Automate the mapping of patient discharge summaries and diagnosis descriptions to standardized diagnosis codes (e.g., CCSR codes) using Natural Language Processing (TF-IDF with cosine similarity).

**Objective:**

Build a **rapid ML prototype** in 2 weeks to:

* Reduce the manual effort and time medical coders spend assigning codes.
* Build a working prototype in 2 weeks using cleaned hospital datasets.
* Present results and an interactive dashboard to stakeholders for validation.
* **Prototyping Process**

| **Phase** | **Activity** |
| --- | --- |
| **Week 1** | * Collected and cleaned patient discharge data including diagnosis descriptions and CCSR codes. * Implemented TF-IDF vectorization and cosine similarity in a Jupyter notebook to match diagnosis descriptions with the most similar known records. * Created a Matched Code column by linking predicted descriptions to their actual codes. * Evaluated model performance using accuracy, precision, recall, and F1-score. * Visualized similarity score distribution and top predicted codes using histograms and bar plots. |
| **Week 2** | * Developed a Streamlit dashboard enabling users to upload cleaned CSVs and run real-time code prediction * Added performance metrics like the Precision, recall and F1-score * Added interactivity: download final results. * Implemented a live diagnosis prediction input that suggests the best-matched code with its similarity and estimated cost. |

**Outcome:**

* **Dashboard**: Streamlit interface enabled easy diagnosis testing, live prediction, and CSV export.
* **Time/Cost Saved**: Estimated ~60% reduction in manual diagnosis code lookup time.
* **Evaluation metrics** : Accuracy : 1.0000,Precision: 1.0000, Recall : 1.0000, F1 Score : 1.0000