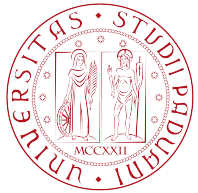


Differential Privacy in Medical Analysis

PhD Course Stat4Engineers

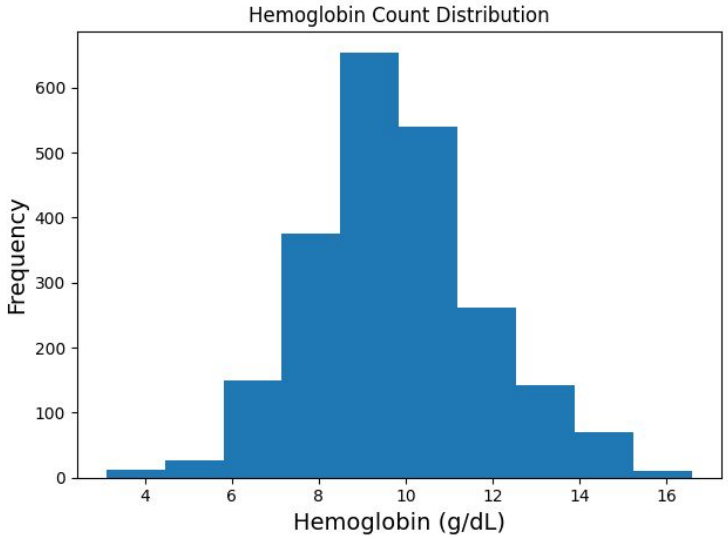
Monguelfo, June 25-28 2024

Francesco L. De Faveri
Laura Menotti
Department of Information Engineering, UniPD

Scenario and Problem Definition

Privacy Risk

When analyzing personal information, there is a Privacy risk of disclosing the sensitive information of the outliers patients.



ε-Differential Privacy & Classification

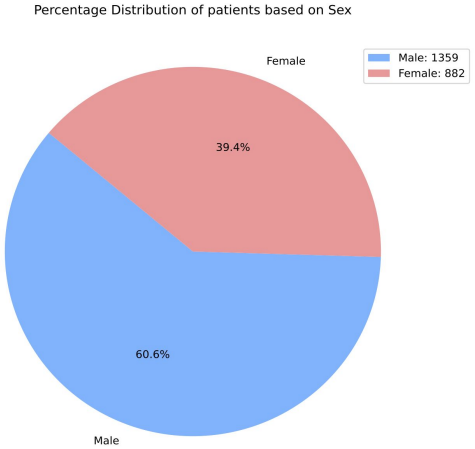
RQ: "Is it possible to apply Differential Privacy during a Classification Task to protect data confidentiality? If ***yes****, at what cost?"*

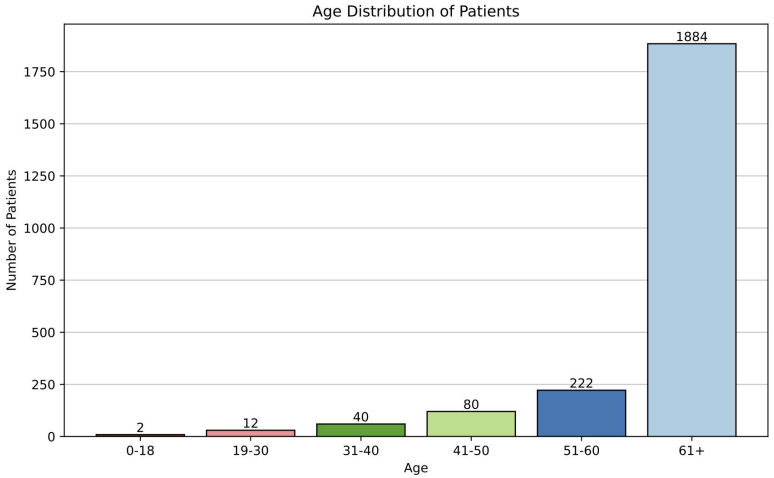
Dataset description

We used a dataset of ***2,241 Myelodysplastic Syndromes patients*** collected for developing a clinical-molecular prognostic model.

The dataset is available in cBioportal .

The dataset contains extended information regarding blood analysis, bone marrow analysis, genomic test results, and genome mutations represented as numeric values and string expressions.





Bernard et al. Molecular International Prognostic Scoring System for Myelodysplastic Syndromes. NEJM Evid 2022;1(7). https://evidence.nejm.org/doi/full/10.1056/EVIDoa2200008

One Sample Test

We applied Laplacian noise parametrized by the privacy budget **ε** to the variable "Overall Survival (Months)" and check whether there is statistical difference between the obfuscated values and the real population mean.

"Overall Survival (Months)": Life expectancy after diagnosis expressed in months.

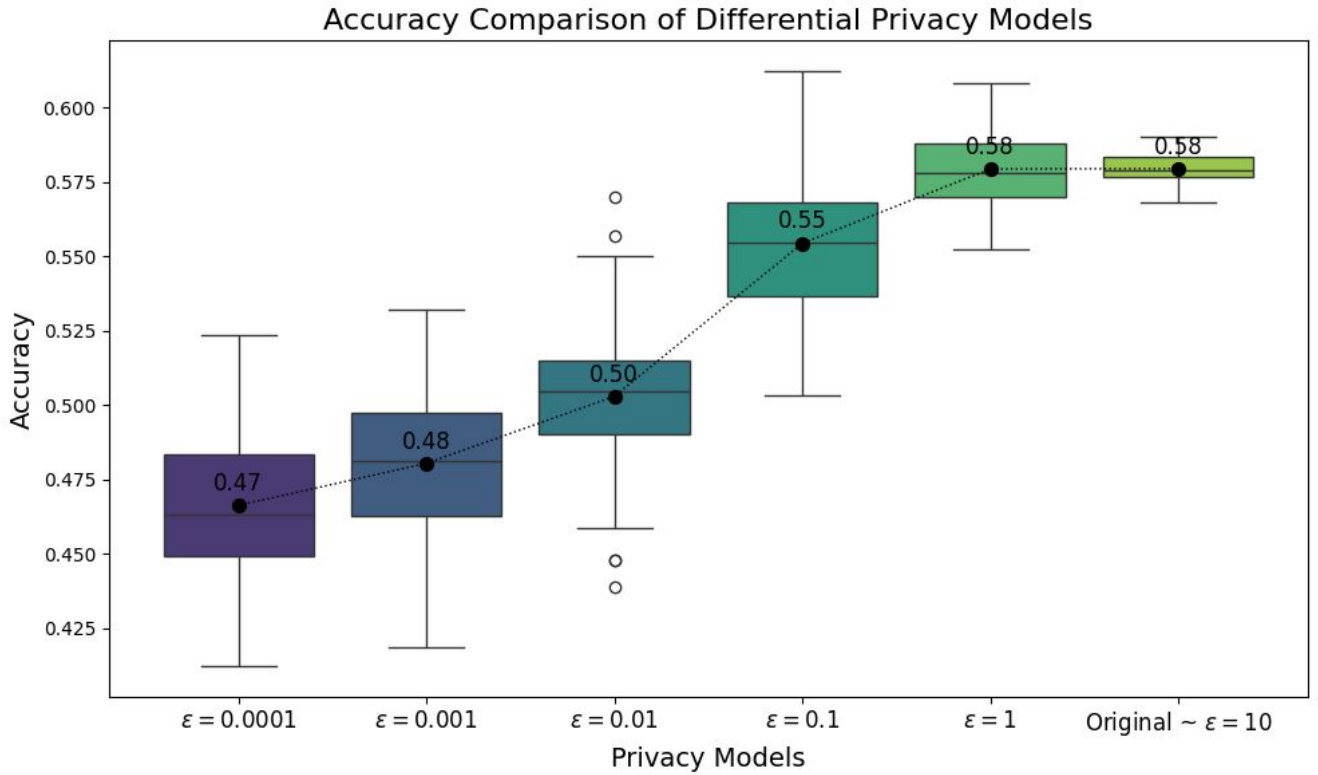
We set as null hypothesis. H_0: all means are equal; H_1: not all means are equal.

We consider $\alpha = 0.05$.

Approximate
Original

ε	0.0001	0.001	0.01	0.1	1	10
p-value	0.12	0.45	0.59	0.90	0.96	0.99

Box plot of the Classification Accuracy



Paired sample Test

We conducted a paired t-Test to ***check whether the classification accuracy of each privacy model is statistically different from the classification accuracy of the original model***, i.e. without privacy perturbation.

We consider α = 0.05.

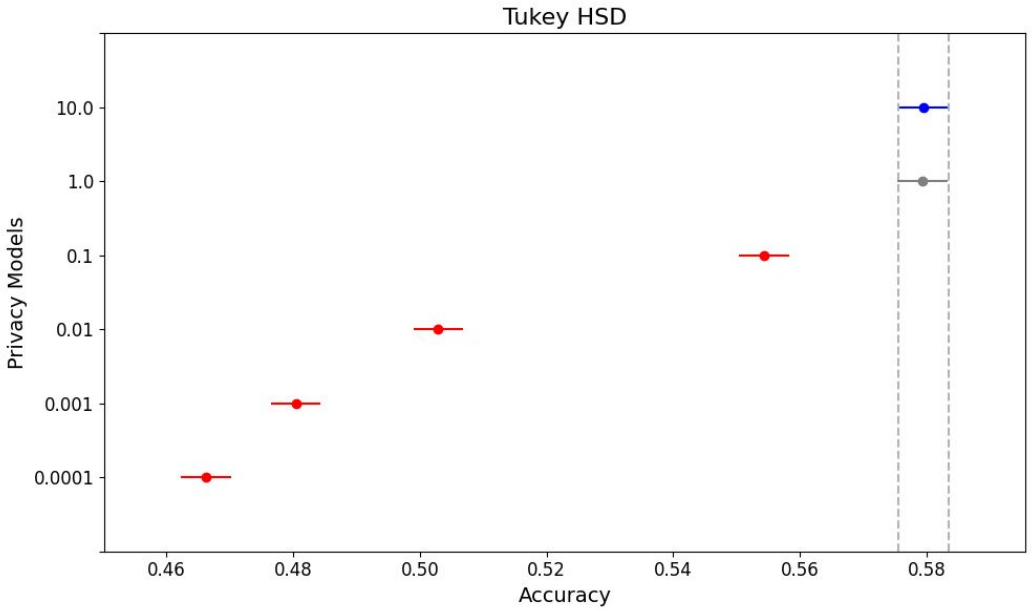
* Results in bold show statistical difference wrt to the original model.

ε	0.0001	0.001	0.01	0.1	1	Original
Accuracy	0.47	0.48	0.50	0.55	0.58	0.58
p-value	**2.41e-70**	**2.27e-63**	**5.09e-55**	**2.94e-18**	0.93	- - -

ANOVA and Tukey HSD tests

ANOVA 1-way table.

p-value	***α***
3.07e-44	0.05



DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Differential Privacy in Medical Analysis

Thank for the attention
Question Time!

Francesco L. De Faveri
Laura Menotti
Department of Information Engineering, UniPD