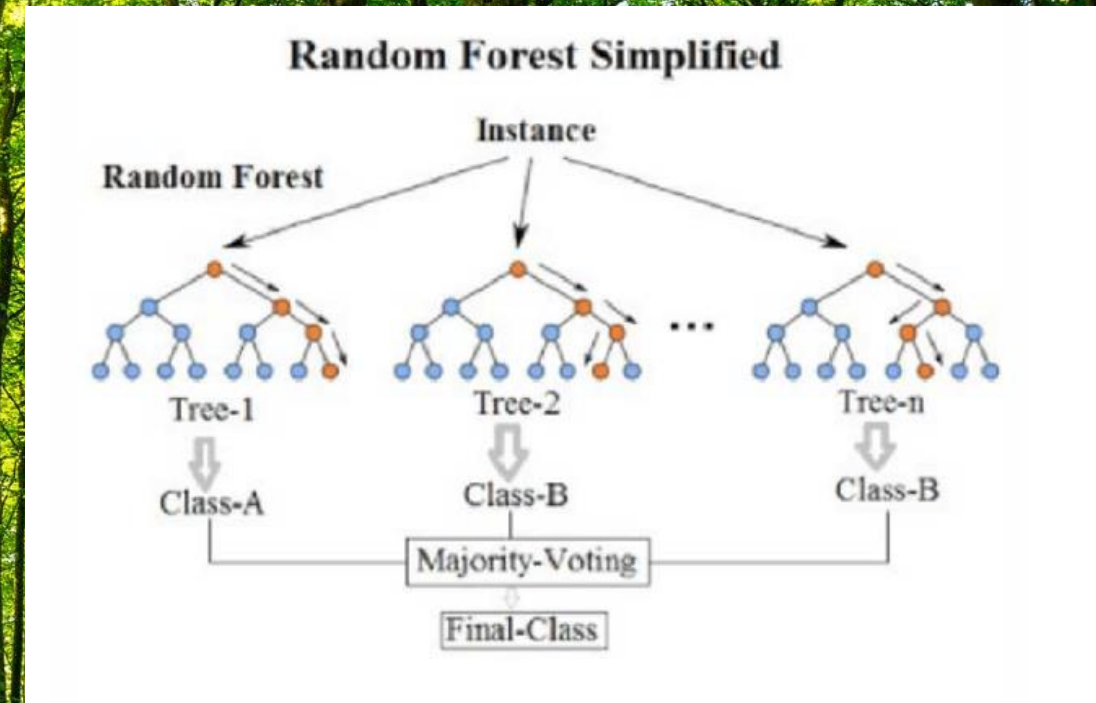
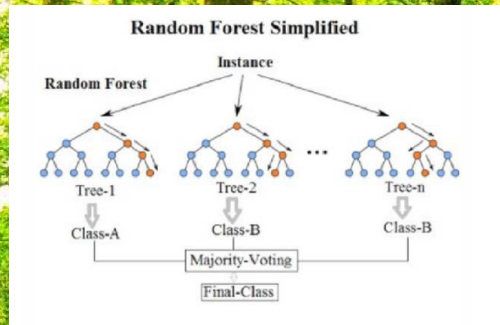


