Differential Privacy in HealthCare

Group 3
Madhav Sood(IMT2021009)
Nilay Kamat(IMT2021096)
Shlok Agrawal(IMT2021103)

Dataset Used

Predicting Show-Up/No-Show

This is a Kaggle dataset to implement ML models that can determine whether a patient shows up to their appointment based on attributes such as Age, Gender, Neighbourhood, Hypertension, Diabetes, Alcoholism, etc.

Attributes

PatientId	- Directly Identifiable
AppointmentID	- Quasi-Identifier
ScheduledDay	- Quasi-Identifier
AppointmentDay	- Quasi-Identifier
Gender	- Quasi-Identifier
Age	- Quasi-Identifier
Hypertension	- Sensitive
Diabetes	- Sensitive
Alcoholism	- Sensitive
Handicap	- Sensitive
No-show	- Label to be predicted

Pre-Processing

- No null values are present in the dataset.
- Label encoded PatientId, AppointmentID, Gender and No-show.
- Split ScheduledDay and AppointmentDay to ScheduledYear, ScheduledMonth and ScheduledDate and AppointmentYear, AppointmentMonth and AppointmentDate respectively.
- Dropped 3 columns - ScheduledDay, AppointmentDay, AppointmentID.

Data Anonymization Tool

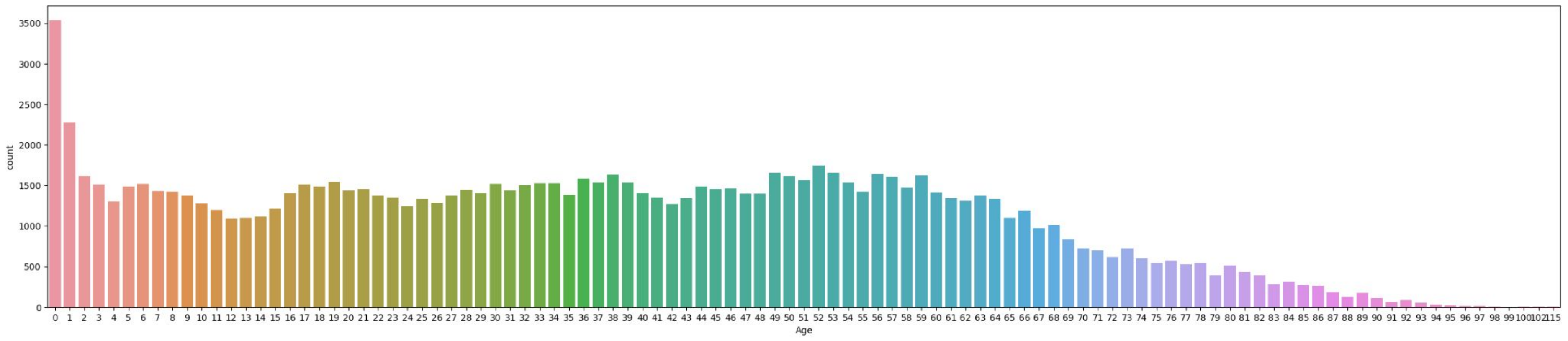
We used the python library - Diffprivlib for the task.

This is IBM’s differential privacy library, and is used for experimenting with, investigating and developing applications in, differential privacy.

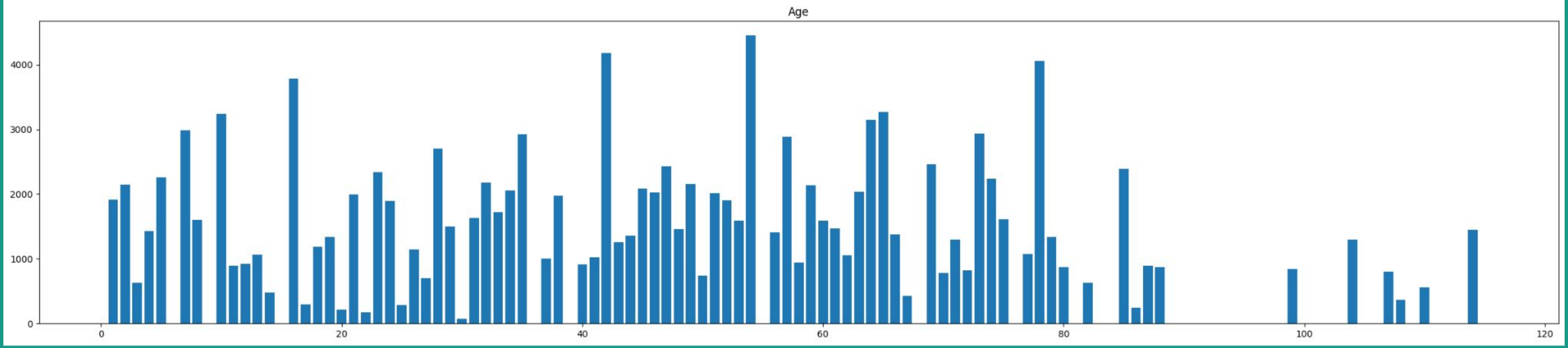
It leverages models and queries from well-known Python libraries like numpy and scikit-learn, commonly employed for data analysis and predictions, and integrates differential privacy into these models.

