Predicting Disease Risks through Historical Insights
Final Report

Kovida Mothukuri,
Lakshmi Bhavani Uppaluri

Semester: Fall 2023
Course: DSA5900 Professional Practicum
Credit Hours: 4

Faculty Supervisor: Dr. Beattie Matt J, Dr. Danala Gopichandh.
Company & Sponsor: Data Institute for Societal Challenges (DISC) – Dr. David Ebert

TABLE OF CONTENTS

Introduction:

Accurately predicting clinical complications ahead of time could be valuable in taking necessary interventions. The Federal Aviation Administration (FAA) uses several metrics to evaluate the pilot's preparedness for license renewal and combat compliance. One of the key factors in this decision-making is evaluating how medically fit a candidate is and how long should his/her license be renewed. Currently, there is no well-streamlined process or system to anticipate such risk metrics, which could be improved by developing robust predictive AI (Artificial Intelligence) enhanced tools. To address this issue, our primary objective in this project is to use a large repository of patients' medical records to predict the likelihood of developing certain disease conditions in the future. Such predictive models developed to identify risk metrics for several medical conditions could be valuable for the FAA to assist in their current license renewal protocols.

The dataset comprises patients' medical histories obtained via IBM Market Scan and purchased by DISC, OU. By extracting insights from historical medical records, we offer valuable insights not just to the FAA, but also to diverse industries like healthcare, government agencies, insurance, and financial services. The prediction of license renewal depends on identifying critical diseases, such as heart diseases, respiratory diseases, and nervous system disorders, listed in the FAA's incapacitating conditions. Selecting 2-3 significant diseases from this list, we will predict their occurrence in patients' future. This prediction aids in determining whether a pilot's license can be renewed. A key aspect of this process involves evaluating medical exams to gauge their potential for license renewal.

Through this project, we engage foundational data science principles encompassing statistics, machine learning, and deep learning. We aim to develop tools that yield insights into predicting the risk of developing specific medical conditions for license renewal potential, thereby offering valuable assistance to both the FAA and relevant authorities.

Objectives:

I. Technical Objectives:

Understanding Data and Previous Work:

1. *Database Familiarity*: Gain a comprehensive understanding of commercial claims and encounters (CCAE) database and several tables enlisted in it. Additionally, if possible, explore supplementary information from the Medicare (MDRC) database.
2. *Previous Work Analysis*: Investigate prior work to grasp the databases used, significant features, and the overall approach taken. Evaluate the strategy employed to address the problem.

Data Enrichment and Expansion:

3. *Feature Enhancement*: Assess the possibility of incorporating additional columns from various tables within the CCAE, moving beyond the previous emphasis solely on the Inpatient admission table.
4. *Extended Dataset*: Expand the dataset timeline by including data from 2017 to 2020. We will analyze patients with two consecutive years of records. Our dataset will include individuals without diseases in the first year and may or may not have developed in the following year. This strategy aims to enhance our dataset's robustness and model development. Select patients with records spanning two consecutive years, considering disease development patterns over time.
5. *Disease Profiling*: Re-evaluate features relevance across tables and pinpoint 2 to 3 critical disease codes as target variables, aligning with FAA recommendations.

Data Preprocessing and Model Development:

6. *Preprocessing Strategies*: Apply tailored preprocessing techniques, encompassing feature engineering and addressing class imbalance.
7. *Model Evaluation and Enhancement*: Evaluate existing models, fine-tune as needed, and introduce new models, while exploring diverse evaluation metrics.
8. *Robust Model Development:* Create refined machine learning models for accurate disease risk prediction, emphasizing assisting in pilot renewal process, leveraging prior insights and novel methodologies.

II. Individual Learning Objectives:

1. Gain familiarity working with remote database (PostgreSQL) on a server to understand, process, access via programming (python) and perform required analysis.
2. Comprehending the dataset, including the significance of each feature. Through this understanding, we aim to identify and prioritize the essential features that will play a pivotal role in the subsequent modeling process.

