Data-Centric-AI-Community/nist-crc-2023

github.com/Data-Centric-AI-Community/nist-crc-2023

Data-Centric-AI-Community

DISCORD

YOUTUBE

MEDIUM

YDATA SYNTHETIC

NIST Privacy Collaborative Reseach

The Data-Centric AI Community just launched a small community project to experiment with the NIST Challenge!

- **Goal:** To learn about Synthetic Data and how it can be used to prepare sensitive private data for public release!
- **Dates:** From April to July. You can also join at any time, follow the weekly plan, and post questions on our Discord.
- **Where:** □***-nist-challenge*** channel in our Discord Server
- **Touch Points:** We meet every Friday around 4 PM GTM on the □***-code-with-me channel*** to discuss the project.

Overview

□ The **overall goal** of the project is to **explore synthetic data** to prepare sensitive private data for public release.

□ NIST has launched a **benchmark of 3 datasets**, `MA`, `TX` (Texas), and `NATIONAL` which you can use in the project.

□ To provide an **evaluation of the de-identified data** against the target/real data, NIST has created the `sdnist` package that can be installed according to the instructions below.

□ To **create the de-identified data**, we'll use `ydata-synthetic` package, explore different model settings and study the effect this has on the final results.

Learning Outcomes

Week	What you will learn
1	**Goal and objectives of the project.** You'll connect with other learners in the DCAI Discord Server and be added to the NIST Team to access the **-nist-challenge** channel and receive permissions to collaborate on the GitHub project.
2	**Basics of Synthetic Data.** You will learn more about what is synthetic data, how is it generated, what are the main applications.
3	**Basics of Data Profiling.** You will learn what is data profiling, how to understand your data with descriptive statistics, and what are common data quality issues. You will also **explore the NIST datasets** with `ydata-profiling` and **preprocess the data** according to your findings.
4 & 5	**Generation of Synthetic Data.** You will explore Deep Learning models (Generative Adversarial Networks -- GAN) to generate realistic synthetic data using `ydata-synthetic`.
6 & 7	**Basics of Evaluating Synthetic Data.** You will explore some strategies to evaluate synthetic data and investigate possible improvements to your solution. We will explore the `sdnist` package to evaluate our synthetic data.
8	**Project Showcase.** You will learn how to best **showcase and publicize your project** in your data portfolio, CV, GitHub, or Medium Account.

Tasks

Week 1:

- Read the instructions and information about the challenge
- Learn about the benchmark data released -- The NIST Diverse Communities Data Excerpts
- Post questions and ideas on the *-nist-challenge* channel

Week 2:

Week 3:

- Learn about the basic aspects of **Data Profiling**:
  - Auditing Data Quality with ydata-profiling: learn about what is **data profiling**, what common **data quality issues** we find in real-world domains (*can you spot a few in the NIST datasets?*), and how `ydata-profiling` can help you diagnose and overcome them
  - Awesome Data Science Tools to Master in 2023: Data Profiling Edition: learn more about **data profiling** and existing **open source tools** to understand your data to the fullest!
  - Auditing Data Quality with YData Profiling: an overview of `ydata-profiling` functionalities and how-to's
- Start profiling the NIST data:
  - Install`ydata-profiling` (check the **Installation Instructions** below) and **don't forget to star it, thank you!**
  - Choose one of the NIST datasets (`MA`, `TX`, or `NATIONAL`):
    - The datasets are available here
    - Run a **Profile Report** on your data (check the **Installation Instructions** below)
    - Create an excel file to register your learnings. **Suggestion for the columns:** `Feature Name | Data Type (Numeric/Categorical) | Missing Values (Y/N) | Notes/Observations`. Your observations should be based on the profiling report, but also on the description of the features provided
- Post questions and comments on the *-nist-challenge* channel.
- Meet us on Friday (**May 12**) to discuss what you've learned (*check the available slots on our Discord Calendar*). **Don't forget to bring your excel file with the data description and your profiling report!**

Weeks 4 & 5:

- Start experimenting with `ydata-synthetic` (check the **Installation Instructions** below and **don't forget to star it, thank you!**). If you prefer a UI experience, you can also **leverage the Streamlit App** in version 1.0.0:
  - How to "pip install ydata-synthetic" without errors!
  - Install ydata-synthetic in 5 min
  - How to Generate Real-World Synthetic Data with CTGAN
  - How to Generate Synthetic Data with ydata-synthetic's Streamlit app

- Compare your **synthetic** data with the **real** data using the `.compare()` functionality of `ydata-profiling`:
    - How to compare 2 datasets with ydata-profiling. *What are the obtained results? Are there any aspects that you can improve?*
- Post questions and comments on the *-nist-challenge* channel! You can upload your profiling reports the the channel so that we can discuss changes and improvements.

Installation Instructions

▶ **How to create and use Virtual Environments?**

▶ **How to install ydata-profiling and create a Profiling Report?**

▶ **How to install ydata-synthetic and create a synthesizer?**