Gaussian Differential Privacy

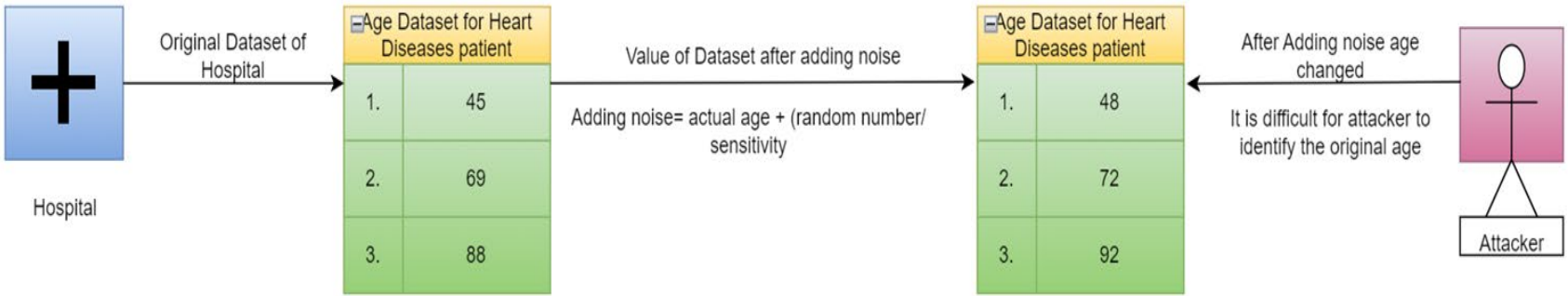
Applying Gaussian Differential Privacy to Heart Disease Data

Lilian Sun | Spring 2023 | College of Computing

Background Information

Sharing a data set by applying GDP

The healthcare statistician are interested in obtaining heart disease data for the community from a specialist hospital, but the specialist hospital is not willing to provide the original data set. Instead, the specialist hospital will use Gaussian differential privacy to add random noise to the original data set, which will protect the privacy of individual patients while still allowing the statistician to analyze the data.


Hospital
Original Dataset of Hospital
Age Dataset for Heart Diseases patient
1. 45
2. 69
3. 88
Value of Dataset after adding noise
Adding noise= actual age + (random number/ sensitivity
Age Dataset for Heart Diseases patient
1. 48
2. 72
3. 92
After Adding noise age changed
It is difficult for attacker to identify the original age
Attacker

Introduction to GDP

Gaussian Differential Privacy is a technique for protecting the privacy of individuals in a dataset.

The core idea of Gaussian Differential Privacy is to add random noise to the data before releasing it. The amount of noise added depends on the sensitivity of the data.

The amount of noise added is determined by the sensitivity of the query and desired level of privacy protection.

Noise added is drawn from gaussian distribution. Amount of noise added depends on standard deviation of GD.

Smaller the epsilon value, stronger the privacy of data.

GAUSSIAN MECHANISM

- Probability of the mechanism outputting a result in S for the original dataset D should not be much larger than the probability of it outputting a result in S for the neighboring dataset D’, where "not much larger" is defined by the parameter ε.

$$Pr[M(D) \in S] \leq Pr[M(D') \in S] \times e^{\varepsilon} + \delta$$

- The parameter δ represents the maximum probability by which this condition may be violated (privacy loss parameter).

Walkthrough - Research Topic

We work on Heart Disease data to study the trade-off between utility and privacy

By adjusting noise (privacy budge) to balance privacy protection to get a optimal privacy budge

1. **Proof the utility & privacy of Heart Disease data after applying GDP**

1. **Verify after applying the optimal privacy budget, if the noise data influence the performance of ML models (verify the utility of the noise data)**

Data Set: Heart Disease

Heart Disease Data Set

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212.000000	0	1	168	0	1.000000	2	2	3	0
53	1	0	140	203.000000	1	0	155	1	3.100000	0	0	3	0
70	1	0	145	174.000000	0	1	125	1	2.600000	0	0	3	0
61	1	0	148	203.000000	0	1	161	0	0.000000	2	1	3	0
62	0	0	138	294.000000	1	1	106	0	1.900000	1	3	2	0