3. Enhance programming skills, focusing on data understanding, and proficiently implementing preprocessing steps. This includes data cleaning, feature extraction, standardization techniques, and implementing robust ML models and their evaluation procedures. Seek to acquire new methods or techniques to optimize results.
4. Gain hands-on experience by working with real-time data, understanding its challenges, and opportunities for improving model performance.

Data:

Ingestion:

The dataset comprises patients' medical histories obtained via IBM Truven Health Market Scan Research database and purchased by DISC, OU. Data captures person-specific clinical utilization, expenditures, and enrollment across inpatient, outpatient, prescription drug, and carve-out services. The data comes from a selection of large employers, health plans, and government and public organizations. The data is stored and managed in PostgreSQL using pgAdmin.

The challenges we encountered in obtaining the desired data included dealing with an extensive dataset, comprehending its vast size, grappling with medical terminology, and conversion of raw data into a meaningful and usable format. Our efforts involved thorough data cleaning and transformation to ensure data quality.

Preparation:

The IBM Market Scan comprises six databases spanning from 2017 to 2020:

1. Commercial Claims and Encounters Database
2. Medicare Supplemental
3. Health and Productivity Management Database
4. Benefit Plan Design Database
5. Medicaid Database
6. Market Scan Lab

Among these databases, we selected the CCAE Database because the other tables lack vital information and features essential to our project's objectives. Additionally, they do not contain relevant information useful for predicting future occurrences of specific diseases based on medical records and behaviors. Including these additional databases would only add complexity to the problem we are addressing.

The Commercial Claims and Encounters Database consists of seven tables:

1. Inpatient Admissions Table (CCAEI)
2. Facility Header Table (CCAEF)
3. Inpatient Services Table (CCAES)
4. Outpatient Services Table (CCAEO)
5. Outpatient Pharmaceutical Claims Table (CCAED)
6. Annual Enrollment Summary Table (CCAEA)
7. Enrollment Detail Table (CCAET)

We have successfully uploaded all the tables data that we have onto the PostgreSQL database, we have decided to work with two distinct sets of data: one pertaining to the years 2017 and 2018, and the other covering 2019 and 2020.

Figure1 presented below depicts the data preparation process, a crucial phase in our analytical endeavors. This visual representation delineates the sequential steps and methodologies employed for the transformation of raw data into a refined and structured format, rendering it primed for thorough analysis. This meticulous process ensures that our data attains a state of cleanliness, consistency, and suitability for the ensuring phases of our research. The figure

serves as a visual guide, elucidating the critical stages and the logical progression from unprocessed data to a well-structured dataset.

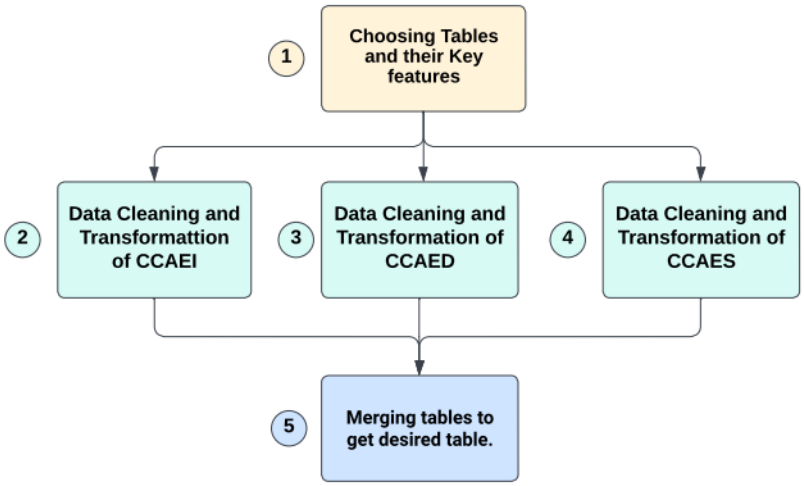


Figure 1: Data Preparation Workflow

1. Choosing Tables and their Key Features:

From the CCAE, our analysis focused on a thorough examination of the key data columns spanning all available tables. Through this meticulous review, it became apparent that the tables pertaining to inpatient admissions, inpatient services, and outpatient pharmaceutical drugs hold paramount relevance to the primary objectives of our project. Following a comprehensive assessment, we have identified a total of 23 pertinent features within the Inpatient Admissions table, 4 significant features within the Outpatient Pharmaceutical Claims table, and an additional 4 critical features within the Inpatient Services table. These discerned features lay the foundation for our subsequent data analysis, contributing to the precision and effectiveness of our research efforts. A more detailed explanation is shown as a flowchart in Figure2.

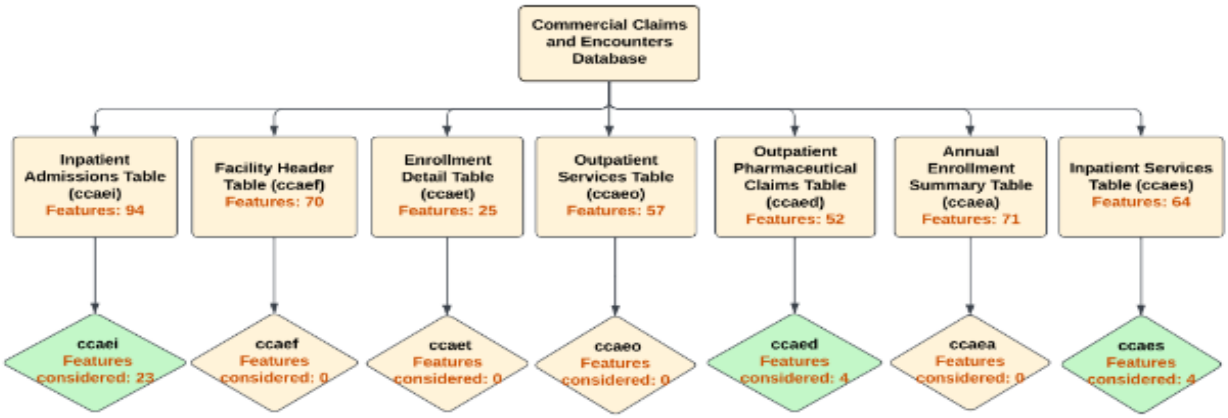


Figure 2: Choosing Tables and their Key Features

2. Data Cleaning and Transformation of CCAEI table:

In the context of the inpatient admissions table, we initiated the process by generating new tables "ccaei_#" derived from the existing ones. The table "ccaei_#" represents data for the years 2017 to 2020, where the symbol "#" is replaced with the corresponding year for each entry. These tables were crafted by extracting essential features such as enrolid, year, age, sex, days, dxver, mdc, agegrp, and an array containing diagnostic codes (dx1, dx2, dx3, dx4, dx5, dx6, dx7, dx8, dx9, dx10, dx11, dx12, dx13, dx14, dx15) designated as dxcodes.

We subsequently removed null values from the following columns: enrolid, sex, age, and dxver (Diagnosis Version). Additionally, we excluded data associated with agegrp equal to 1, as individuals in this age group are ineligible to be pilots. Data corresponding to dxver equal to 9, which is related to ICD-9 codes (since we use ICD-10 library codes), was also eliminated. Furthermore, any null values within the dxcodes array were removed. In the interest of data consistency, we decided to exclude the dxver column due to the presence of data where dxver=0, signifying an association with ICD10.

We introduced a new column called mdc_flag for the categorical feature mdc (Major Diagnosis Category). To focus on mdc relevant to our objective, we identified specific categories of mdc 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 16, 17, and 27. In this column, we assigned a value of 1 when the mdc value corresponds to a major diagnosis relevant to our objective and others as 0. Additionally, we created another column named non_mdc_flag, which contains the opposite values of the mdc_flag column, indicating cases where the mdc value does not align with our objective. A summary of the flow process applied for the CCAEI table transformation is summarized in Figure 3.

