Rishi Gandhi

COMP 3202 - Extra Assignment 3

202014908

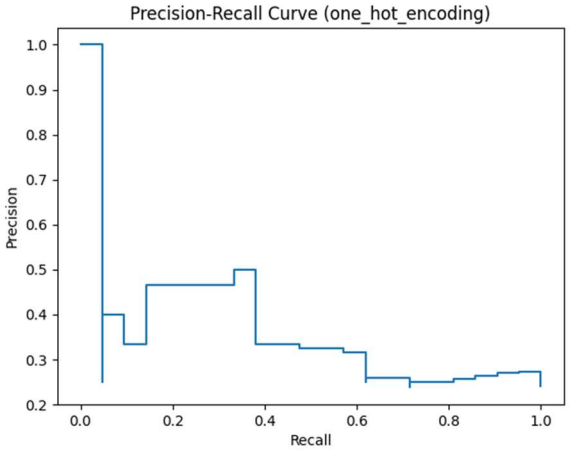
1. **Data Encodings Evaluated:**
    a. **k-mer frequencies (kmer_freq_2 and kmer_freq_3):** These encodings represent the frequency of each possible substring of length k (k-mer) within the DNA sequences. By considering different values of k (2 and 3), we capture different levels of sequence information.
    b. **One-Hot Encoding:** This encoding represents each nucleotide in the DNA sequence as a binary vector, where each position in the vector corresponds to a specific nucleotide (A, C, G, T). It provides a binary representation of the presence or absence of each nucleotide at each position.
    c. **Nucleotide Composition:** This encoding represents the proportion of each nucleotide (A, C, G, T) in the DNA sequence. It captures the overall composition of the sequence in terms of the frequency of each nucleotide.
2. **Tree-Based Ensemble Method Used:**
    a. **Random Forest Classifier:** Random forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It's robust to overfitting, handles high-dimensional data well, and provides feature importance measures.

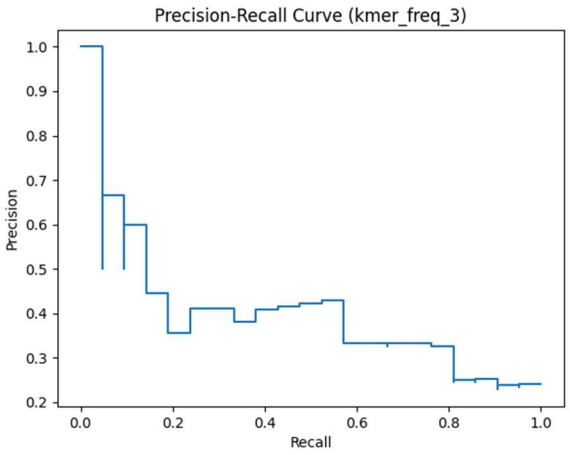
3. **Performance Metric Used to Assess Model Performance:**
    a. **Mean Average Precision:** Mean Average Precision (MAP) is a metric commonly used for evaluating the performance of binary classifiers, particularly in imbalanced datasets. It computes the average precision-recall area under the curve.
    b. **F1-Score:** F1-Score is the harmonic mean of precision and recall, providing a balance between the two metrics. These metrics are suitable for assessing classification performance in this context, especially considering the imbalance in the dataset.
4. **Table with Mean Average-Precision-Score and F1-Score:**

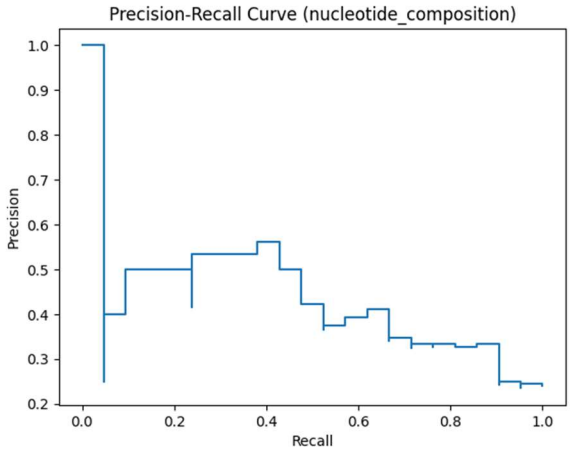
Encoding	Mean Average Precision	Standard Average Precision	Mean F1 Score	Standard F1 Score
kmer_freq_2	0.502237	0.071834	0.287937	0.075299
kmer_freq_3	0.328127	0.079569	0.042091	0.042168
one_hot_encoding	0.331062	0.040108	0.000000	0.000000
nucleotide_composition	0.412528	0.079623	0.324758	0.100233

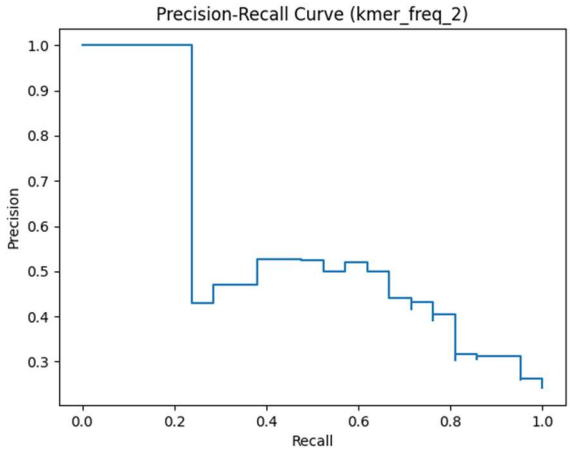
The table displays the mean average-precision-score and F1-score along with their standard deviations per encoding. The best encoding found based on the provided metrics is indicated in the "Best Encoding" column. In this case, the best encoding is `kmer_freq_2`.

5. Figures









6. Acknowledgements

a. NumPy: Oliphant TE. "A guide to NumPy." USA: Trelgol Publishing (2006).

b. Pandas: McKinney, Wes. "Data Structures for Statistical Computing in Python," Proceedings of the 9th Python in Science Conference, vol. 445, pp. 56–61, 2010.

c. ChatGPT – www.chat.openai.com (Used for debugging)

d. scikit-learn: Pedregosa F, Varoquaux G, Gramfort A, et al. "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research 12 (2011): 2825-2830.

e. Matplotlib: Hunter JD. "Matplotlib: A 2D Graphics Environment." Computing in Science & Engineering 9, no. 3 (2007): 90–95.

f. scikit-bio: Amir A, McDonald D, Navas-Molina JA, et al. "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns." mSystems 2, no. 2 (2017): e00191-16.