- According to the Gaussian mechanism, for a function f(x) which returns a number, the following definition of F(x) satisfies (ε, δ) - differential privacy

$$F(x) = f(x) + N(\sigma)$$

NOISE CALCULATION

$$\sqrt{2\log(1.25/\delta)}\ \frac{\Delta_2 f}{\varepsilon} \longrightarrow \textbf{\textit{Sensitivity}}$$

- We are adding this Normal Distribution Scale to the Output value to get the Noisy Result.
- As we know for Counting queries whose neighboring dataset differ by 1 row, Sensitivity is 1.
- Note: Here Delta is the Failure Probability which should be greater than 0.

HOW MUCH AM I DEVIATED ?

- Relative Error :

Xi' : Perturbation value of xi

$$E = \frac{\left| f_i(D) + u_i - f_i(D) \right|}{\left| f_i(D) \right|}$$

$$E = |x_i' - x_i| / |x_i|$$

IT’S ABOUT UTILITY TOO

- UTILITY METRIC

$$U = 1 - E$$

RELATION IDENTIFICATION

- **CASE-1**

  When Epsilon is small, High Privacy, Less Utility / Accuracy

- **CASE-2**

  When Epsilon is large, Less Privacy, More Utility / Accuracy

So, When Epsilon value is increasing, we can conclude below facts:

High Utility => High Accuracy => Less Errors => Less Utility loss

EPSILON VS UTILITY LOSS

epsilon	UtilityLoss
0.1	0.41
0.3	0.16
0.5	0.14
0.9	0.13

Figure 5.2

EPSILON VS UTILITY LOSS

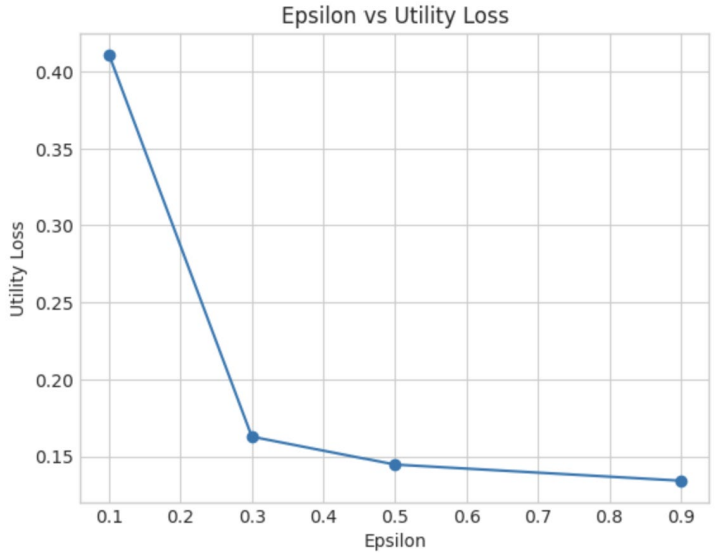


Figure 1.2

EPSILON VS ACCURACY

epsilon	Accuracy
0.1	58.92
0.3	83.73
0.5	85.54
0.9	86.58

Figure 1.3

EPSILON vs ACCURACY

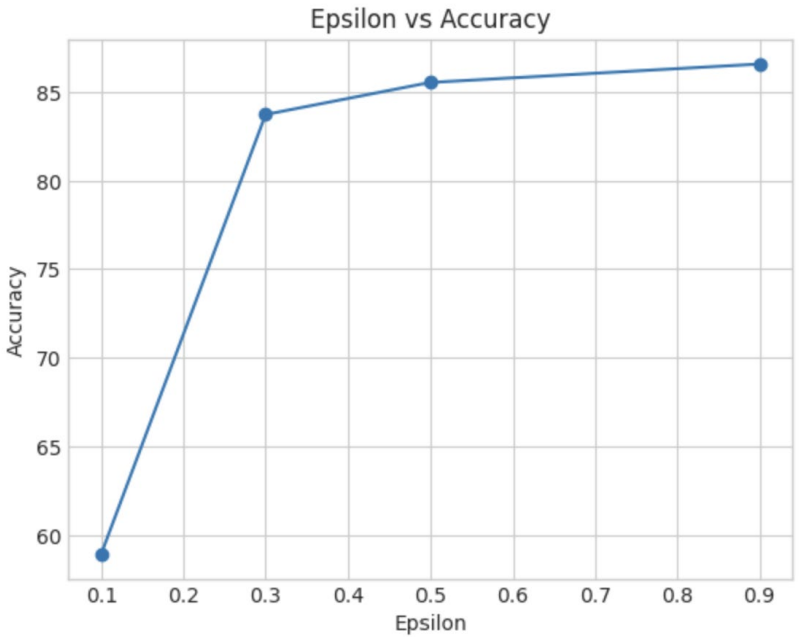


Figure 1.4

THE SIGNIFICANCE of THE OPTIMAL BUDGET

Sharing a data set by applying GDP

Based on prior research and code implementation, we will determine the optimal privacy budget to be added to the original heart disease dataset before sharing it with healthcare statisticians. The healthcare statisticians will then collect and analyze the noisy dataset to make informed decisions by using different machine learning models

Lab 2 MACHINE LEARNING with GDP

ML models are trained on Independent Variables

Independent Variables (Features)

Dependent Variables (Response)

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212.000000	0	1	168	0	1.000000	2	2	3	0
53	1	0	140	203.000000	1	0	155	1	3.100000	0	0	3	0
70	1	0	145	174.000000	0	1	125	1	2.600000	0	0	3	0
61	1	0	148	203.000000	0	1	161	0	0.000000	2	1	3	0
62	0	0	138	294.000000	1	1	106	0	1.900000	1	3	2	0

Lab 2 MACHINE LEARNING with GDP

There are 13 independent variables in the dataset are classified into two types:
1st. continuous features 2nd. categorical features

The change in continuous or categorical features can have varying degrees of influence on ML models' performance

Independent Variables (Features)

Dependent Variables (Response)

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212.000000	0	1	168	0	1.000000	2	2	3	0
53	1	0	140	203.000000	1	0	155	1	3.100000	0	0	3	0
70	1	0	145	174.000000	0	1	125	1	2.600000	0	0	3	0
61	1	0	148	203.000000	0	1	161	0	0.000000	2	1	3	0
62	0	0	138	294.000000	1	1	106	0	1.900000	1	3	2	0

Lab 2 MACHINE LEARNING with GDP

We research how the optimal budget influence the performance of ML models by adding noise to **continuous and categorical features respectively.**

Independent Variables (Features)

Dependent Variables (Response)

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212.000000	0	1	168	0	1.000000	2	2	3	0
53	1	0	140	203.000000	1	0	155	1	3.100000	0	0	3	0
70	1	0	145	174.000000	0	1	125	1	2.600000	0	0	3	0
61	1	0	148	203.000000	0	1	161	0	0.000000	2	1	3	0
62	0	0	138	294.000000	1	1	106	0	1.900000	1	3	2	0