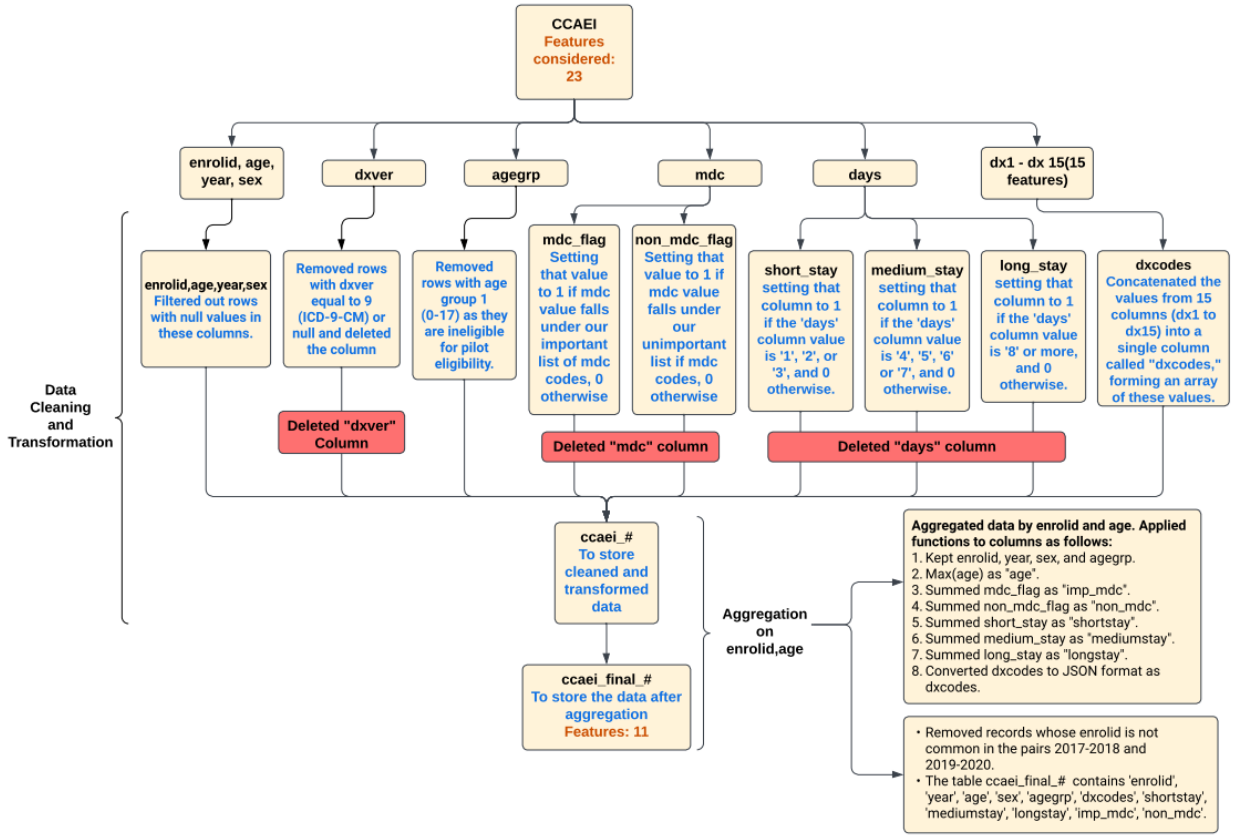


Figure 3: Data Cleaning and transformation of CCAEI table

Next, we performed a transformation on the days column, generating three new columns: short_stay, medium_stay, and long_stay. In this transformation, short_stay was assigned a value of 1 when the days column held values of 1, 2, or 3, signifying a brief hospital stay. medium_stay was set to 1 when days contained values of 4, 5, 6, or 7, indicating a moderate hospital stay. long_stay was assigned a value of 1 when days exceeded 7, representing an extended hospital stay.

Subsequently, we generated a new table "ccaei_final_#" by aggregating data based on enrolid values. Additionally, we filtered and selected the enrolids common to both the 2017 and 2018 dataset, as well as for the 2019 and 2020 dataset. The resulting data in the new tables adheres to the following format: enrolid, year, age, sex, agegrp, merged_dxcodes, tot_shortstay, tot_mediumstay, tot_longstay, tot_mdc_imp, and tot_non_mdc. During the aggregation process, we summed the values within the columns for dxcodes, short_stay, medium_stay, long_stay, mdc_flag, and non_mdc_flag.