Klasyfikacja Metagenomiczna: MinHash

Piotr Kupidura

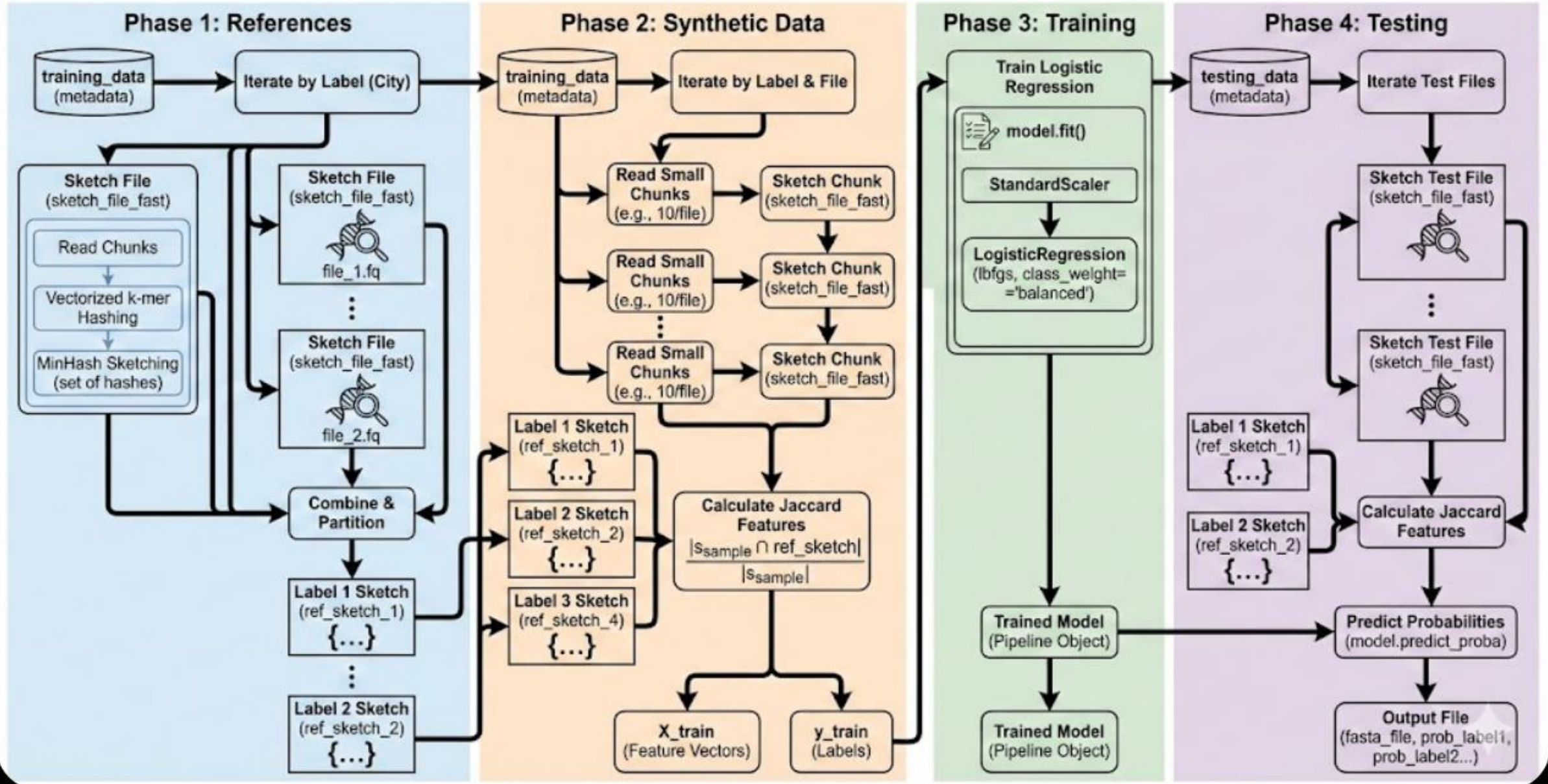
Michał Stanowski

Jakub Giezgala



RANDOM FOREST (AUC: 0.78)

OVERVIEW OF THE MINHASH + LOGISTIC REGRESSION APPROACH



FASTER & FASTER (Wang & NumPy)

```
def vectorized_hash(arr: np.ndarray) -> np.ndarray:
    """ Invertible integer mixing function (Wang Hash) for randomization.
    # MurmurHash3 64-bit a bit faster but get worse AUC
    key = arr.astype(np.uint64)
    key ^= key >> np.uint64(33)
    key *= np.uint64(0xff51afd7ed558ccd)
    key ^= key >> np.uint64(33)
    key *= np.uint64(0xc4ceb9fe1a85ec53)
    key ^= key >> np.uint64(33)
    """
    key = arr.copy()
    key = (~key) + (key << 21)
    key = key ^ (key >> 24)
    key = (key + (key << 3)) + (key << 8)
    key = key ^ (key >> 14)
    key = (key + (key << 2)) + (key << 4)
    key = key ^ (key >> 28)
    key = key + (key << 31)
    return key
```



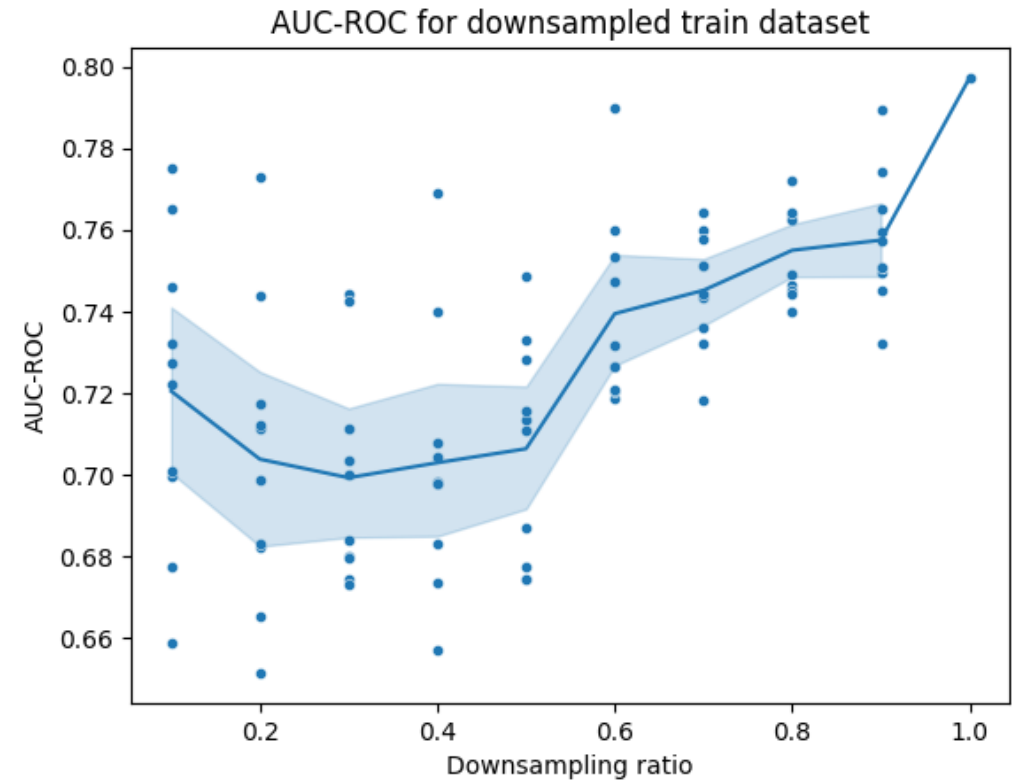
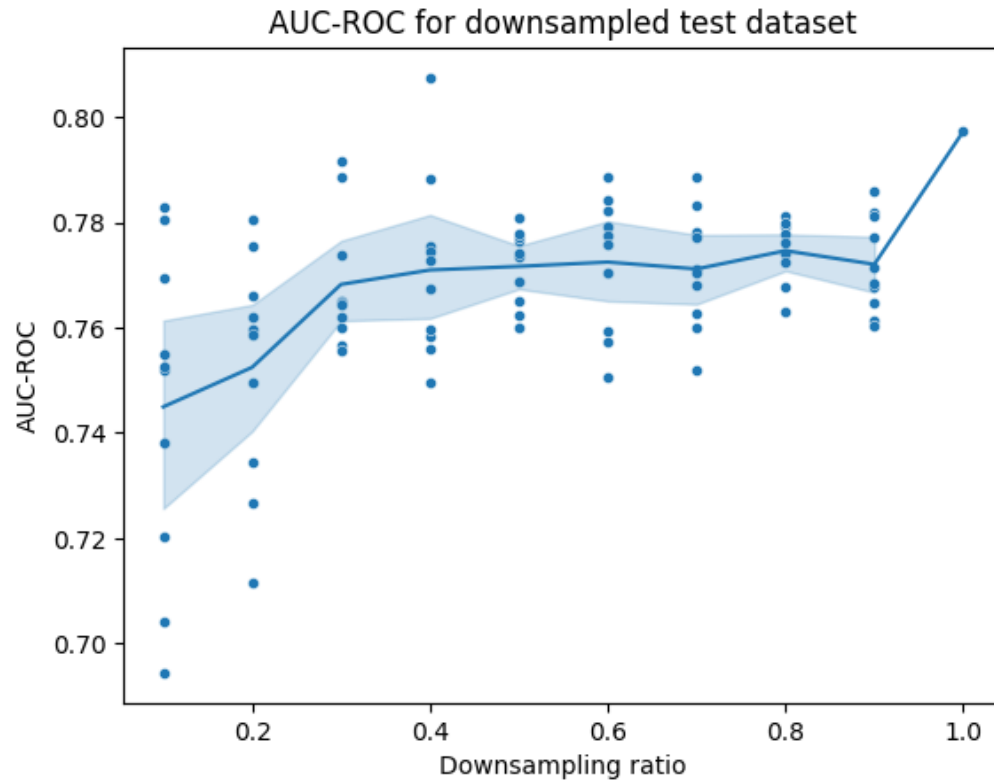
My friend showing me how his 1000 lines of C++ code is 100x faster than my 10 line Python code.



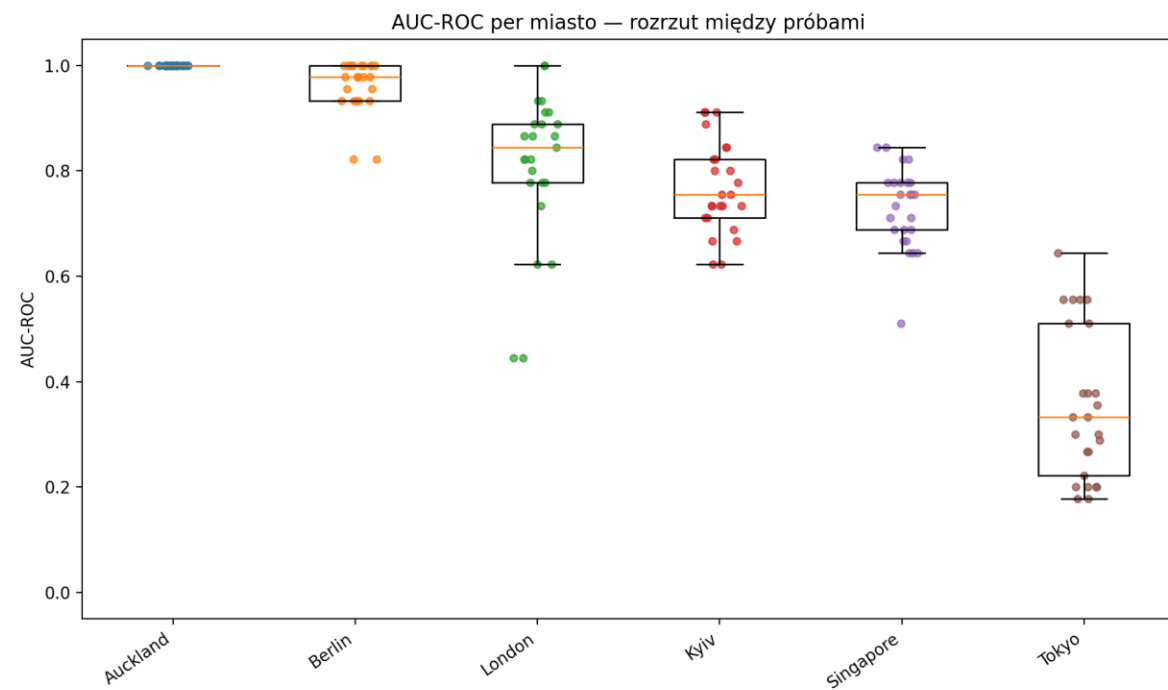
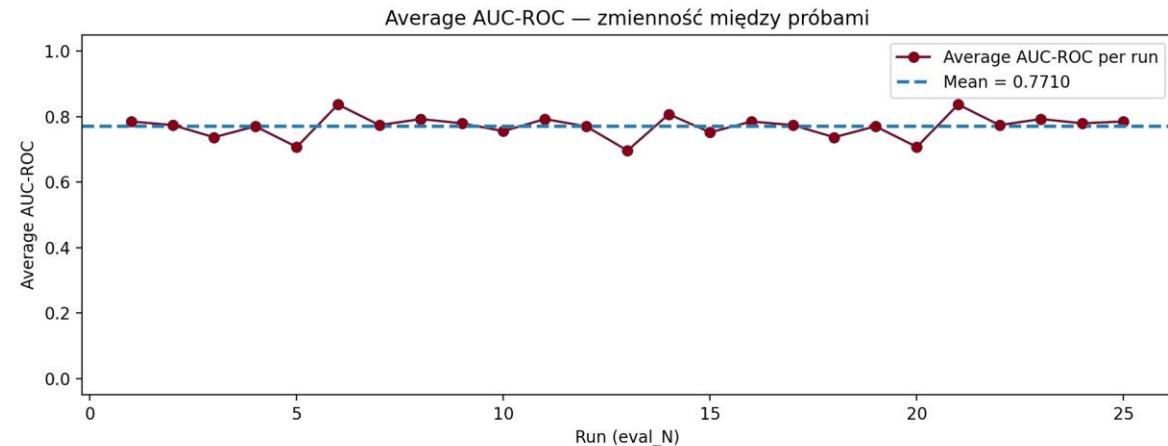
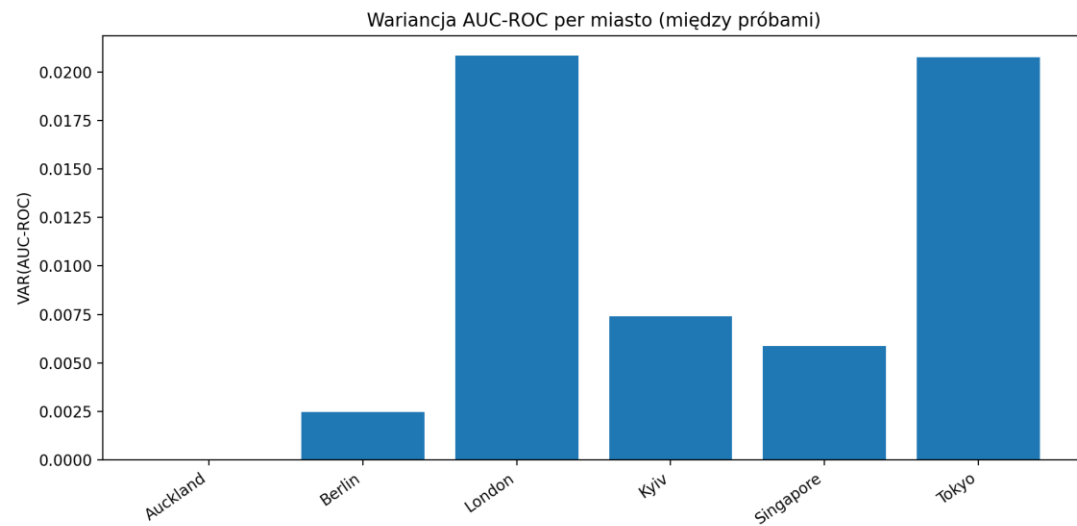
GRID SEARCH

- **DEFAULT_K = 25** -> Długość k-meru, na którym operuje algorytm. Wysoka wartość pozwala na dużą specyficzność biologiczną. Pozwala na precyzyjne odróżnianie sygnatur, minimalizując ryzyko przypadkowych dopasowań.
- **DEFAULT_SEED = 424242** -> Zapewnia determinizm i powtarzalność przy inicjacji modelu klasyfikatora (Regresji Logistycznej).
- **CHUNK_SIZE_KB = 75** -> Rozmiar fragmentu (w kilobajtach), na jakie dzielone są pliki treningowe w fazie generowania syntetycznego podziału, w celu uczenia na mniejszym podzbiorze.
- **DEFAULT_SKETCH_SIZE = 7500** -> Liczba najmniejszych wartości funkcji skrótu (hashy) przechowywanych w "szkicu" dla każdej próbki. Szkic zajmuje w pamięci dokładnie 7500×8 bajtów (dla 64-bitowych liczb całkowitych) ≈ 60 KB na próbkę. Im większy rozmiar, tym mniejszy błąd statystyczny przy porównywaniu próbek ($\text{Error} \approx 1/\text{SketchSize}$)

DOWNSAMPLING



EVALUATION ON DIFFERENT SUBSETS



REFERENCES



Mourad Azhari, Altaf Alooui, Zakia Acharoui, Badia Ettaki - *"Adaptation of the random forest method: solving the problem of pulsar search"*.



Nicholas A. Bokulich - *"Integrating sequence composition information into microbial diversity analyses with k-mer frequency counting"*.



Debopriya Ghosh, Javier Cabrera - *"Enriched Random Forest for High Dimensional Genomic Data"*.



Kenneth S Katz, Oleg Shutov, Richard Lapoint, Michael Kimelman, J Rodney Brister, Christopher O'Sullivan - *"STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions"*.