Lab 2 MACHINE LEARNING PIPELINES

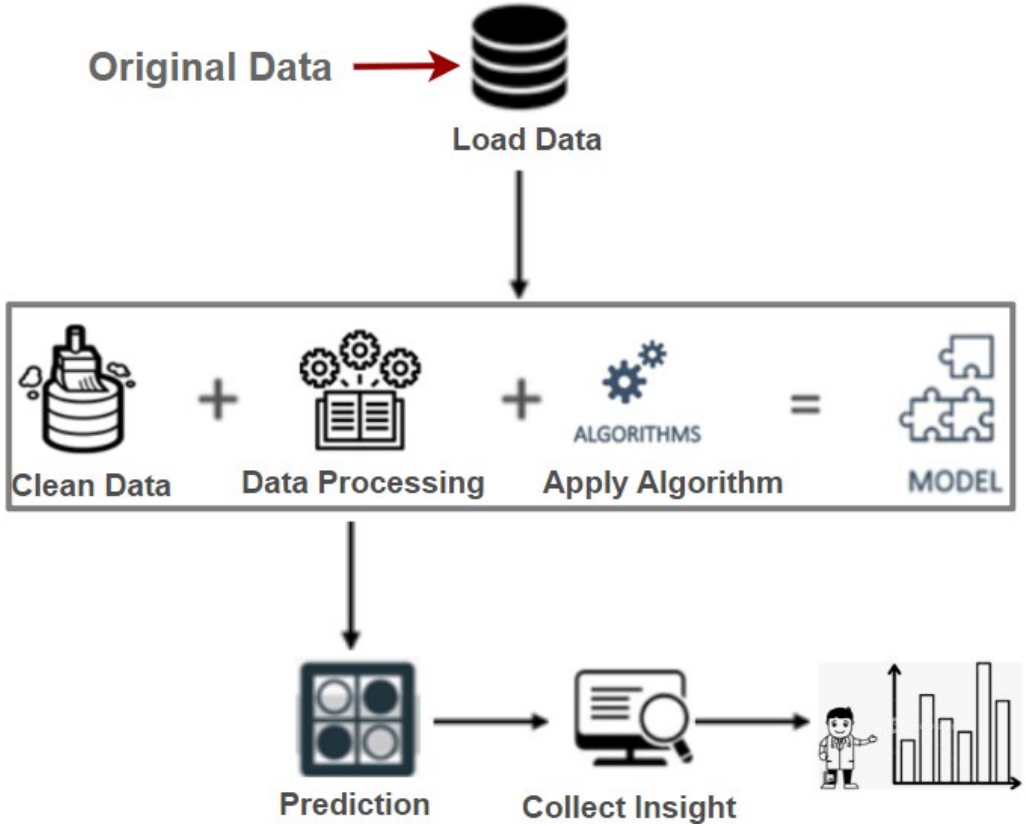


Figure 3.1 Model training process

Model Accuracy is a measure of how well a machine learning model is able to make accurate predictions on new, unseen data.

Model	Accuracy
K-Nearest Neighbour	95.609756
Random Forest	88.780488
Gradient Boosting	86.829268
Logistic Regression	83.902439
Support Vector Machine	83.902439
Decision Tree	83.902439
Gaussian Naive Bayes	82.439024