3. **Data Cleaning and Transformation of CCAED table:**

We started with the table of outpatient pharmaceutical claims, which included a column called 'thergrp' (Therapeutic Group). This column categorizes the drugs taken by patients during their hospital visits. We created a new table "ccaed_#" by extracting the columns enrolid, year, age, and thergrp. We then removed any rows with null values in the enrolid, year, and age columns to ensure data quality. To focus on drugs relevant to our objective, we identified specific therapeutic groups of categories 1, 5, 6, 7, 8, 14, 15, 16, and 27. We filtered the thergrp column to retain only the values corresponding to the identified set of therapeutic groups, discarding others. In Figure 4, we present a concise summary of the procedural flow applied to transform the CCAED table.

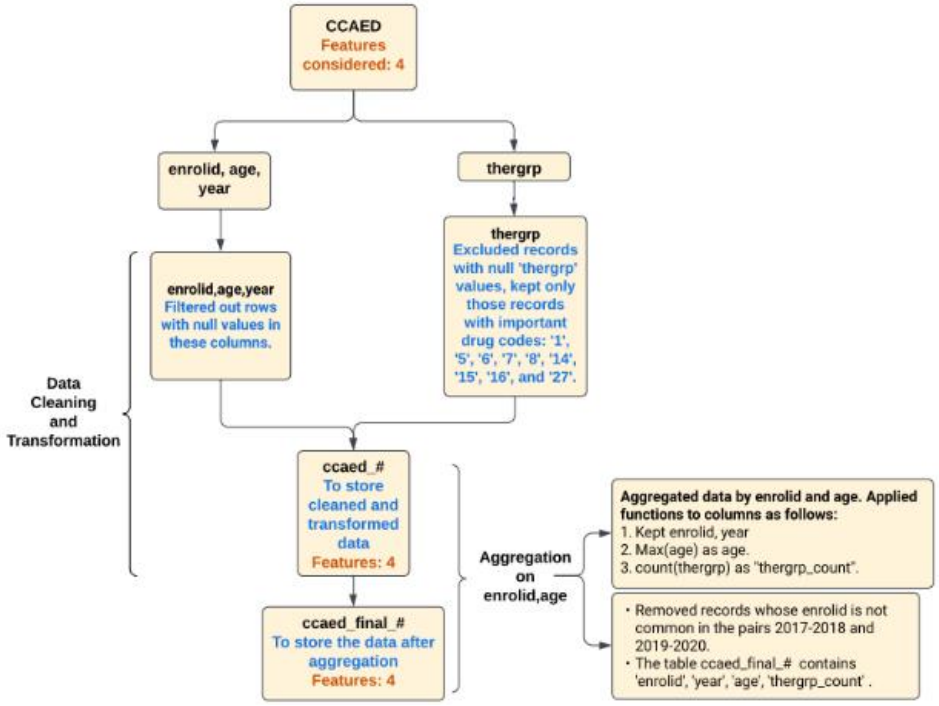


Figure 4: Data Cleaning and transformation of CCAED table

After this filtering process, we created a new table called "ccaed_final_#" containing enrolid, year, age, and a new feature called thergrp_count. We performed an aggregation based on enrolid, selecting enrolids that were common in both outpatient and inpatient admissions. The thergrp_count feature was generated by counting the occurrences of therapeutic group values for each enrolid during the aggregation, providing a measure of drug categories used by patients on their hospital visits.

4. Data Cleaning and Transformation of CCAES table:

Within the inpatient services table, we identified a critical feature labeled "svcscat" (Service Sub-Category Code). This code categorizes the type of service provided comprehensively. We initiated by creating a new table "ccaes_#" and extracting the columns enrolid, year, age, and svcscat. To ensure data quality, we removed any rows containing null values in the enrolid, year, and age columns. We introduced two new columns: 'emer_flag' and 'non_emer_flag'. The emer_flag column takes a value of 1 when the code in svcscat ends with 20, indicating an emergency-related service, otherwise, it is set to 0. Similarly, the non_emer_flag is assigned a value of 1 if the svcscat code does not end with 20, signifying a non-emergency service, otherwise, it is set to 0.

After refining the data, we performed an aggregation based on enrolids. Additionally, we filtered the enrolids to include only those that were also found in the inpatient admissions table. Subsequently, we constructed a new table "ccaes_final_#". In this table, we retained the columns enrolid, year, and age. Additionally, we introduced two new informative features: 'emer_visits' and 'non_emer_visits'. The emer_visits feature was calculated by summing the counts of emer_flag values for each enrolid. Similarly, the non_emer_visits feature was generated by aggregating the total counts of non_emer_flag values for each enrolid. The flow process employed for the transformation of the CCAES table is depicted in Figure 5, as summarized below.

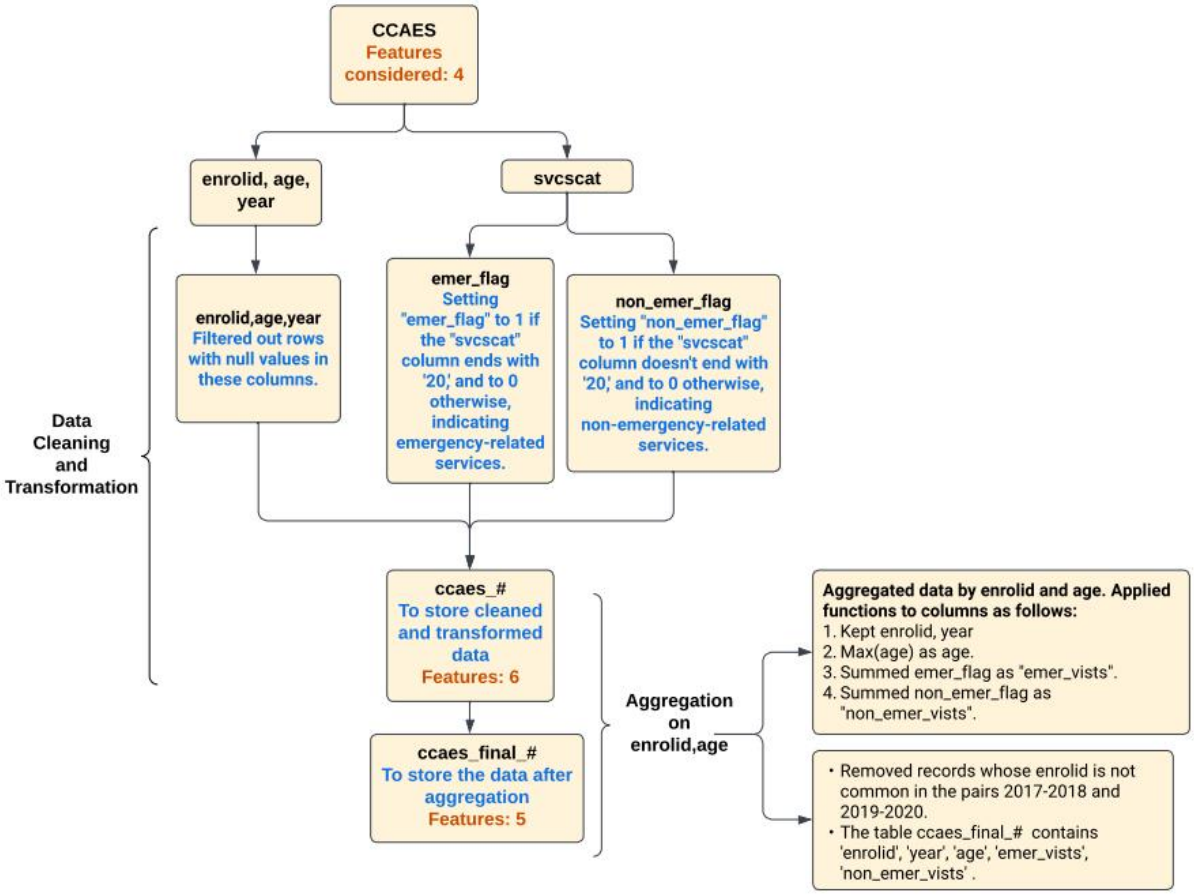


Figure 5: Data Cleaning and Transformation of CCAES table

5. Merging tables to get desired table:

To initiate our modeling process, we undertook the creation of a final data table, by combining the information from three primary data sources: inpatient admissions, inpatient services, and outpatient pharmaceutical claims. The objective was to construct a comprehensive dataset for our analysis.

1. *Merging Final Inpatient Admissions and Final Inpatient Services Data:*

We commenced by consolidating data from two key sources: inpatient admissions, which encompassed 11 distinct features, and inpatient services, containing 4 features. This consolidation resulted in the formation of a new table referred to as 'merged_in_#'. We accomplished this merger through an inner join operation, based on the common 'enrolid' column, which facilitated the synchronization of relevant information across the datasets. Furthermore, we pruned the redundant features 'year' and 'age' to enhance the dataset's conciseness.

2. *Combining the 'merged_in_#' Data with Final Outpatient Pharmaceutical Claims Data:*

Building upon the 'merged_in_#' table, we integrated data from the outpatient pharmaceutical claims, which brought an additional 5 features to our dataset. This amalgamation resulted in the creation of a composite table 'merged_inout_#'. The 'merged_inout_#' table presented a comprehensive view, comprising 14 features in total. This encompassed essential information, including 'enrolid,' 'year,' 'age,' 'sex,' 'agegrp,' 'dxcodes,' 'shortstay,' 'medium stay,' 'longstay,' 'imp_mdc,' 'non_mdc,' 'emer_visits,' 'non_emer_visits,' and 'thergrp_count.'

The establishment of the 'merged_inout_#' table represents a significant milestone in our analytical journey. This table now encompasses a diverse and comprehensive dataset, setting the stage for the forthcoming phases of modeling and data analysis. The integration of this dataset holds the potential to yield valuable insights and serve as a strong foundation for achieving our overarching research objectives. A more comprehensive description of the process is visually presented in Figure 6 through a flowchart.

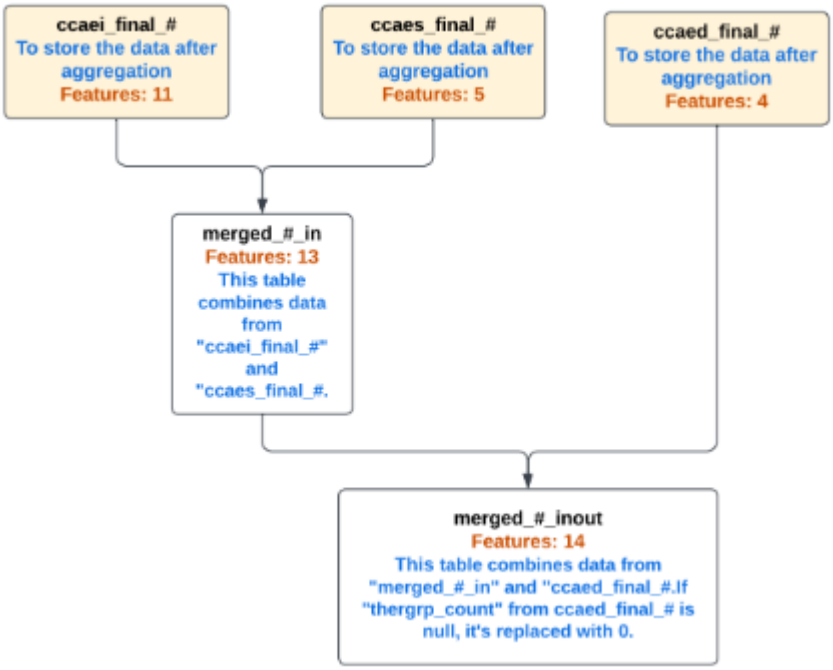


Figure 6: Merging tables to get desired table.

Following the creation of the 'merged_inout_#' table, the subsequent step in our data processing involved the generation of new features derived from the 'dxcodes' column. The original diagnosis codes within this column were recorded in the ICD-10-CM format, which presented a formidable challenge for model training. This challenge stemmed from the sheer volume of potential code combinations within the ICD-10-CM classification system.

To effectively address this complexity, we undertook a transformative approach to these features. We systematically tallied the occurrences of diagnosis codes and categorized them into their respective disease categories. This process served to simplify the data and provide a more manageable foundation for subsequent analysis and model training.

We imported data from PostgreSQL into Google Colab using pgAdmin. Afterward, we loaded this data into a Python Data Frame. Below is the Figure 7 which depicts of the appearance of the data in 'merged_inout_#':

	enrolid	year	age	sex	agegrp	dxcodes	shortstay	mediumstay	longstay	imp_mdc	non_mdc	emer_visits	non_emer_visits	thergrp_count
0	1005412302	2017	47	2	4	[[C9200, D6481, D65, I959, K521, R079, C9500, C9590, I498, I348, C9250, R9431, D709, I4581, C9201], [Z5111, C9200, E063, F419, M3500, C9201]]	0	1	1	2	0	2	128	4
1	1005478506	2017	64	1	5	[[J189, E871, E872, I509, J9601, J181, R918, R0602, E785, I10, J9801]]	1	0	0	1	0	6	45	0
2	1005479606	2017	57	2	5	[[M170, D62, I10, M069, Z87891, M179, Z471, G8918, Z96652], [A4151, D638, G9341, N10, R6520, J90, M069, R1033, R1084, R51, G9340, J189, M25551, M25552, R112], [M1612, D62, I10, M069, M810, M169, Z471, K219, Z96642], [M1711, I10, M069, Z79899, Z87891, M179, Z471, G8918, Z96651]]	3	1	0	3	1	4	90	0
3	1005485701	2017	46	2	4	[[N1330, G40909, N859, N200, N920, Z01818, N924, R109, R52, N8320]]	1	0	0	1	0	5	26	2
4	1005495501	2017	62	2	5	[[I471, E039, E785, G4733, I480, Z7901, Z8249, I499, R0902, R079, E0590, Z98890]]	1	0	0	1	0	7	10	2

Figure 7: Sample data in 'merged_inout_#' table

The ICD-10-CM codes were effectively organized into 22 major disease categories, each representing a distinct medical condition. This categorization not only streamlined the data but also facilitated a more intuitive and interpretable framework for our research and analysis. As a result, new tables labeled 'transformed_#' were established, wherein the new features represented the counts of diseases in their respective categories.
The data in the table 'transformed_#' looks as below in the Figure 8:

	enrolid	age	sex	agegrp	shortstay	mediumstay	longstay	imp_mdc	non_mdc	emer_visits	non_emer_visits
0	1005412302	47	2	4	0	1	1	2	0	2	128
1	1005478506	64	1	5	1	0	0	1	0	6	45
2	1005479606	57	2	5	3	1	0	3	1	4	90
3	1005485701	46	2	4	1	0	0	1	0	5	26
4	1005495501	62	2	5	1	0	0	1	0	7	10

thergrp_count	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX	XXI	XXII
4	0	0	3	1	1	0	0	0	4	0	1	0	1	0	0	0	0	2	0	0	1	0
0	0	0	0	3	0	0	0	0	2	4	0	0	0	0	0	0	0	2	0	0	0	0
0	1	0	3	0	0	4	0	0	3	2	1	0	13	1	0	0	0	5	0	0	9	0
2	0	0	0	0	0	1	0	0	0	0	0	0	0	6	0	0	0	2	0	0	1	0
2	0	0	0	3	0	1	0	0	3	0	0	0	0	0	0	0	0	2	0	0	3	0

Figure 8: Sample data in 'transformed_#' table.

The roman numbers in the columns represent each disease category or chapter as described in a python ICD-10-CM package here: https://pypi.org/project/icd10-cm/ .

After transformation of dxcodes to chapter codes, we can see the 22-chapter codes related to each major diagnosis category. