassemia is of the Reversible Defect type.

--------------------------------------------------------------------------------

Row 3:

A male patient with elevated fasting blood sugar and ECG showing Left Ventricular Hypertrophy is 56 years old. He presents with type 4 chest pain and exercise-induced angina. His maximum heart rate reached 144, and the ST segment is flat with a depression of 1.2 mm. He has one major vessel affected, a resting blood pressure of 125 mm Hg, serum cholesterol level of 249 mg/dL, and his thalassemia type is Normal.

--------------------------------------------------------------------------------

Row 4:

A 50-year-old male with chest pain type 4 has a resting blood pressure of 150 mm Hg and cholesterol at 243 mg/dL. ST depression is 2.6 mm with a flat slope, and maximum heart rate reached is 128. Fasting blood sugar is normal. No exercise-induced angina was reported. ECG shows Left Ventricular Hypertrophy, no major vessels are affected, and thalassemia is of the Reversible Defect type.

**Results for UCI HD Dataset:**

*GPT2 as Classification Model*

For the GPT-2 model, we evaluated its performance on raw tabular data formatted as sequences, using different sampling strategies to better represent the majority class. We also assessed its performance when the tabular data was converted into clinical text without applying paraphrasing techniques to enhance the representation of the majority class. These settings were included as part of our ablation study.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Sampling Technique** | **Recall for Healthy Class (%)** | **Recall for Heart Disease Class (%)** |
| ANN | CopulaGAN | 0.41 | 0.63 |
| ANN | TVAE | 0.12 | 0.95 |
| ANN | CTGAN | 0.16 | 1 |
| CNN | SMOTE-NC | 0.19 | 1 |
| CNN | CTGAN | 0.25 | 1 |
| **GPT2 (Input Sequence)** | TVAE | 0.38 | 0.95 |
| RF | CopulaGAN | 0.47 | 0.89 |
| ANN | SMOTE-NC | 1 | 0.37 |
| MLP | CopulaGAN | 0.5 | 0.89 |
| CNN | CopulaGAN | 0.41 | 1 |
| Transformer (Input Sequence) | CTGAN | 0.44 | 1 |
| CNN | TVAE | 0.5 | 0.95 |
| DT | SMOTE-NC | 0.66 | 0.79 |
| AdaBoost | CTGAN | 0.94 | 0.53 |
| MLP | SMOTE-NC | 0.59 | 0.89 |
| MLP | TVAE | 0.53 | 0.95 |
| DT | TVAE | 0.66 | 0.84 |
| RF | TVAE | 0.66 | 0.84 |
| RF | CTGAN | 0.62 | 0.95 |
| Transformer (Input Sequence) | CopulaGAN | 0.62 | 0.95 |
| AdaBoost | TVAE | 0.78 | 0.79 |
| Transformer (Input Sequence) | TVAE | 0.69 | 0.89 |
| Transformer (Input Sequence) | SMOTE-NC | 0.66 | 0.95 |
| **GPT2 (Input Sequence)** | CTGAN | 0.94 | 0.68 |
| MLP | CTGAN | 0.78 | 0.84 |
| RusBoost | TVAE | 0.88 | 0.74 |
| DT | CopulaGAN | 0.69 | 0.95 |
| RF | SMOTE-NC | 0.75 | 0.89 |
| DT | CTGAN | 0.75 | 0.89 |
| RusBoost | SMOTE-NC | 0.88 | 0.77 |
| **GPT2 (Input Sequence)** | SMOTE-NC | 0.97 | 0.68 |
| AdaBoost | SMOTE-NC | 0.88 | 0.79 |
| GNB | TVAE | 0.72 | 0.95 |
| RusBoost | CopulaGAN | 0.88 | 0.79 |
| AdaBoost | CopulaGAN | 0.88 | 0.79 |
| **GPT2 (Table to Text)** | Raw | 0.88 | 0.79 |
| GNB | CTGAN | 0.84 | 0.84 |
| GNB | CopulaGAN | 0.75 | 0.95 |
| **GPT2 (Input Sequence)** | CopulaGAN | 0.91 | 0.79 |
| GNB | SMOTE-NC | 0.78 | 0.95 |
| RusBoost | CTGAN | 0.94 | 0.89 |
| **GPT2 (Table to Text)** | **GPT4** | 0.91 | 0.95 |

Similar results and improvements were also found for the other two datasets.