Lab 2 MACHINE LEARNING PIPELINES

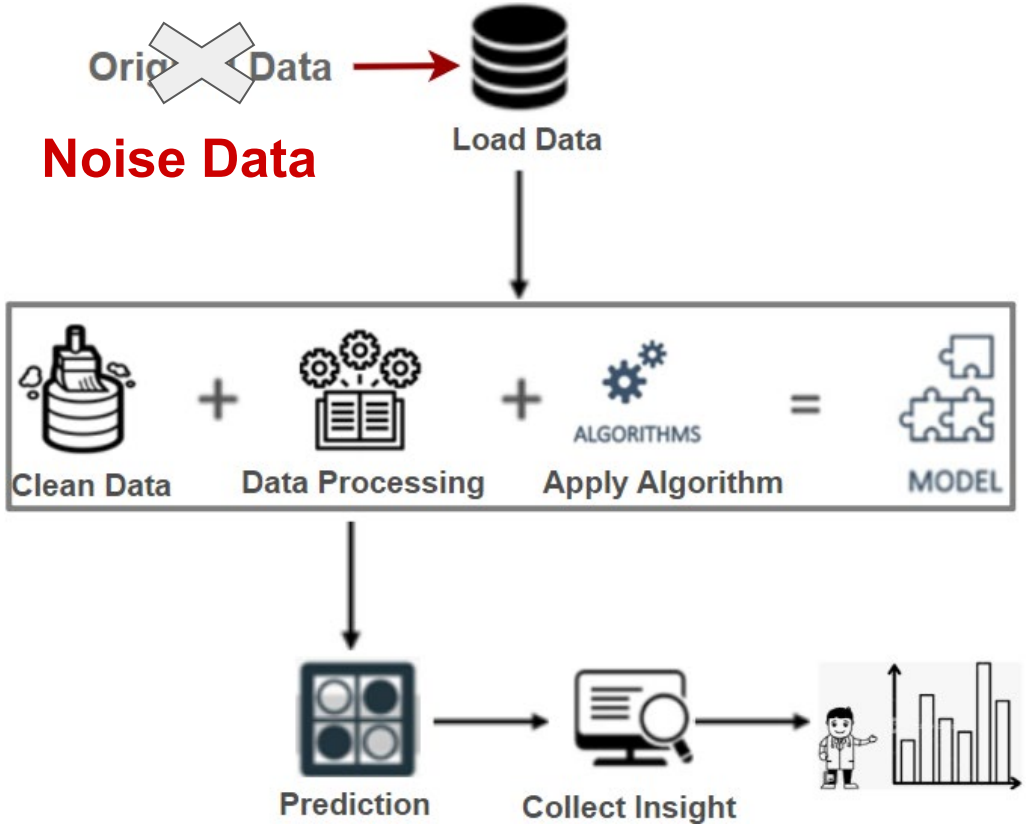


Figure 3.2 Model training process

The only change in the ML pipelines is data quality

Model	Accuracy
K-Nearest Neighbour	
Random Forest	
Gradient Boosting	
Logistic Regression	?
Support Vector Machine	
Decision Tree	
Gaussian Naive Bayes	

APPLYING GDP to CONTINUOUS FEATURES

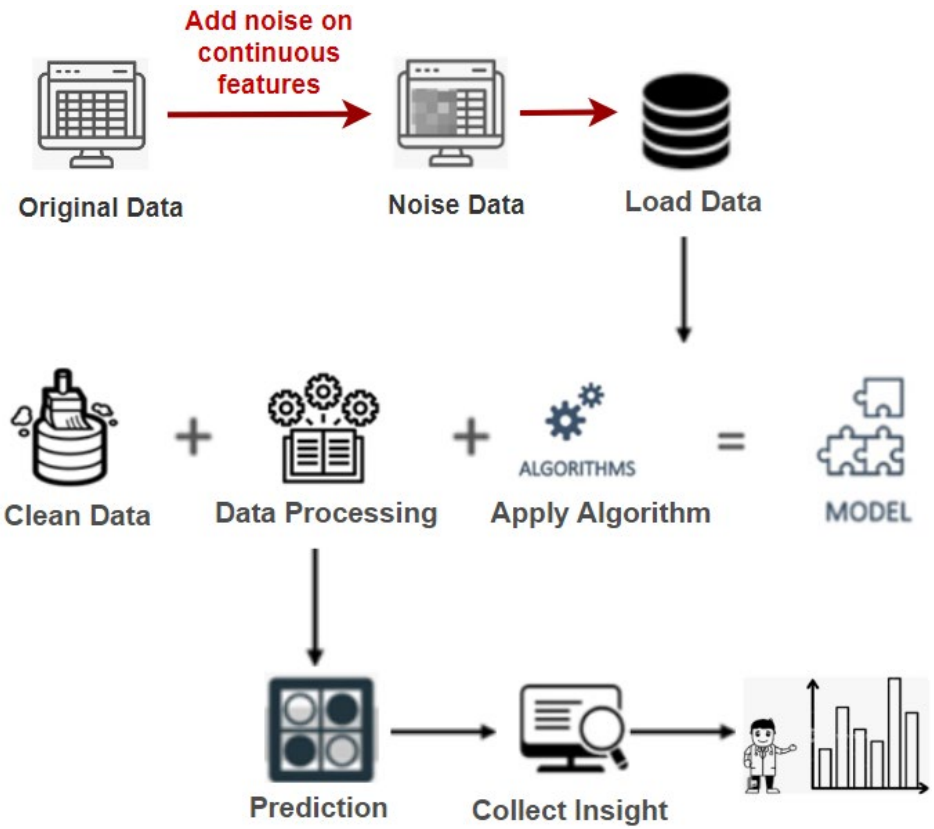


Figure 3.3 Model training process

The only change in the ML pipelines is data quality

Model	Accuracy
K-Nearest Neighbour	
Random Forest	
Gradient Boosting	
Logistic Regression	?
Support Vector Machine	
Decision Tree	
Gaussian Naive Bayes	

APPLYING GDP to CONTINUOUS FEATURES

GDP ~~X~~ No noise

age	trestbps	thalach	oldpeak
52	125	168	1.000000
53	140	155	3.100000
70	145	125	2.600000
61	148	161	0.000000
62	138	106	1.900000

GDP Data

age	trestbps	thalach	oldpeak
64.799258	137.799258	180.799258	13.799258
65.799258	152.799258	167.799258	15.899258
82.799258	157.799258	137.799258	15.399258
73.799258	160.799258	173.799258	12.799258
74.799258	150.799258	118.799258	14.699258

sensitivity=1, epsilon=0.5, delta = 1e-5

APPLYING GDP to CONTINUOUS FEATURES

df_v1_mean	df_v2_mean	mean_diff
54.434146	53.872515	-0.561631
131.611707	131.078570	-0.533137
149.114146	148.561968	-0.552178
1.071512	0.532964	-0.538548

df_v1_variances	df_v2_variances	variance_diff
54.434146	53.872515	81.808681
131.611707	131.078570	306.792678
149.114146	148.561968	529.101013
1.071512	0.532964	1.379628

We are not sure if the change (variance difference) could result in overfitting or underfitting the noise data. **Let's directly compare ML performance!**

ML PERFORMANCE COMPARISION

GDP ✗ No noise

Model	Accuracy
K-Nearest Neighbour	95.609756
Random Forest	88.780488
Gradient Boosting	86.829268
Logistic Regression	83.902439
Support Vector Machine	83.902439
Decision Tree	83.902439
Gaussian Naive Bayes	82.439024

GDP

Model	Accuracy
K-Nearest Neighbour	98.048780
Random Forest	91.707317
Gradient Boosting	91.707317
Gaussian Naive Bayes	88.780488
Logistic Regression	87.804878
Decision Tree	86.829268
Support Vector Machine	85.853659

Models performance does not significantly change after applying GDP with the optimal ε on the continuous features

APPLYING GDP to CATEGORICAL FEATURES

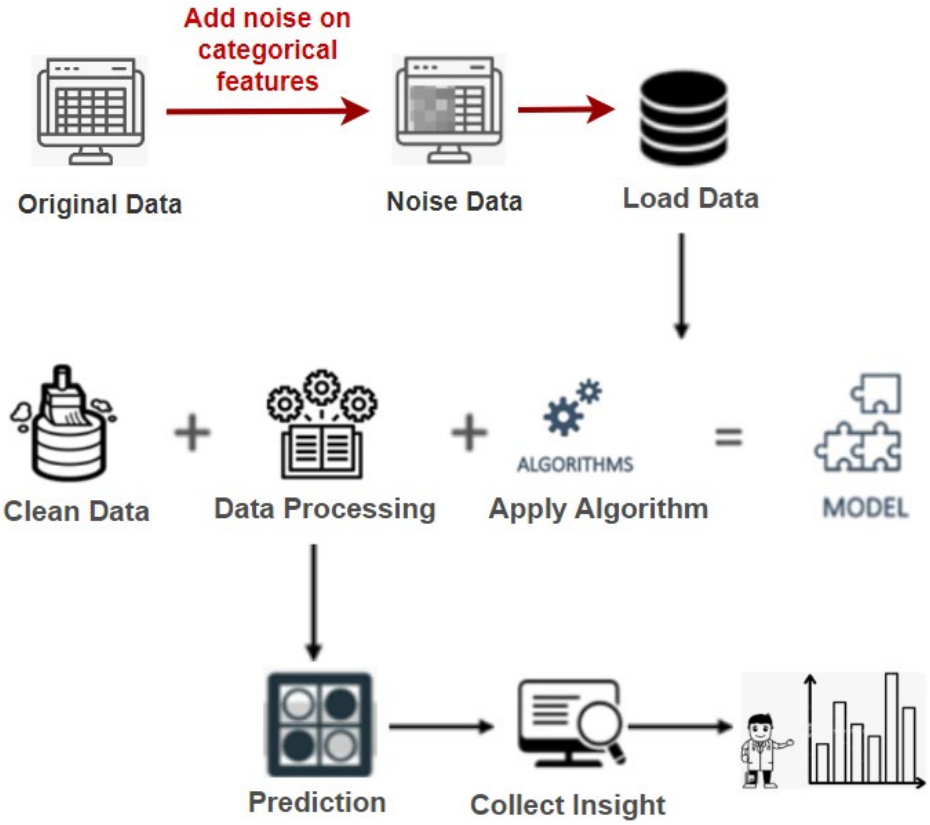


Figure 3.4 Model training process

The only change in the ML pipelines is data quality

Model	Accuracy
K-Nearest Neighbour	
Random Forest	
Gradient Boosting	
Logistic Regression	?
Support Vector Machine	
Decision Tree	
Gaussian Naive Bayes	

APPLYING GDP to CATEGORICAL FEATURES

GDP ~~X~~ No noise

sex	chol	fbs	restecg	exang	ca
1	212	0	1	0	2
1	203	1	0	1	0
1	174	0	1	1	0
1	203	0	1	0	1
0	294	1	1	0	3

GDP Data

sex	chol	fbs	restecg	exang	ca
9.565626	220.565626	8.565626	9.565626	8.565626	10.565626
9.565626	211.565626	9.565626	8.565626	9.565626	8.565626
9.565626	182.565626	8.565626	9.565626	9.565626	8.565626
9.565626	211.565626	8.565626	9.565626	8.565626	9.565626
8.565626	302.565626	9.565626	9.565626	8.565626	11.565626

sensitivity=1, epsilon=0.5, delta = 1e-5

ML PERFORMANCE COMPARISION

GDP ✗ No noise

Model	Accuracy
K-Nearest Neighbour	95.609756
Random Forest	88.780488
Gradient Boosting	86.829268
Logistic Regression	83.902439
Support Vector Machine	83.902439
Decision Tree	83.902439
Gaussian Naive Bayes	82.439024

GDP

Model	Accuracy
K-Nearest Neighbour	98.048780
Random Forest	91.707317
Gradient Boosting	91.707317
Gaussian Naive Bayes	88.780488
Logistic Regression	87.804878
Decision Tree	86.829268
Support Vector Machine	85.853659

Models performance does not significantly change after applying GDP with the optimal ε on the categorical features

Conclusion

- ❑ Hospital shared noise data with optimal privacy budget
- ❑ Achieved balance between privacy and utility

- ❑ Optimal privacy budget did not compromise statistical information

- ❑ Privacy budget's impact on continuous and categorical features deemed negligible in these 7 models

Challenge

The accuracy of the machine learning models on the original heart disease dataset was below 50%, it was difficult to meaningfully compare the performance of the models before and after applying Gaussian differential privacy. After performing proper feature engineering, the accuracy of the models has been increased to a level where meaningful comparisons can be made.

Reference

[1] Dong, J., Roth, A., & Su, W. J. (2022). Gaussian differential privacy. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(1), 3–37. https://doi.org/10.1111/rssb.12454

[2] Dwork, C., & Smith, A. (2010). Differential privacy for statistics: What we know and what we want to learn. Journal of Privacy and Confidentiality, 1(2). https://doi.org/10.29012/jpc.v1i2.570

[3] Borja Balle 1 Yu-Xiang Wang 2 3. Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising