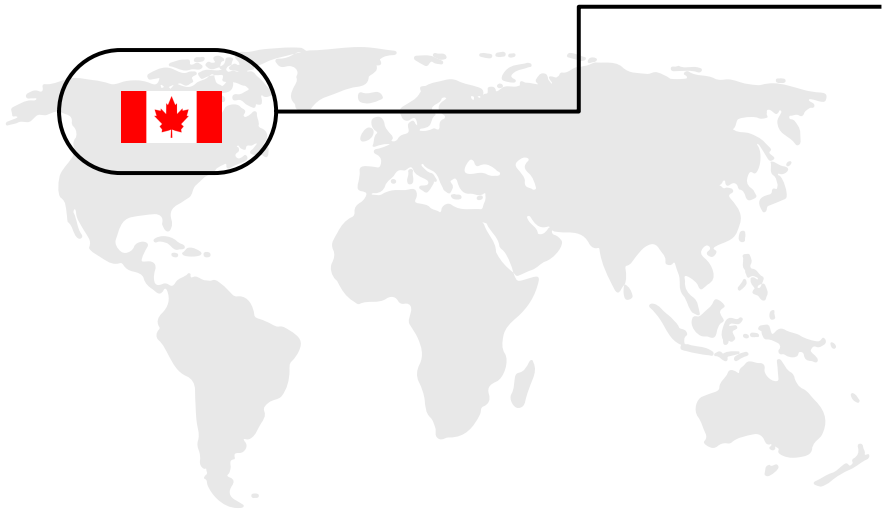
Privacy-Preserving Record Linkage Methods for Homelessness Data

Felipe Castaño Gonzalez
University of Calgary
ENEL 645
Fall 2024

UNIVERSITY OF CALGARY

Table of contents

1. Introduction



Canada

235,000 — People Experienced Homelessness [1]

18,467 — Shelter space [2]

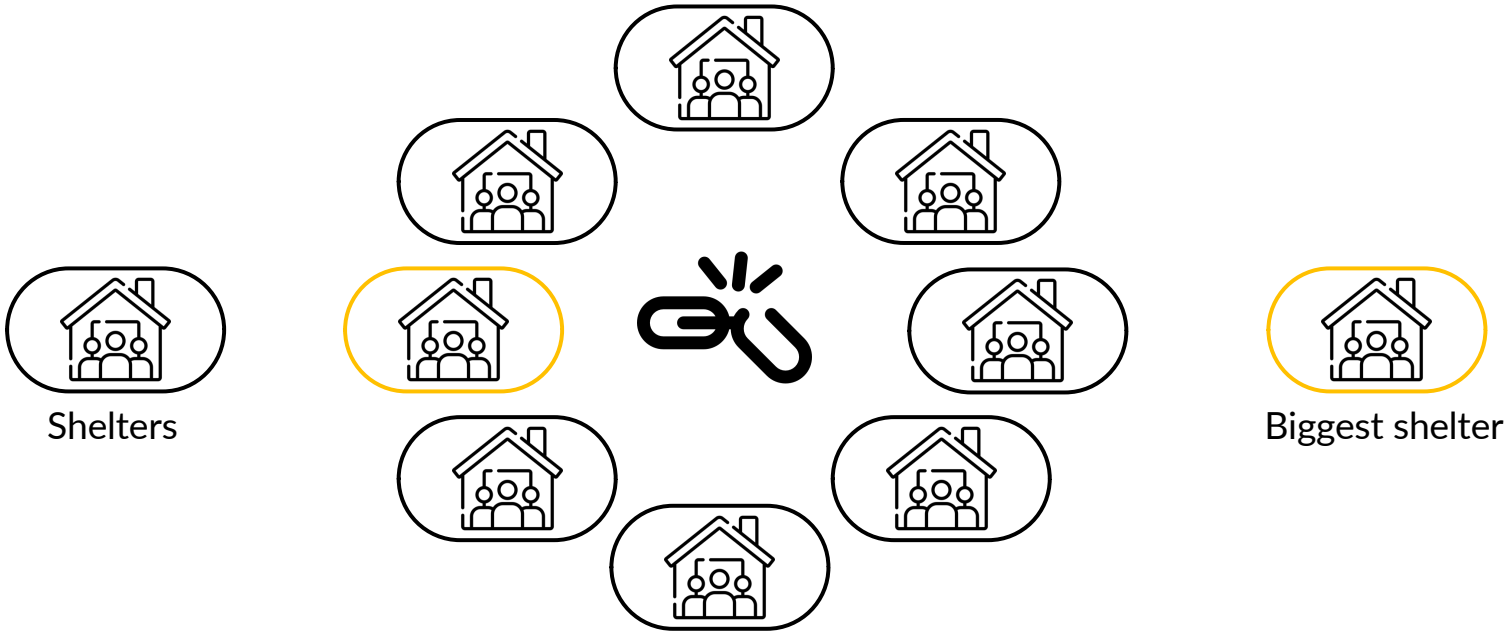
88,6 % — Occupancy rate [3]