Attribute “Age”

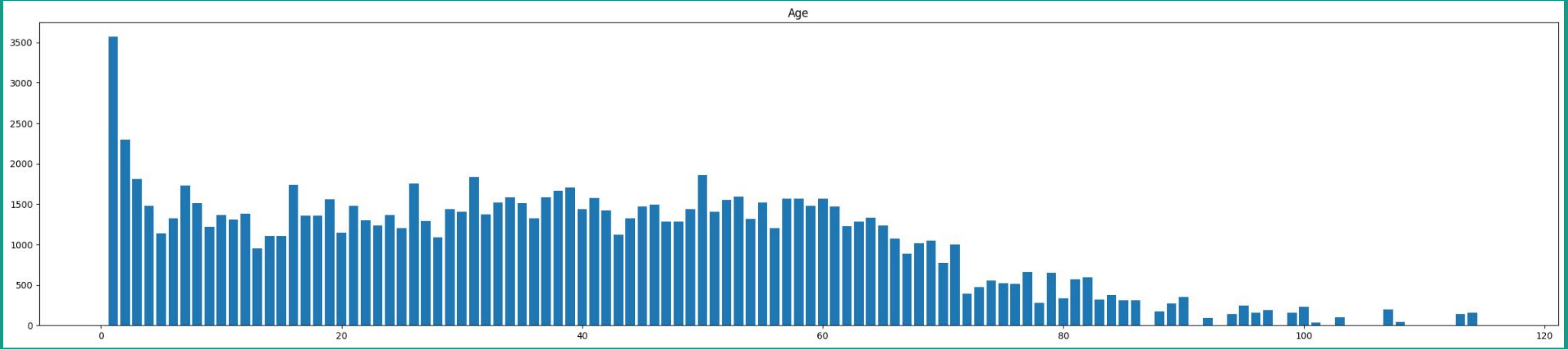
Original distribution of “Age” attribute.



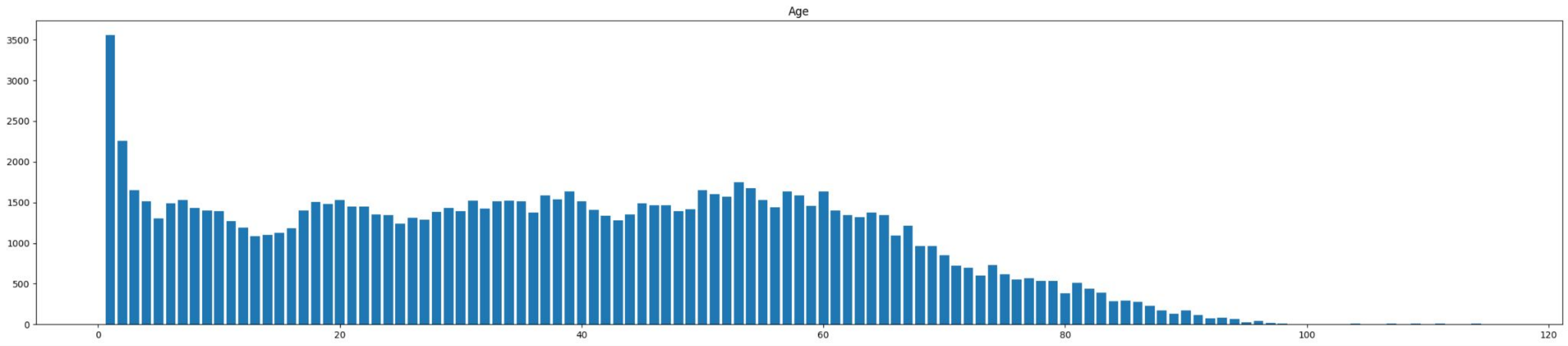
Distribution of "Age" attribute with epsilon = 0.0001



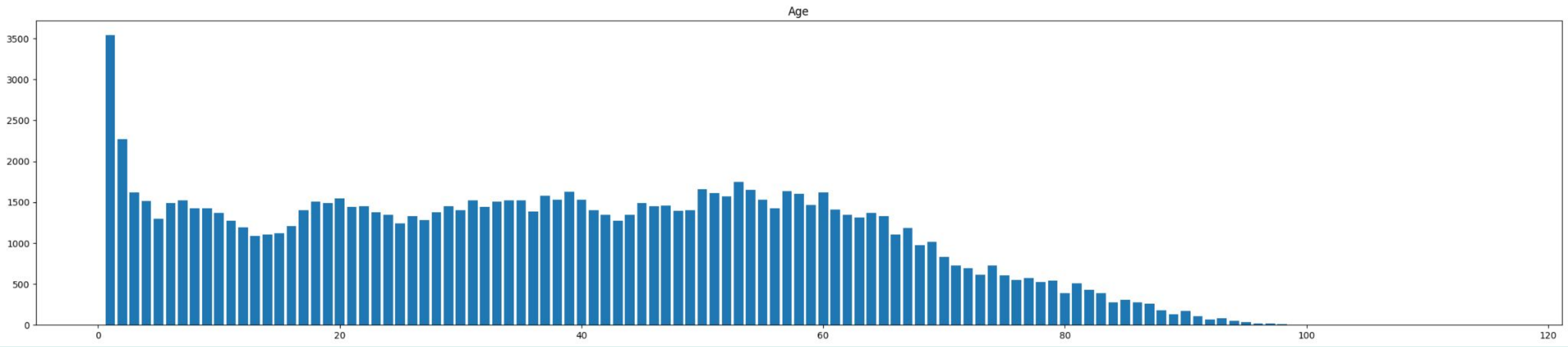
Distribution of "Age" attribute with epsilon = 0.001



Distribution of “Age” attribute with epsilon =0.1

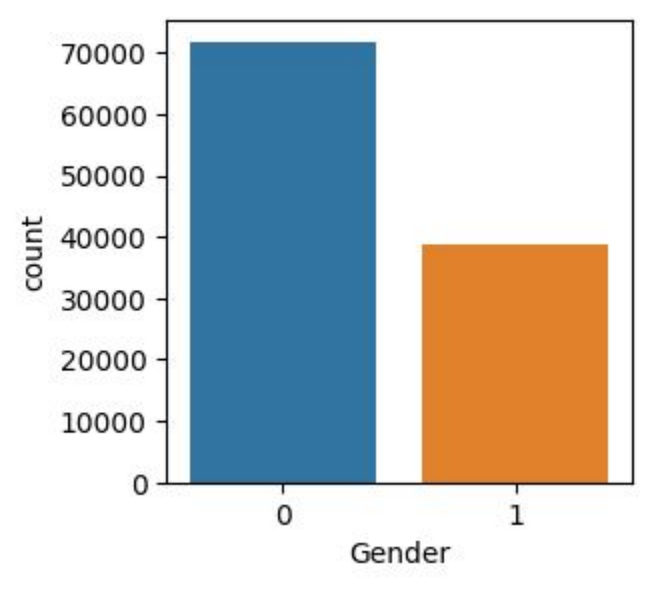


Distribution of “Age” attribute with epsilon =1

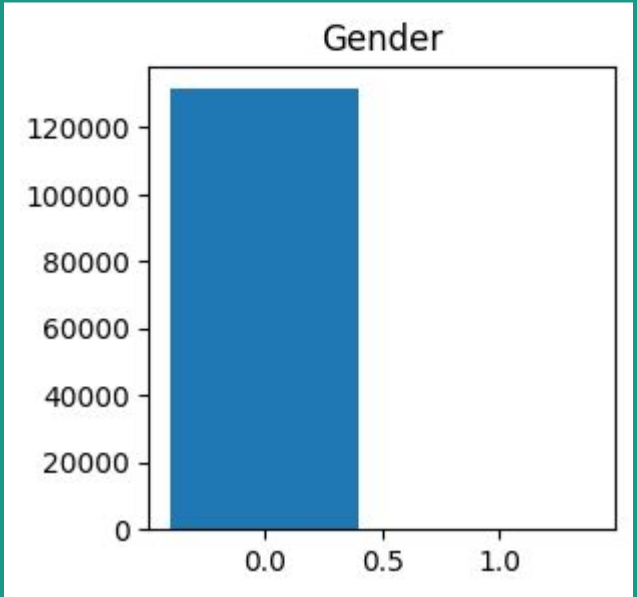


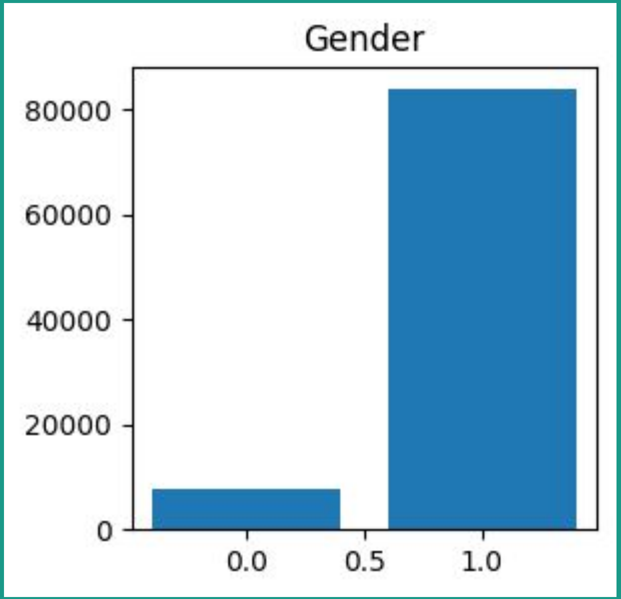
Attribute “Gender”

Original distribution of “Gender” attribute

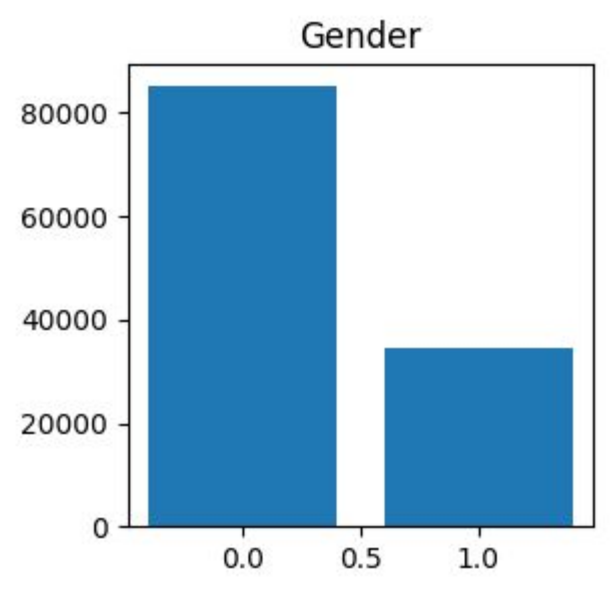


Distribution of “Gender” attribute with epsilon =0.00001

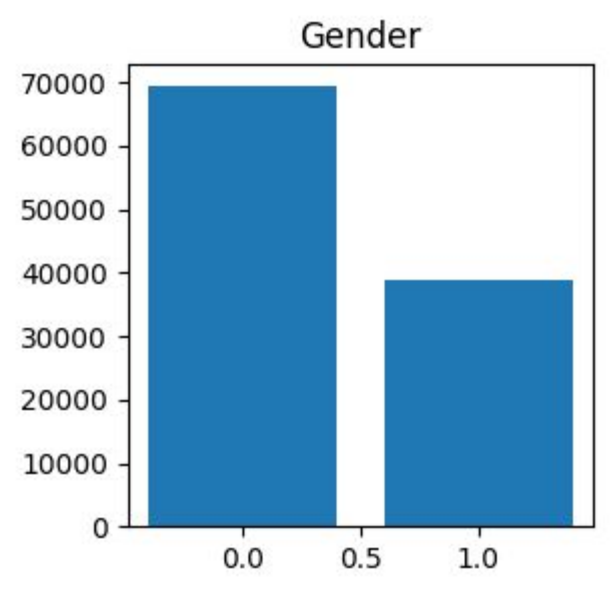




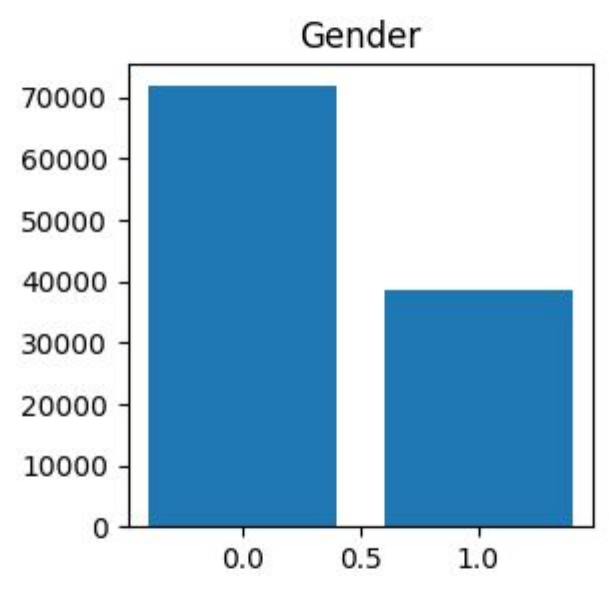
Distribution of “Gender” attribute with epsilon =0.0001



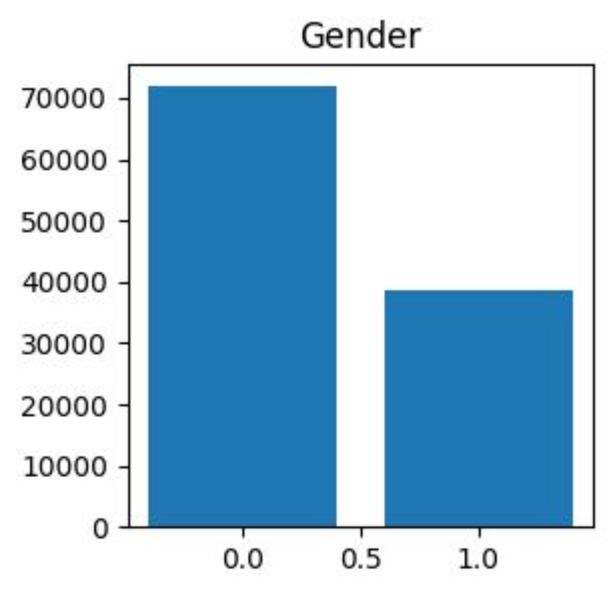
Distribution of “Gender” attribute with epsilon =0.001



Distribution of “Gender” attribute with epsilon =0.01

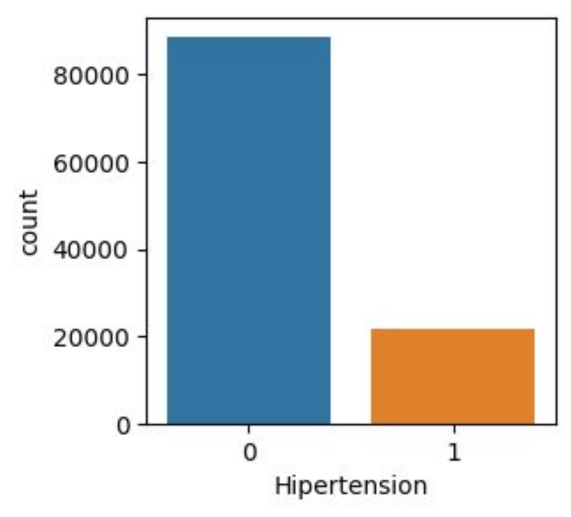


Distribution of “Gender” attribute with epsilon =0.1

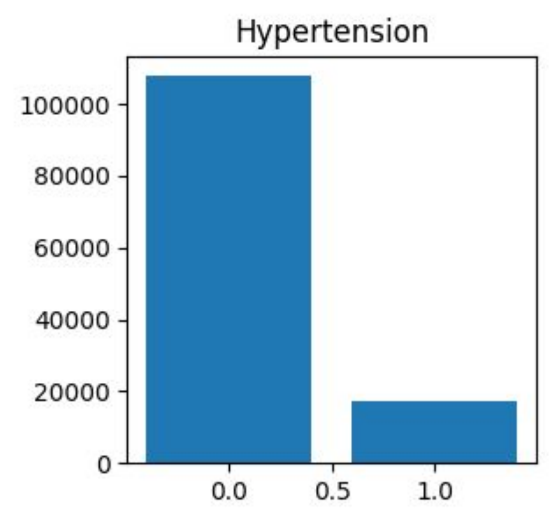


Attribute “Hypertension”

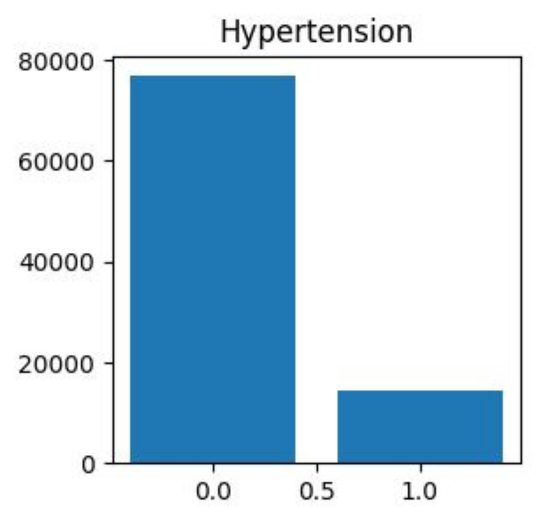
Original distribution of “Hypertension” attribute



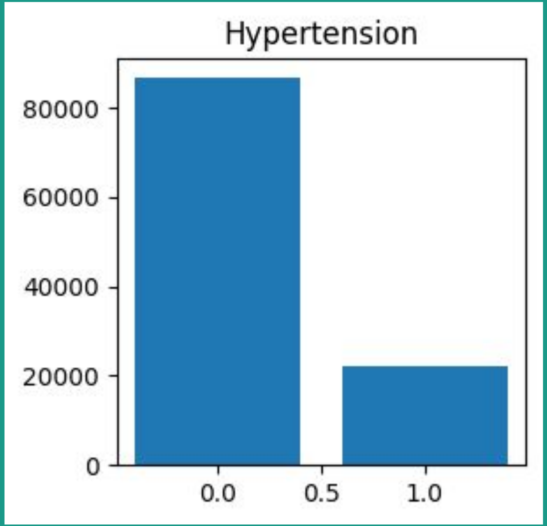
Distribution of “Hypertension” attribute with epsilon =0.00001



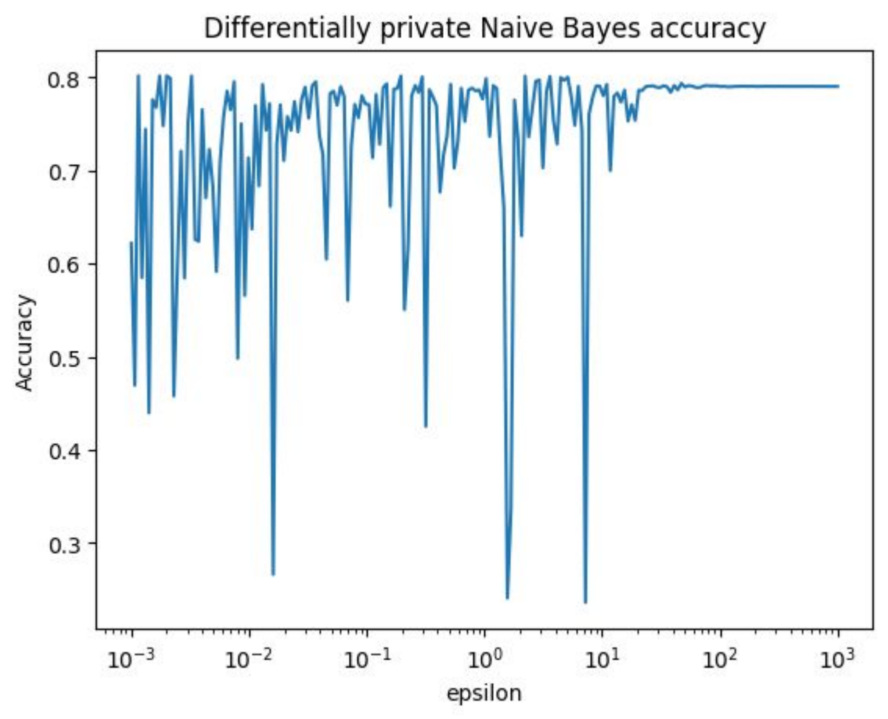
Distribution of “Hypertension” attribute with epsilon =0.0001



Distribution of “Hypertension” attribute with epsilon =0.001



Accuracy values versus Epsilon



The accuracy on the original dataset after preprocessing was 80.11%.

Demo