Task 4: Critical Analysis Report

# 1. TF-IDF vs. One-Hot Encoding: Performance Insights

* Why TF-IDF Performed Better

- Informative Weighting: Unlike one-hot encoding, TF-IDF gives higher weights to important but less common terms. This enhances the model’s ability to differentiate diseases that share common symptoms but differ in rare ones.

- Denser Representations: TF-IDF produced a feature matrix of shape (25, 1020) with ~93% sparsity, compared to one-hot’s (25, 496) with ~95% sparsity. Though both are sparse, TF-IDF’s larger and more expressive feature set led to better performance.

- Vocabulary Sensitivity: TF-IDF supports frequency-based learning, whereas one-hot encoding considers only presence or absence. This makes TF-IDF better suited to capture nuanced patterns in symptoms and risk factors.

* When One-Hot Might Still Be Useful

- For simple rule-based systems or highly structured symptom vocabularies, one-hot encoding remains interpretable and computationally lighter.

# 2. Clinical Relevance of the Results

- In model evaluations, KNN with TF-IDF and cosine similarity consistently yielded the highest performance (up to 64% accuracy and 49% F1-score).

- Dimensionality reduction using PCA and SVD revealed that diseases naturally clustered more clearly with TF-IDF vectors than with one-hot encoded data.

- These clusters corresponded well to manually labeled clinical categories (e.g., cardiovascular, respiratory), indicating that the representation aligns with real-world medical groupings.

# 3. Limitations of Encoding Approaches

* One-Hot Encoding:

- No Frequency Context: It cannot distinguish between commonly occurring and rare symptoms.

- High Dimensionality: Resulted in a matrix of 496 features that’s highly sparse and less informative.

- Zero Context Awareness: Treats all symptoms equally, ignoring clinical importance.

* TF-IDF Encoding:

- Bag-of-Words Limitation: Doesn’t account for word order or negation (e.g., “no fever” = “fever”).

- Needs Textual Data: Relies on availability of free-text symptom descriptions.

- Sensitive to Vocabulary Noise: Without stop-word filtering or stemming, less relevant terms might skew weights.

# Conclusion

TF-IDF, especially when paired with cosine similarity, emerged as the more effective encoding method for disease classification based on symptom text. It captured symptom relationships better and aligned closely with real-world diagnostic groupings.

However, both encoding methods have trade-offs. For future work, integrating context-aware embeddings (like BioBERT) or combining structured symptom data with free-text features could significantly improve diagnostic accuracy and clinical reliability.