**Medical Billing Data Analysis**

An in-depth exploration and automation of Diagnosis coding via machine learning

**Project Executed by Group 7**

Team Members:

**Sruthi Gangadharan (0812533)**

**Harpreet Kaur (0803155)**

**Tanveer Kaur (0825924)**

**Priteshkumar Surendrabhai Rana (0792969)**

**Executive Summary**

Analysis of medical billing data is crucial to healthcare providers' capacity to remain financially viable. The primary goal of this project is to guarantee the best possible financial health and level of patient care through the extraction and processing of billing and claims data. This project intends to improve operational efficiencies and lower human error by automating the assignment of medical procedure codes from clinical narratives, opening the door for more dependable and efficient healthcare service billing.

**Introduction**

Medical billing is essential to keeping healthcare organizations financially stable in the intricate and ever-changing world of healthcare. The procedure entails carefully reviewing and processing billing data related to patient care services. Our work demonstrates how machine learning can improve efficiency and reduce mistake rates of conventional billing procedures.

**Objective**

**Purpose:**

The goal of the study is to build a predictive machine learning model that can identify medical procedure codes by using keywords taken out of clinical narratives. This innovation aims to replace the laborious and prone to error manual coding process.

**Problem Statement:**

As manual coding relies on human input, it is inherently inefficient and prone to errors. In healthcare settings, timely reimbursements and financial planning depend heavily on accuracy and processing time, both of which can be greatly increased by automating this work.

**Data Overview**

Three key datasets form the foundation of our model training and testing:

dxcode.xlsx: This dataset houses clinical keywords linked with their corresponding medical codes, providing a foundational vocabulary for the model.

proc.xlsx: Contains detailed clinical narratives sourced directly from patient records, offering real-world data for training.

proc2.xlsx: A synthesized dataset that combines elements from the first two, equipped with narratives aligned with procedural codes, ready for direct application in training our model.

**Methodology**

**Data Preprocessing:**

To guarantee the quality of the data input into the models, a comprehensive cleanup was carried out on the datasets. This involved normalizing text data, removing trailing spaces, and making sure all entries were consistent.

**Model Exploration and Selection:**

Decision Trees and Random Forests: These more straightforward models, which are renowned for their interpretability but are constrained by overfitting, were used in the initial exploration phases. To overcome some of these restrictions, we extended to random forests, but we discovered that their accuracy was insufficient for our purposes.   
Final Decision: Classifier Chain technique with Logistic Regression Model: Following basic model exploration, we switched to a Classifier Chain technique with a logistic regression model. We were able to skillfully handle the multi-label nature of medical coding through chained predictions while retaining the ease of use and efficacy of logistic regression for binary classification.

A graph of different colored lines

Description automatically generated with medium confidence

**Implementation Details**

Model Training using TF-IDF Vectorization: To highlight the significance of distinct phrases present in clinical narratives, we utilized TF-IDF vectorization to transform textual data into a format that can be numerically analyzed. Then, in order to account for interdependencies among codes, each label was handled successively using logistic regression, incorporating results from earlier labels into later training phases.

**Technological Stack:**

* Python as the core programming language.
* Pandas and NumPy for data manipulation and numerical operations.
* Scikit-learn for implementing machine learning models.
* Model Performance and Accuracy

The logistic regression model's training accuracy of 99% was excellent. It demonstrated its effectiveness in managing the intricate requirements of medical procedure coding by correctly predicting complex medical codes from a provided narrative with a high probability in a practical setting.

**Challenges Encountered and Resolutions**

Data Sparsity and Dimensionality: Training was difficult because of scant data because some medical procedures are uncommon. Moreover, it was initially difficult to handle the increased dimensionality that TF-IDF introduced.   
Solutions Put into Practice: The understanding and prediction of related labels in our model was much enhanced using Classifier Chains. Despite the enormous feature set, regularization approaches were improved to maximize model performance.

**Conclusion and Future Directions**

The potential to simplify medical procedure coding has been effectively illustrated by this study, greatly reducing the labor of medical coders and improving the precision and effectiveness of medical billing systems. In order to further improve accuracy and dependability, we plan to broaden our dataset to encompass a wider range of medical scenarios and investigate increasingly complex models, such as neural networks.

**Acknowledgments**

We would like to express our gratitude to the data scientists and healthcare experts whose contributions of data and insights made this project possible.