Table of Contents

[Task 4: 2](#_Toc195547957)

[Question 1: 2](#_Toc195547958)

[Advantages of TF-IDF: 2](#_Toc195547959)

[Advantages of One-Hot Encoding 2](#_Toc195547960)

[Question 2: 2](#_Toc195547961)

[Implications: 3](#_Toc195547962)

[Question 3: 3](#_Toc195547963)

[TF-IDF Limitations: 3](#_Toc195547964)

[One-Hot Encoding Limitations: 3](#_Toc195547965)

# Task 4:

## Question 1:

TF-IDF (Term Frequency-Inverse Document Frequency) captures not only the presence of a term but also how *unique or important* it is within the corpus. This is crucial in clinical text data, where certain terms like “acute,” “chronic,” or “syndrome” may appear across multiple disease descriptions but are more informative when weighted properly.

### Advantages of TF-IDF:

* Down-weights common medical terms that may appear in many diseases but don't help in distinguishing them.
* Captures importance of disease-specific terminology (e.g., "dissection" vs. "asthma").
* Sparse but informative: It leads to lower dimensionality than one-hot encoding while retaining meaningful word distinctions.

### Advantages of One-Hot Encoding

* For extremely short text where word frequency isn’t meaningful, one-hot could be just as good.
* If only presence/absence matters (and not frequency), one-hot may suffice.
* With a small dataset, one-hot is simpler and less sensitive to outliers or rare terms.

In our case, TF-IDF generally outperformed one-hot encoding, especially in terms of F1-score and precision, suggesting that the frequency-weighted approach helps the models capture more nuanced patterns in symptom descriptions and disease definitions.

## Question 2:

One interesting observation is that diseases grouped under the same clinical category (e.g., "Cardiovascular") often had similar TF-IDF feature profiles. For example:

* Aortic Dissection, Stroke, and Pulmonary Embolism share terms like *“acute,” “chest pain,” “sudden onset,”* which are clinically overlapping and were clustered together in the feature space.
* This suggests that TF-IDF is able to cluster diseases with shared symptomatology, which aligns well with how physicians think about differential diagnosis.

### Implications:

* The model could be helpful in *triaging patients* based on symptom descriptions.
* The semantic closeness revealed by TF-IDF could support clinical decision support tools, guiding doctors toward a probable disease category even before lab or imaging data.

## Question 3:

### TF-IDF Limitations:

* **Context ignorance**: TF-IDF doesn't understand word order or negation. “No chest pain” is treated similarly to “chest pain.”
* **Sensitivity to noise**: Rare or misspelled terms may get inappropriately high weights.
* **Interpretability**: Clinicians may find it harder to interpret which exact features led to the decision, especially in high-dimensional TF-IDF space.

### One-Hot Encoding Limitations:

* **Ignores frequency and importance**: Every word is treated equally, reducing sensitivity to crucial medical terms.
* **Very high dimensionality**: Increases computational load and may lead to overfitting, especially on small datasets.
* **Sparse and less informative**: Makes it harder for models to generalize.