Calgary

2,782 — Homeless individuals PiT [4]

6,839 — Unique individuals accessed Calgary Drop-In Centre [4]

1. Introduction



2.000 shelters in Calgary [5]

2. Related Work

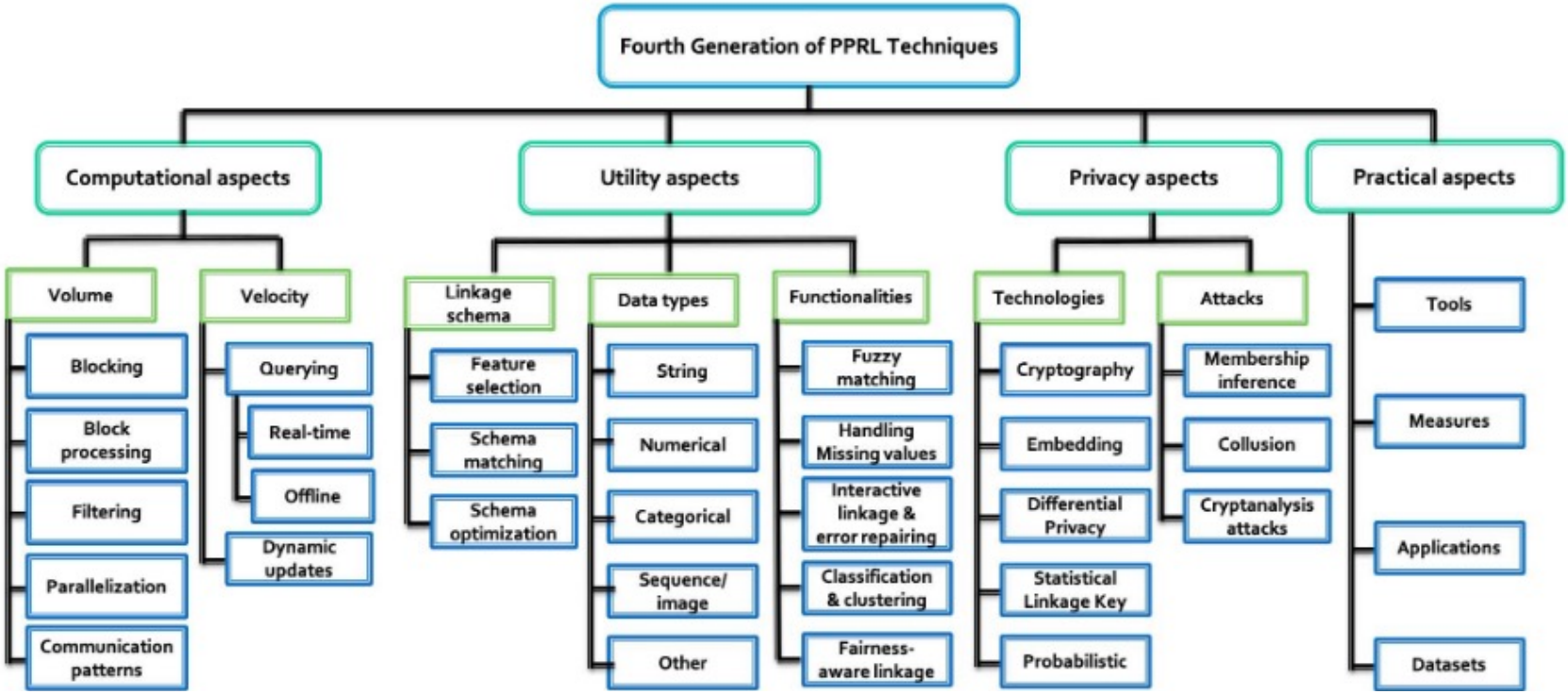
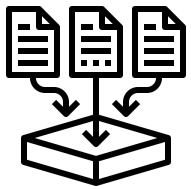
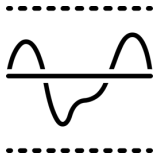


Fig. 1. A taxonomy of the data-driven (fourth) generation of privacy-preserving record linkage techniques. [6]

3. Materials and Methods

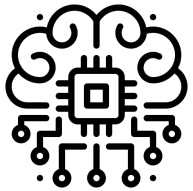


Dataset



Threshold

Privacy Preserving



Machine Learning

3.1 Dataset

Id	IdLinked	FirstName	LastName	DobDay	DobMonth	DobYear
1070	1070	michaela	neumann	11	11	1915
1016	1016	courtney	painter	14	12	1916
4405	4405	charles	green	30	9	1948
1288	1288	vanessa	parr	19	11	1995
3585	3585	mikayla	malloney	8	2	1986

Tab. 1. Raw synthetic dataset based on the Calgary Drop in Centre structure.

Records = 15.991
Unique = 4.750
Duplicated 11.241

ConcatenatedDetails	IsDuplicated
aabar3031973	1
aaliyahottens14121918	0
aaliyahottens14121918	1
aarobbstovm1931973	1
aaronbarsoum3031973	1
...	...
zoewastell15121998	0
zoewastell15121998	1
zoewebb17121910	1
zoewebb17121910	0
zoeyxlepaterson1291993	1

Tab. 2. Synthetic dataset after organization.

3.2 Privacy preserving

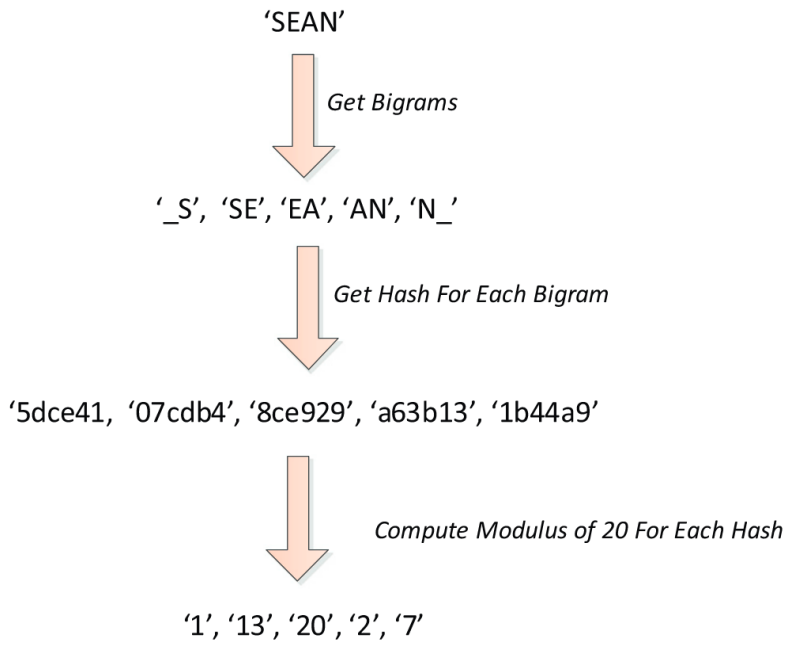
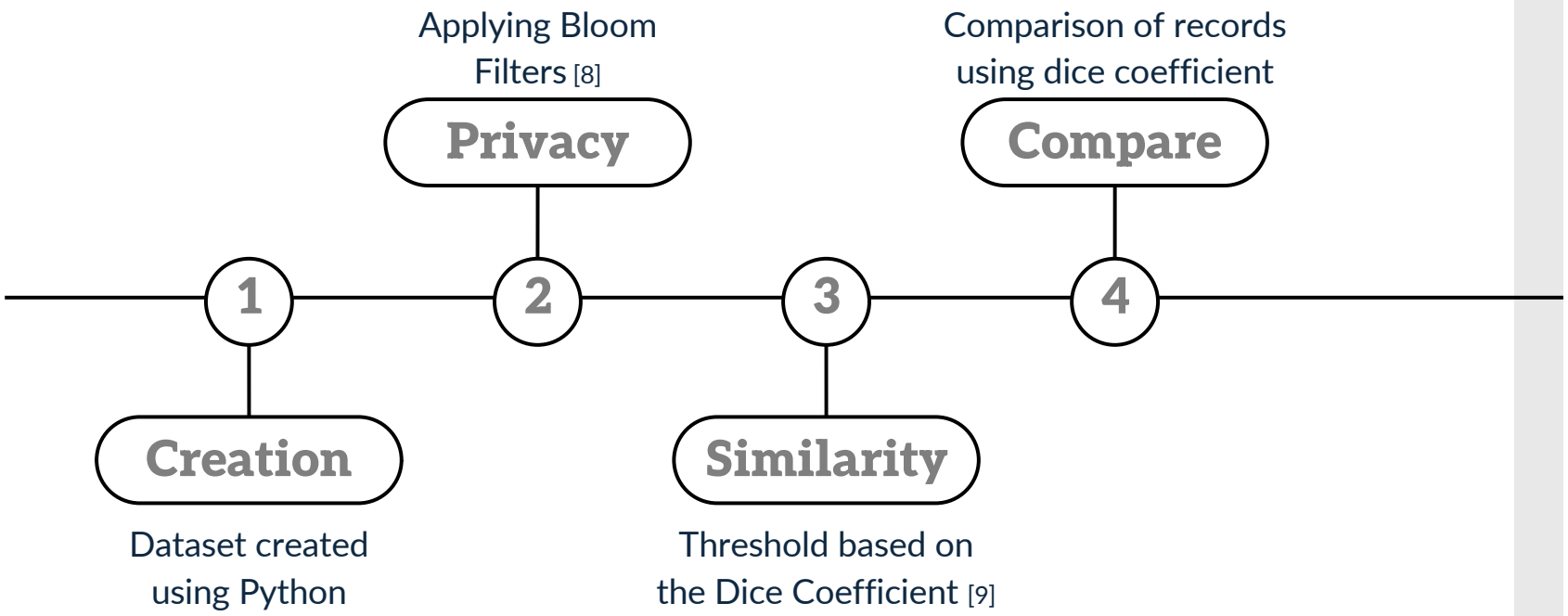
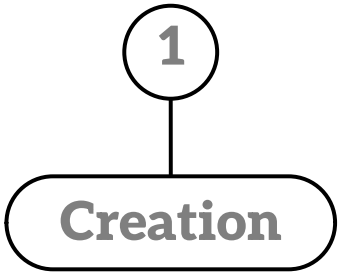


Fig. 2. Example of encoding the name 'Sean' into a Bloom Filter of length 20. [7]

3.3 Threshold





	FirstName	LastName	DobDay	DobMonth	DobYear	FullDetailsOrganized
0	rachael	dent	22	7	1928	rachaeldent2271928
1	rac	dent	22	7	1928	racdent2271928
2	rachaelintux	dent	22	7	1928	rachaelintuxdent2271928
3	isabella	everett	16	8	1911	isabellaeverett1681911
4	isabella	ewfsfut	16	8	1901	isabellaewfsfut1681901

Tab. 3. Complete dataset concatenating the information

ConcatenatedDetails	IsDuplicated
rachaeldent2271928	0
racdent2271928	1
rachaelintuxdent2271928	1
isabellaeverett1681911	0
isabellaewfsfut1681901	1
...	...
victoriavfitzpatrickjf1141981	1
victoriafitzpatrick1141981	1
victoriafitzpatrick1341981	1
victoriafitzpatrick2041981	1
victoriafitzpatrick1211955	1

Tab. 4. Dataset concatenated with a single or duplicate record

ConcatenatedDetails	IsDuplicated	BloomFilter32Bit
rachaeldent2271928	0	427125415
racdent2271928	1	2177565218
rachaelintuxdent2271928	1	2067881650
isabellaeverett1681911	0	2637064116
isabellaewfsfut1681901	1	2235145841
...	...	...
victoriavfitzpatrickjf1141981	1	1078740363
victoriafitzpatrick1141981	1	588773085
victoriafitzpatrick1341981	1	3120261611
victoriafitzpatrick2041981	1	1718509726
victoriafitzpatrick1211955	1	325097743

Tab. 5. Dataset concatenated with single or duplicate record and bloom filter the 32 bits.

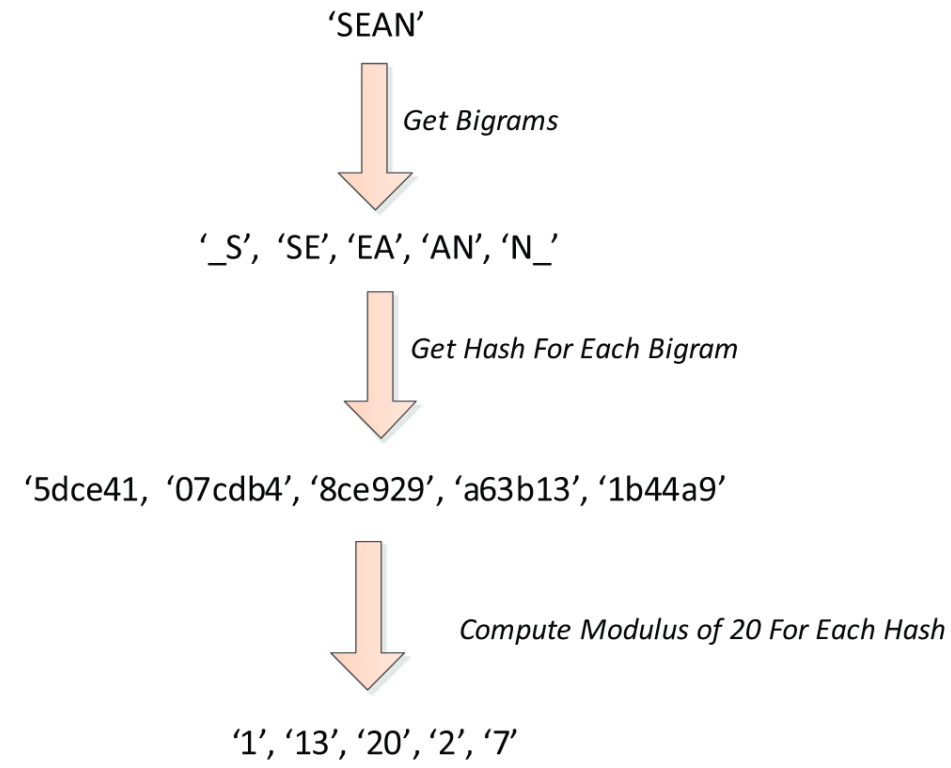


Fig. 2. Example of encoding the name 'Sean' into a Bloom Filter of length 20. [6]

3

Similarity

$$Dice\ Coefficient_{A,B} = \frac{2h}{a+b}$$

where h is the number of positions set to 1 in both bloom filters,
a is the number of bit positions set to 1 in bloom filter A,
and b is the number of bit positions set to 1 in bloom filter B.

An example....

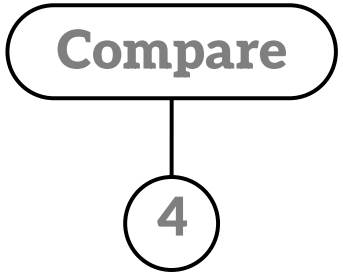
Bloom Filter 1: 5 positions set to 1

1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1
1	1	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	1

Bloom Filter 2: 6 positions set to 1

$$= \frac{2 \times 4}{5 + 6} = 0.727...$$

Fig. 3. Example of calculating string similarity by comparing two bloom filters. [6]



PrincipalDetails	IsDuplicated	BloomFilter32Bit
zoepaterson1291993	0	3921124910
jessicabeams2131904	1	1651423879
adamcra1051927	1	761781273
bertiewhite8101961	1	1526532195
jasminezrkwebbanupxbho1061944	1	4013317649
...	...	...
jackcarbone1411940	1	503599820
jackraward1991980	1	1574844469
nedstephenson1811975	0	735912613
brinleybauman2641936	1	491570036
michaeladunstone18101912	1	3176224164

Tab. 6. Dataset concatenated with single or duplicate record and bloom filter the 32 bits.

BloomFilter1	BloomFilter2
3921124910	1651423879
3921124910	761781273
3921124910	1526532195
3921124910	4013317649
3921124910	3144879481
...	...
1574844469	491570036
1574844469	3176224164
735912613	491570036
735912613	3176224164
491570036	3176224164

Tab. 7. Comparison of records with bloom filter

3.4 Machine Learning

XG Boost

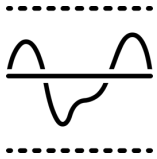
BloomFilter32Bit	BloomFilter32Bit_Binary	BF_CountOf1s	BF_Density	Hamming_Similarity
1643689933	01100001111110001011011111001101	19	0.59375	1.00000
3567187259	11010100100111101111010100111011	20	0.62500	0.46875
3567187259	11010100100111101111010100111011	20	0.62500	0.46875
48879422	00000010111010011101011100111110	17	0.53125	0.56250
2877932438	10101011100010011100001110010110	16	0.50000	0.46875
...	...	...	...	...
539910774	00100000001011100110001001110110	13	0.40625	0.43750
539910774	00100000001011100110001001110110	13	0.40625	0.43750
3715518538	11011101011101100101000001001010	16	0.50000	0.40625
3715518538	11011101011101100101000001001010	16	0.50000	0.40625
3761702588	11100000001101110000011010111100	15	0.46875	0.50000

Tab. 8. Dataset defined as X for the execution of XG Boost.

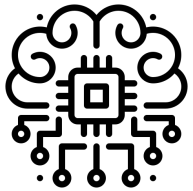
IsDuplicated
1
0
1
1
1
...
0
1
0
1
1

Tab. 9. Dataset defined as Y for the execution of XG Boost.

4. Results and Discussion



Threshold



Machine Learning

4.1 Threshold

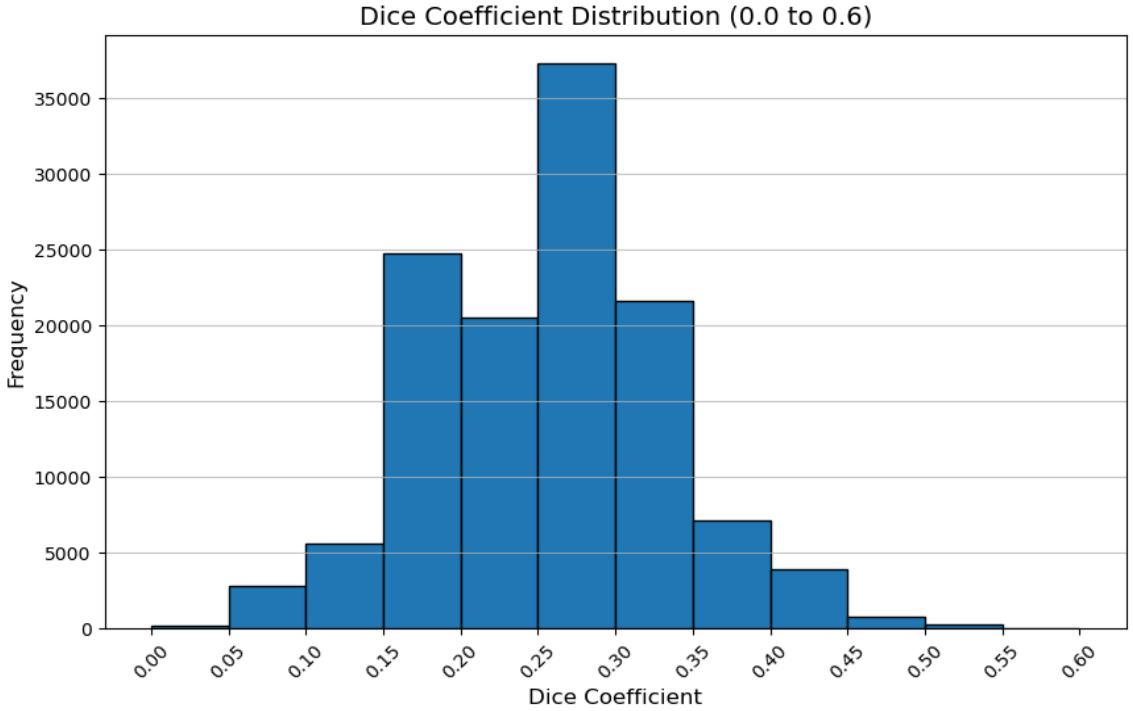
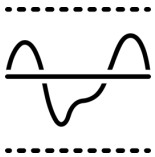


Fig. 4. Histogram of the dice coefficient of all records



4.1 Threshold

BloomFilter1	BloomFilter2	DiceCoefficient	Threshold_0.4	Threshold_0.3	Threshold_0.2	Threshold_0.1
3921124910	1651423879	0.343750	1	0	0	0
3921124910	761781273	0.375000	1	0	0	0
3921124910	1526532195	0.312500	1	0	0	0
3921124910	4013317649	0.406250	0	0	0	0
3921124910	3144879481	0.375000	1	0	0	0

Tab. 8. Comparison of records with dice coefficient and value of thresholds

4.1 Threshold

Threshold	True Positives	True Negatives	False Positives	False Negatives	Total Positives	Total Negatives	Precision	Recall	Accuracy
Threshold_0.1	1461	12197	332	59942	61403	12529	0.814835	0.023794	0.184737
Threshold_0.2	16829	9238	3291	44574	61403	12529	0.836431	0.274075	0.352581
Threshold_0.3	45931	3590	8939	15472	61403	12529	0.837088	0.748025	0.669818
Threshold_0.4	59281	578	11951	2122	61403	12529	0.832224	0.965441	0.809649

Tab. 9. Results for threshold, precision, recall and accuracy.

4.2 Machine Learning

	precision	recall
0	0.804006	0.627057
1	0.676584	0.836173
accuracy	0.727994	0.727994

Tab. 10. Results for XG Boost, precision, recall and accuracy.

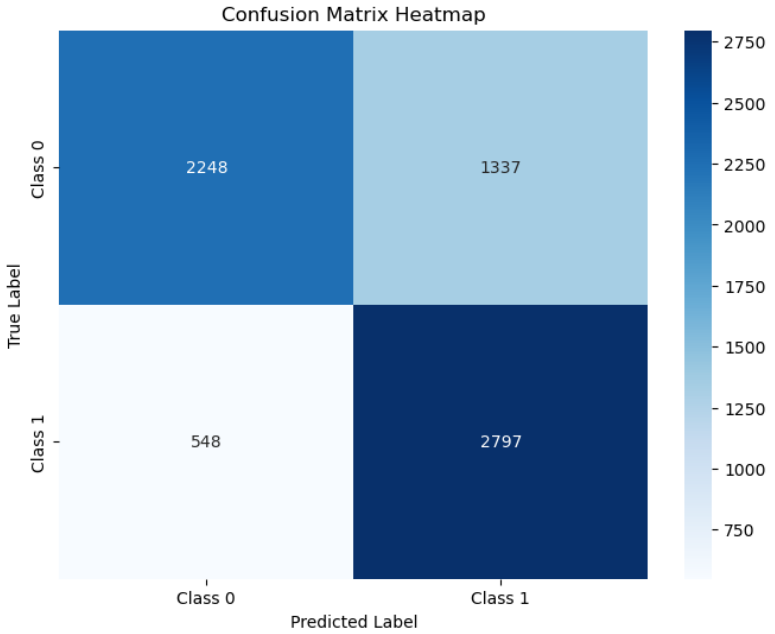


Fig. 5. Confusion matrix for the XG Boost Classifier.

5. Conclusions

1. The use of 32-bit bloom filters can compromise large databases, leading to more false positives.

2. Threshold adjustments are crucial in imbalanced datasets, where lower thresholds mitigate the imbalance's effect by improving recall, but at the cost of precision.

3. Effective handling of imbalanced datasets in machine learning models, through techniques like resampling, weighting, or specialized algorithms, is essential to ensure fair representation of both classes and robust performance across metrics.

6. References

[1] H. Canada, “Substance-related poisonings and homelessness in Canada: a descriptive analysis of hospitalization data.” Accessed: Aug. 27, 2024. [Online]. Available: https://www.canada.ca/en/health-canada/services/opioids/hospitalizations-substance-related-poisonings-homelessness.html

[2] S. C. Government of Canada, “Homeless Shelter Capacity in Canada from 2016 to 2022, Infrastructure Canada.” Accessed: Aug. 27, 2024. [Online]. Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410035301

[3] *Homelessness data snapshot: the National Shelter Study 2022 update*, [Cat. No.: T94-60/2024E-PDF]. Ottawa, Ontario: Infrastructure Canada, 2024.

[4] “2022-2023 Report To The Community | Calgary Drop-In Centre.” Accessed: Aug. 28, 2024. [Online]. Available: https://calgarydropin.ca/2022-2023-report-to-the-community/

[5] C. Housing, “Supporting Calgarians experiencing homelessness,” https://www.calgary.ca. Accessed: Dec. 02, 2024. [Online]. Available: https://www.calgary.ca/content/www/en/home/social-services/homelessness-in-calgary.html

[6] A. Gkoulalas-Divanis, D. Vatsalan, D. Karapiperis, and M. Kantarcioglu, “Modern Privacy-Preserving Record Linkage Techniques: An Overview,” *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 4966–4987, 2021, doi: 10.1109/TIFS.2021.3114026.

[7] S. M. Randall, A. M. Ferrante, J. H. Boyd, J. K. Bauer, and J. B. Semmens, “Privacy-preserving record linkage on large real world datasets,” *J. Biomed. Inform.*, vol. 50, pp. 205–212, Aug. 2014, doi: 10.1016/j.jbi.2013.12.003.

[8] R. Schnell, T. Bachteler, and J. Reiher, “Privacy-preserving record linkage using Bloom filters,” *BMC Med. Inform. Decis. Mak.*, vol. 9, no. 1, p. 41, Dec. 2009, doi: 10.1186/1472-6947-9-41.

[9] L. R. Dice, “Measures of the Amount of Ecologic Association Between Species,” *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945, doi: 10.2307/1932409.