Project Title:-

Predicting Treatment Outcomes from Patient Clinical data

GitHub link : https://github.com/Nimit99/Capstone_606

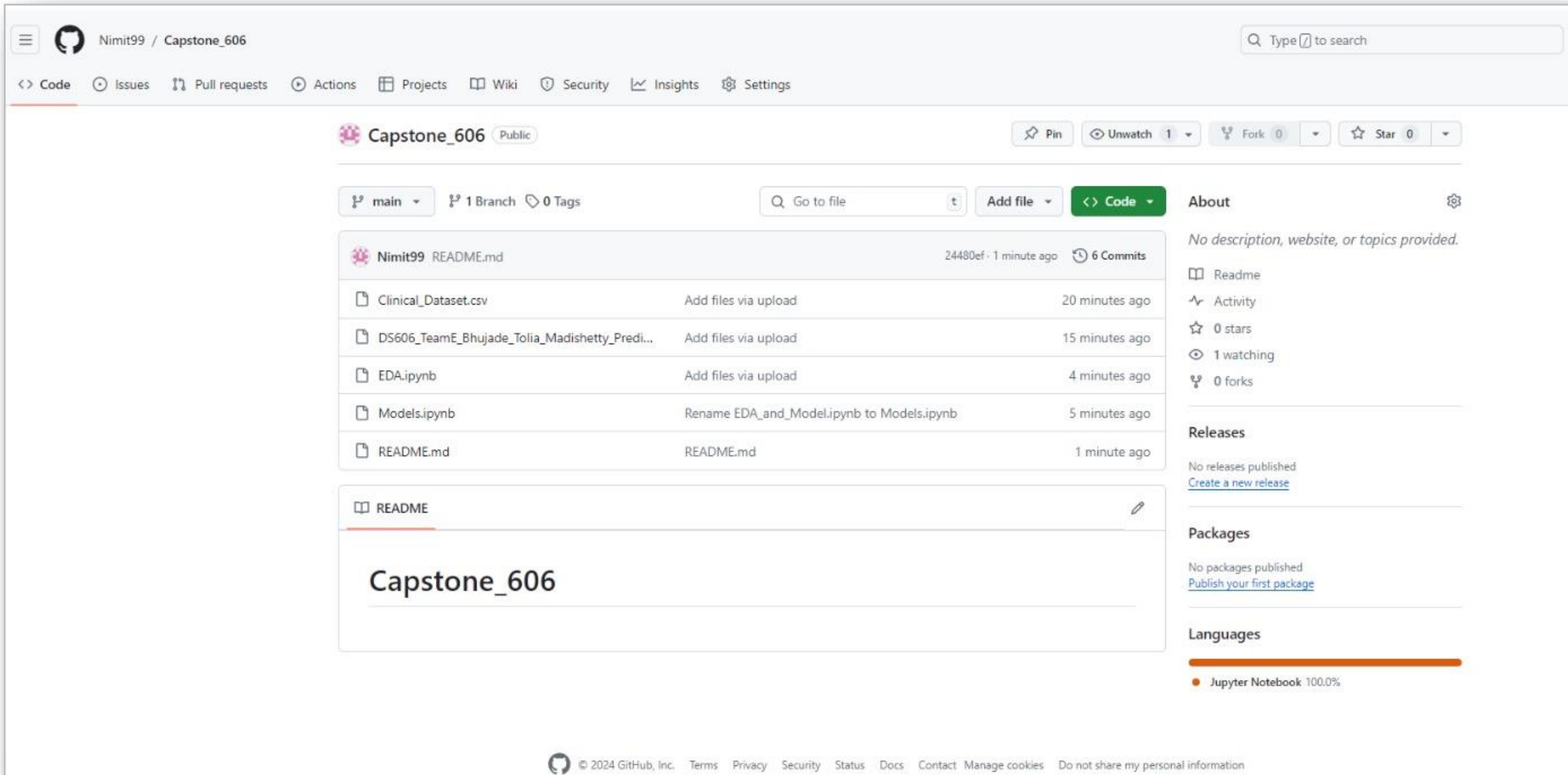
Prepared by- Team E

1. Ayushi Bhujade (ZG28331)
2. Nimit Tolia (YT29650)
3. Sriteja Madishetty (VT76695)

TABLE OF CONTENT

01. GITHUB DETAILS

GitHub link: https://github.com/Nimit99/Capstone_606



02. PROJECT OVERVIEW (1)

Introduction

- In the healthcare industry, the ability to accurately predict treatment outcomes based on patient clinical data can significantly enhance patient care and optimize medical resources.
- This **project leverages advanced machine learning techniques - Natural Language Processing** to predict the most suitable treatment for patients based on their comprehensive clinical history and current medical condition.

Objective

- The primary objective of this project is to **develop a robust predictive model** that can analyze extensive **textual clinical data** and accurately **predict the appropriate treatment name**.
- The predictive model is designed to process and interpret complex medical information like patient demographics, physiological context, visit motivation, diagnosis results, and related conditions.
- The ultimate goal is to provide precise treatment recommendations based on data-driven insights.

Data Source

- The dataset used in this project is sourced from Hugging Face, specifically the augmented clinical notes dataset (AGBonnet/augmented-clinical-notes · Datasets at Hugging Face). **This dataset comprises a wide array of textual and numerical clinical features of patients.**

02. PROJECT OVERVIEW - Methodology (2)

Data Selection

- A subset of the original dataset was selected to ensure manageable data size and relevance.
- The selected subset consists of 17,000 records, 11 columns and **100 treatment classes**.

Data Pre - processing

- Given that the dataset is primarily textual, several preprocessing steps were undertaken
- **Removing stop words, eliminating punctuation,** and **performing lemmatization** to standardize the text.

Feature Engineering

- Creating relevant features using **TF-IDF Vectorizer**. This converts text data into numerical features and captures the importance of words in documents relative to the entire dataset.

Model Development

- Training different ML models, that can handle the datasets with high dimensionality.
- Fine-tuning Model Parameters and choosing the high performance model for our dataset.

Evaluation:

- Splitting the dataset into an 80-20% train-test split to evaluate model performance on unseen data.
- Performing cross-validation to ensure the model's robustness and reliability.
- Assessing the model's performance using **accuracy, precision, recall, F1-score and AUC.**

03. LITERATURE SURVEY (1)

Project/Study	Objective	Techniques Used	Key Contributions	Citation
MediNote	Generate clinical notes from dialogues	Fine-tuning large language models (LLMs) such as MediNote-7B and MediNote-13B	Automated clinical documentation to enhance efficiency and accuracy	[4]
MediTron	Clinical LLM development	Pre-trained on PubMed articles and clinical guidelines, fine-tuned with the dataset	Improved generation of detailed and structured clinical notes	[1]
Literature-Augmented Clinical Outcome Prediction	Enhance clinical outcome predictions	Sparse and dense retrieval models for integrating biomedical literature with clinical notes	Improved accuracy in outcome predictions by using both clinical notes and relevant literature	[2]
NoteChat	Extend PMC-Patients with synthetic dialogues	Synthetic dialogues generated using ChatGPT and GPT-4	Realistic training data mimicking real-world clinical interactions	[3]
Structured Patient Information Extraction	Extract structured data from clinical notes	Using GPT-4 to extract structured patient information	Enhanced structured data extraction from unstructured text for better training of predictive models	[1]

03. LITERATURE SURVEY (2)

Given below are the aspects of our approach and objectives that differs from the previous efforts utilized on Augmented Clinical Notes Dataset:

Aspect	Other Works	Our Work
Objective	Generate clinical notes from dialogues (e.g., MediNote, MediTron)	Develop a predictive model specifically for predicting appropriate treatment names based on structured data
Data Utilization	Utilized Synthetic dialogues and structured patient information for note generation and structured data extraction	Leverage structured data from clinical notes summary for building a treatment prediction model
Modelling Techniques	NLP techniques and large language models (LLMs) for text generation and integration with literature	NLP techniques (vectorization) and ML statistical modeling techniques for predictive analytics (e.g., decision trees, random forests)
Feature Engineering	Text processing and synthesis of patient-doctor dialogues	Detailed feature engineering from structured data fields (e.g., patient demographics, diagnosis, medical history)
Enhanced Model Accuracy	Evaluated on the quality of generated notes and literature integration	Evaluated on classification and prediction accuracy for treatment names

04. EXPLORATORY DATA ANALYSIS - Data Overview (1)

#Columns: 11 #Rows: 17,000 #Treatment Classes: 100

	age	sex	visit_motivation	physiological_context	admission_reason	diagnosis_test	diagnosis_result	related_condition	reason_for_treatment	treatment_name
0	19	male	nocturnal cough and bilateral lower limb swelling	diagnosed with hypertension and stage 5 chronic kidney disease	further evaluation and management of hypertension and stage 5 chronic kidney disease	creatinine level	elevated	hypertension	to manage hypertension	amlodipine
1	13	female	referred by orthopedic surgeon due to persistent pain after in situ screw fixation for scfe	pain in left hip and knee after injury while doing gymnastics	correction osteotomy according to southwick with re-screw fixation	anteroposterior radiograph of the pelvis	no abnormalities seen	pain in left hip and knee	to alleviate pain from injury	physical therapy
2	44	male	recurrent postprandial epigastric pain	third hospitalization for the same complaint, gastric erosions found in gastroduodenoscopy three months back	recurrent postprandial epigastric pain	abdominal radiograph	multiple air fluid levels suggestive of intestinal obstruction	subacute intestinal obstruction	to manage intestinal obstruction	intravenous fluids
3	51	female	increased watery diarrhea with occasional blood and cramping abdominal pain	ulcerative colitis for 5 years	nonradiating chest pain located at the midsternal region, shortness of breath, and worsening fatigue	stool studies including stool cultures, stool ova, and parasites	negative	ulcerative colitis	lack of response to oral prednisone	infliximab

04. EXPLORATORY DATA ANALYSIS - Data Overview (2)

```
--------Datatypes of all columns-----------
age                       object
sex                       object
physiological_context     object
visit_motivation          object
admission_reason          object
diagnosis_test            object
diagnosis_result          object
diagnosis_condition       object
related_condition         object
reason_for_treatment      object
treatment_name            object
dtype: object
```

Fig (i)

```
--------Null values in dataframe:-----------
age                         9
sex                       104
physiological_context    2826
visit_motivation          338
admission_reason         2488
diagnosis_test            701
diagnosis_result         1412
diagnosis_condition      6383
related_condition         678
reason_for_treatment     1192
treatment_name              0
dtype: int64
```

Fig (ii)

Action: Handle missing values –

- Drop column *diagnosis_condition.*
- Delete records with null input

Result: The shape of the new data is : Rows - 10,093, Columns - 10

04. EXPLORATORY DATA ANALYSIS - Gender Distribution (3)

```
---------Value count of Sex column-----------
 sex
male                    5110
female                  3710
woman                    997
man                      156
boy                       59
girl                      52
gentleman                  6
male (geminus a)           2
sex not specified          1
Name: count, dtype: int64
```

Fig (i)

Action: Normalize Gender Representation to two major categories: Male, Female

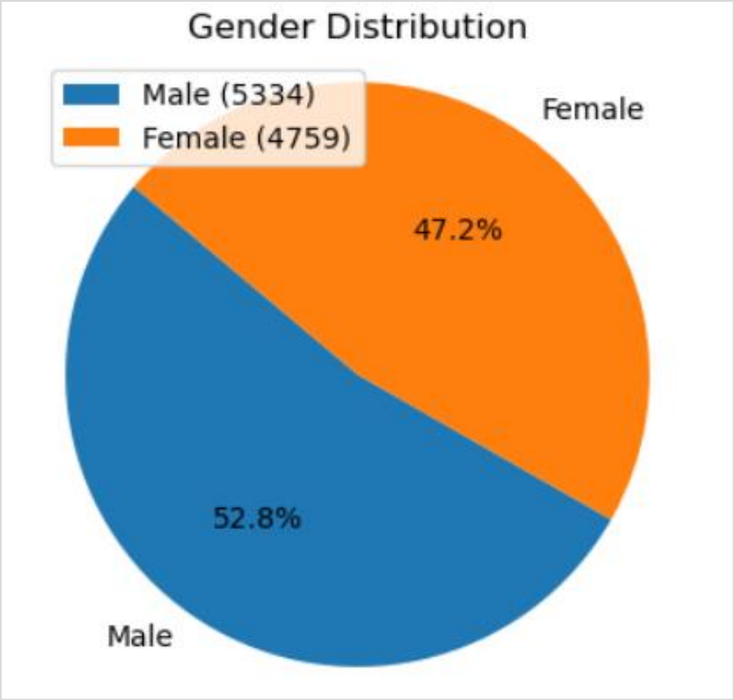


Fig (ii)

04. EXPLORATORY DATA ANALYSIS - Age Distribution (3)

```
array(['one year', '35', '74', '64', '85', '10 years old', '72', '44',
       '56 years old', '4 years old', '83', '51', '62', '16', '46', '56',
       '53', '58', '26', '57', '67', '40', '19', '37', '50 years old',
       '54', '45', '10-year-old', '61', '55 years old', '79', '27', '20',
       '32', '48', '17', '63', '8 years old', '68 years old', '39',
       '36 years old', '24', '29', '82', '65', '59', '70 years old', '68',
```

Fig (i)

Action: Extract the numerical value from textual information of age using regular expression match

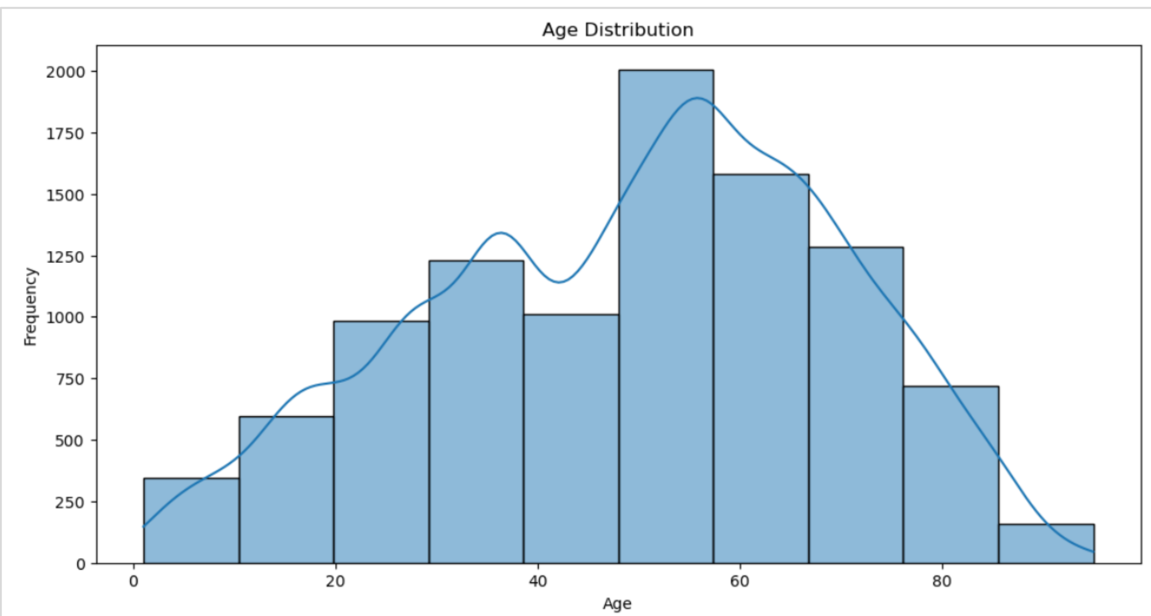


Fig (ii)

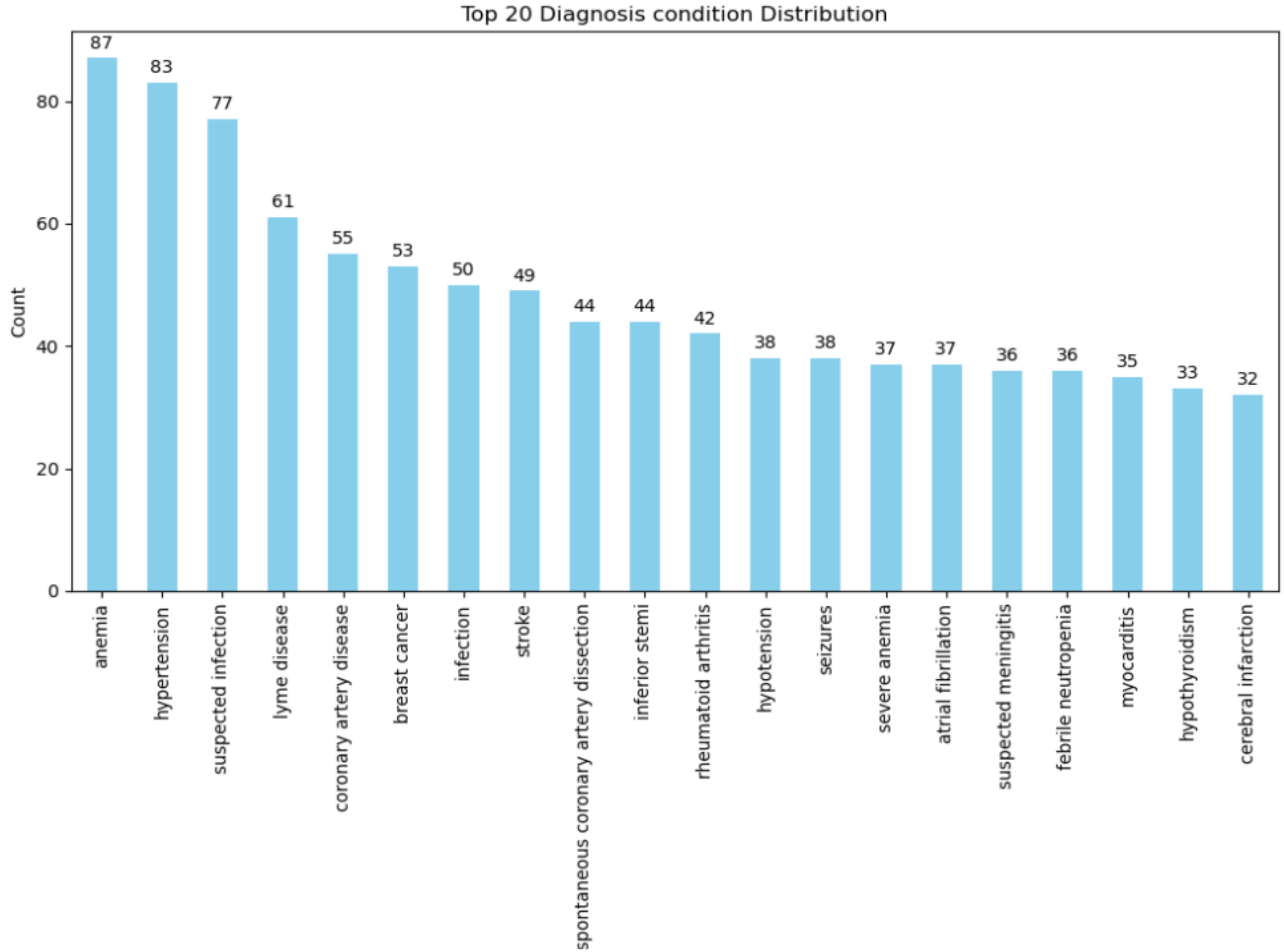
```python
import regex as re

def extract_age(age):
    if pd.isna(age):
        return None
    match = re.search(r'\d+', str(age))
    if match:
        return int(match.group())
    return None

# Apply the extraction function
df['age_num'] = df['age'].apply(extract_age)
```

Fig (iii)

04. EXPLORATORY DATA ANALYSIS - Top Diagnosis Conditions (4)



04. EXPLORATORY DATA ANALYSIS - Major visit motivations (5)

Topic Discovery: Applied Latent Dirichlet Allocation (LDA) to the "visit_motivation" column to uncover hidden topics or themes within the text data.

Insight Generation: LDA allows to gain insights into the common reasons behind medical visits, in understanding patterns and trends in patient concerns or clinical focuses

```
Topics in LDA model:
Topic #1: pain cough fever, dyspnea abdominal symptoms chills, weakness generalized worsening
Topic #2: left numbness limb transient routine blood follow-up treatment discovered injury
Topic #3: progressive weakness right loss consciousness decreased headache, eye recurrent right-sided
Topic #4: right pain fall left flank lower limb complaints mass chest
Topic #5: pain left swelling right headache complaints knee upper persistent epigastric
Topic #6: weight loss lower worsening months episodes bilateral extremity significant syncope
Topic #7: pain abdominal onset sudden lower pain, acute vomiting upper severe
Topic #8: shortness chest breath mass pain left worsening progressive discomfort evaluation
Topic #9: seizure progressively fever episode complaints fatigue headaches mild arm generalized
Topic #10: right respiratory management pain, severe distress evaluation left fever increased
```

Fig (i)

04. EXPLORATORY DATA ANALYSIS - Identify diagnosis tests (6)

Identifying Medical Topics: Applying LDA to diagnosis texts helps identify common themes or topics within medical diagnoses.

Insight Generation: LDA-generated topics helps to understand the distribution of diagnoses across patient populations.

```
Diagnosis Tests Topics in LDA model:
Topic #1: ct scan mri chest abdomen brain contrast head x-ray pelvis
Topic #2: imaging ultrasound resonance magnetic angiography (mri) investigations echocardiogram radiographs echocardiography
Topic #3: blood cultures urine serum level test count tests levels troponin
Topic #4: tomography computed (ct) examination abdominal laboratory scan tests culture histopathological
Topic #5: biopsy angiography laboratory analysis coronary liver function renal fluid aspiration
```

Fig (i)

04. EXPLORATORY DATA ANALYSIS - Code for LDA (7)

```
# Step 1: Tokenize and preprocess the text
# Example function to preprocess text (you can customize this based on your needs)
def preprocess_text(text):
    tokens = text.lower().split()  # Split text into tokens
    tokens = [token for token in tokens if token not in ENGLISH_STOP_WORDS]  # Remove stopwords
    return tokens

# Apply preprocessing to your 'visit_motivation' column
processed_docs = df['diagnosis_test'].apply(preprocess_text)

# Step 2: Create Document-Term Matrix using CountVectorizer
vectorizer = CountVectorizer(tokenizer=lambda x: x, lowercase=False)
doc_term_matrix = vectorizer.fit_transform(processed_docs)

# Step 3: Train LDA model
num_topics = 5  # You can adjust this number based on your specific needs
lda = LatentDirichletAllocation(n_components=num_topics, random_state=42)
lda.fit(doc_term_matrix)

# Print topics and their top words
def print_top_words(model, feature_names, n_top_words):
    for topic_idx, topic in enumerate(model.components_):
        message = f"Topic #{topic_idx + 1}: "
        message += " ".join([feature_names[i] for i in topic.argsort()[:-n_top_words - 1:-1]])
        print(message)

n_top_words = 10  # Number of top words to display per topic
print("\nDiagnosis Tests Topics in LDA model:")
print_top_words(lda, vectorizer.get_feature_names_out(), n_top_words)
```

Fig (i)

04. EXPLORATORY DATA ANALYSIS - Popular Treatments (8)

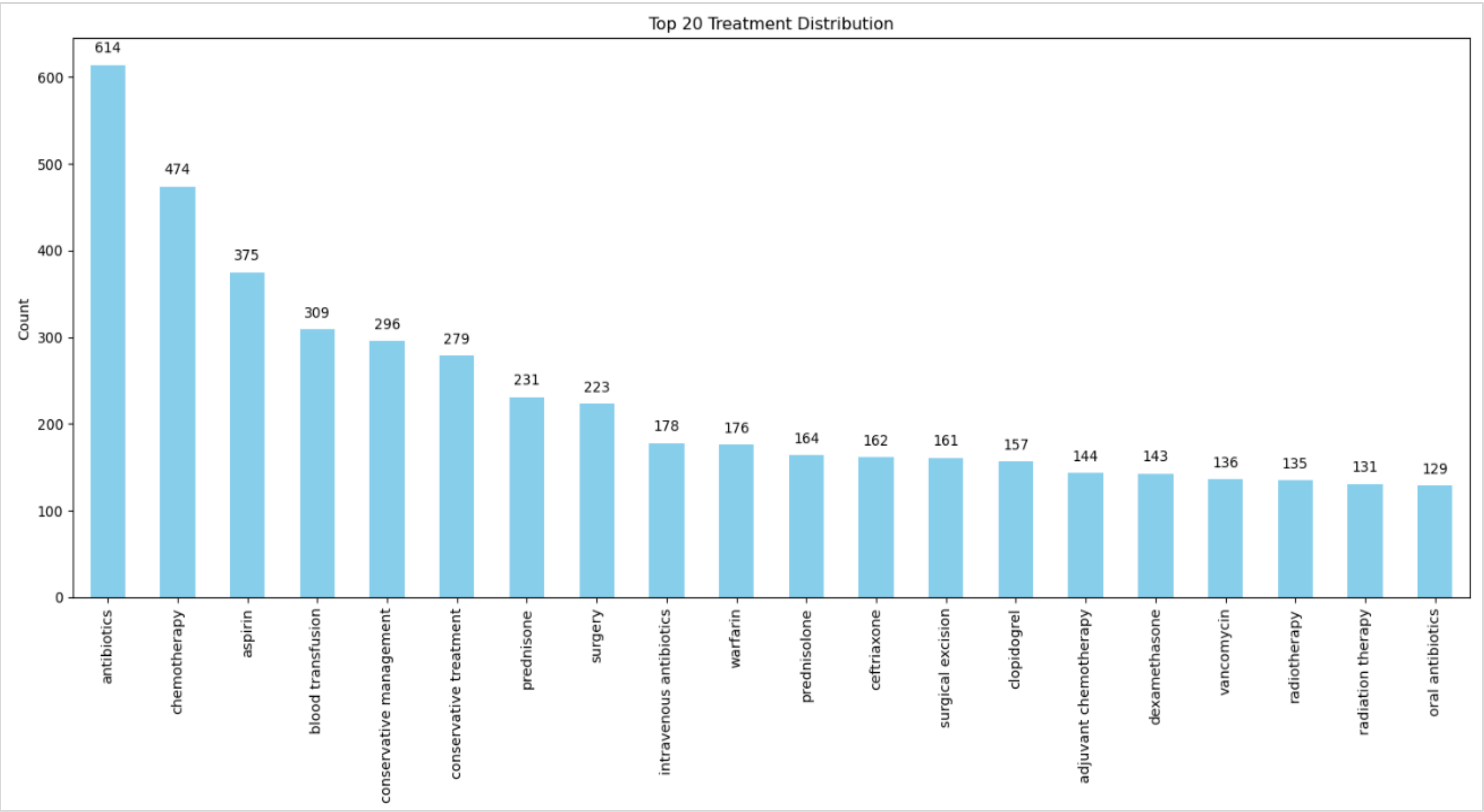


Fig (i)

05. FEATURE ENGINEERING

Standardize Treatment Names

- Objective: Reduce variability in treatment names.
- Approach: Group synonyms and variations under a standard name.
- Eg: ("antibiotics", "antibiotic therapy", "antibiotic treatment" → "antibiotics").
- Keep the treatment classes with over 50 records; remove the rest.

Encode Target Variable

- Objective: Convert categorical target variable into numerical format.
- Approach: Use **LabelEncoder** for transformation.

Text Data Preprocessing

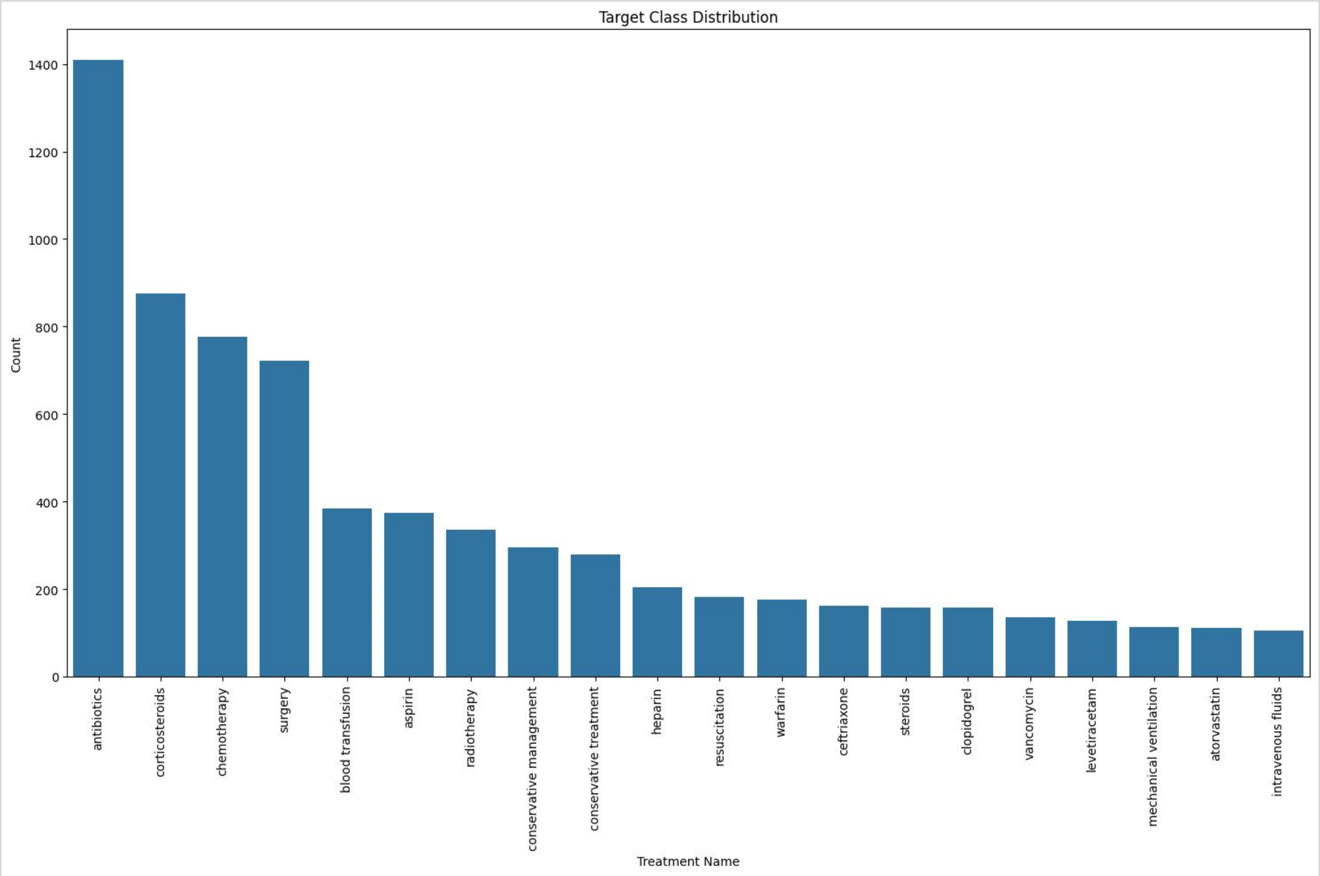
- Objective: Clean and prepare text data for modeling while retaining clinical terms
- Stop Words Removal: Eliminate common, non-informative words.
- Punctuation Removal: Remove punctuation marks.
- Lemmatization: Reduce words to their base form.
- Tokenization: Split text into tokens.
- Tool: Utilize **spacy** for preprocessing.

Text Vectorization

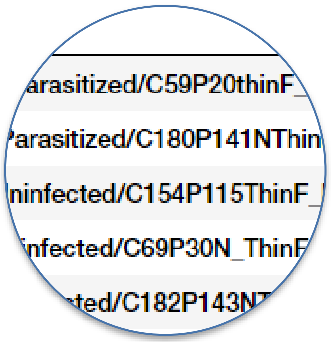
- Objective: Convert textual data into numerical vectors.
- Approach: Use **TFIDFVectorizer** for transformation.

06. PREDICTIVE MODEL DEVELOPMENT – Class Imbalance (1)

- After Feature Engineering the final **target variable (Treatment)** has 20 classes.
- **Note:** The data is highly imbalance.

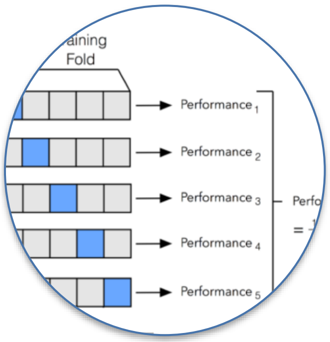


06. PREDICTIVE MODEL DEVELOPMENT (2)



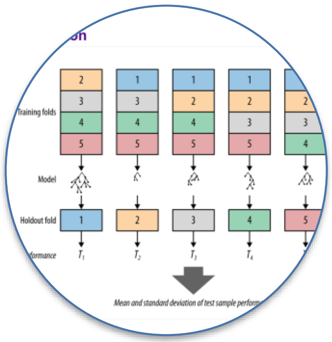
Data

- Shape : (7020, 10)
- X variables : 9
- Y variable : 1
- Y variable: "Treatment"
- No. of target classes: 20



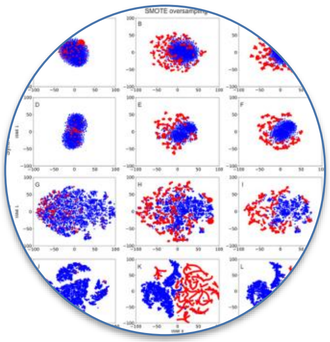
Train Test Split

- Train set: 20%
- Test set: 80%
- Method: Stratify
- X_train (5616, 9)
- X_test (1404, 9)



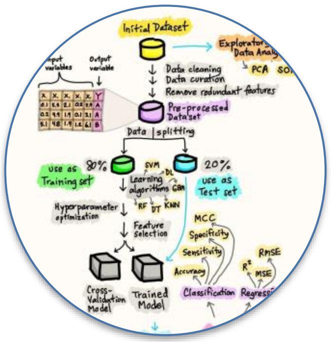
Cross Validation

- Method: Stratified K-fold



Oversampling technique

- Objective: Address class imbalance
- Method: SMOTE
- Scoring: F1_weighted



ML models

- KNN
- Logistic Regression
- SVM
- Random Forest

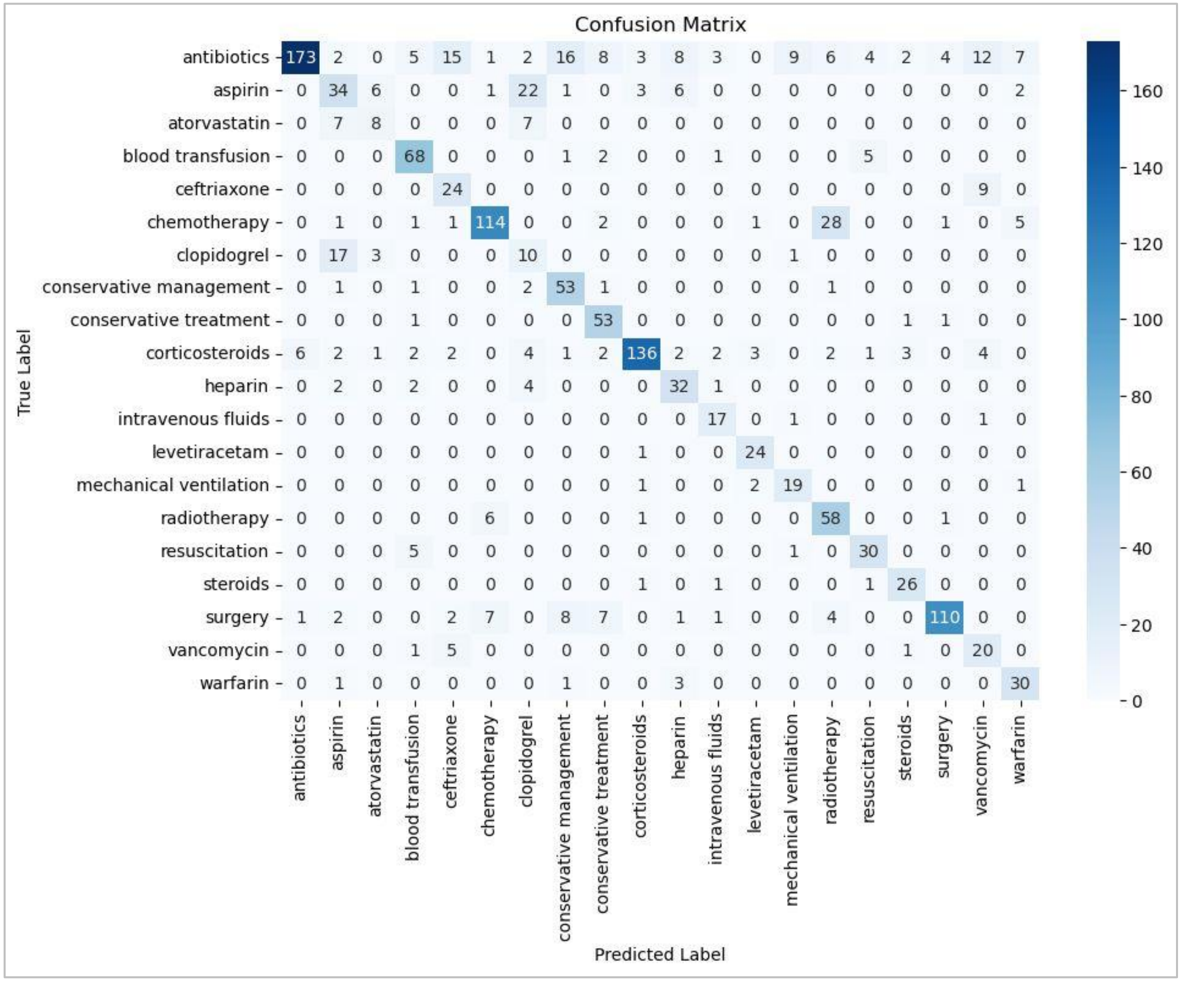
07. EVALUATION RESULTS (1)

Model	Parameters	Grid Search CV	Accuracy	Precision	Recall	F1 score (Weighted)
SVM	random_state=42	folds = 5	85%	0.85	0.85	0.85
Random Forest	Class_weight = 'balanced'	folds = 5	83%	0.83	0.83	0.83
Logistic Regression	max_iter=1000 random_state=42	folds = 5	82.7%	0.83	0.82	0.83
KNN	n_neighbors: 5	folds = 5	74%	0.78	0.74	0.74

07. EVALUATION RESULTS – Hyperparameter Tuning (2)

Model	Parameters (after hyperparameter tuning)	Grid Search CV	Accuracy	Precision	Recall	F1 score (Weighted)	AUC score
SVM	**C: 10** **class_weight: 'balanced'** **gamma: 'scale'** **kernel: 'linear'**	**folds = 5**	**86.6%**	**0.8597**	**0.8668**	**0.86**	**0.98**
Random Forest	class_weight = 'balanced' max_depth: None features: 'sqrt' samples_leaf: 1 samples_split: 2 n_estimators: 200	folds = 5	83.69%	0.7586	0.75	0.8373	0.9699

07. EVALUATION RESULTS – Confusion Matrix of Target variable (3)



08. CONCLUSION

- **Objective:** Predict treatment names based on patients' visit motivation, diagnosis condition, and associated clinical context.

- **Nature of Problem:** An NLP task involving predictive model building and feature engineering using TF-IDF vectorization.

- **Best Performing Model:** SVM with:

  - 86.6% accuracy
  - 0.98 AUC score

- **Reasons for SVM's Superior Performance:**

  - *High-Dimensional Feature Space:* The TF-IDF vectorizer creates a high-dimensional feature space. SVMs excel in such spaces, effectively separating different classes.
  - *Effective for Sparse Data:* TF-IDF transforms text data into a sparse format. SVMs handle sparse data well, making them ideal for text classification.
  - *Robust to Overfitting:* With proper regularization, SVMs are less likely to overfit, crucial for complex datasets with diverse textual descriptions and medical terms.
  - *Versatility with Kernels:* SVMs can use different kernel functions to handle non-linear relationships. For text data, a linear kernel often works well.

09. REFERENCES

1. Bonnet, A. (n.d.). AGBonnet/augmented-clinical-notes · datasets at hugging face. AGBonnet/augmented-clinical-notes · Datasets at Hugging Face. https://huggingface.co/datasets/AGBonnet/augmented-clinical-notes.
2. Zhao, Z., Jin, Q., Chen, F., Peng, T., & Yu, S. (2023, April 19). PMC-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems. arXiv.org. https://arxiv.org/abs/2202.13876.
3. Wang, J., Yao, Z., Yang, Z., Zhou, H., Li, R., Wang, X., Xu, Y., & Yu, H. (2024, June 28). NoteChat: A dataset of synthetic doctor-patient conversations conditioned on clinical notes. arXiv.org. https://arxiv.org/abs/2310.15959.
4. https://huggingface.co/datasets/AGBonnet/augmented-clinical-notes/blob/main/report.pdf.
5. Centers for Disease Control and Prevention. (n.d.). Health Insurance Portability and accountability act of 1996 (HIPAA). Centers for Disease Control and Prevention. https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html?CDC_AAref_Val=https%3A%2F%2Fwww.cdc.gov%2Fphlp%2Fpublications%2Ftopic%2Fhipaa.html

THANK YOU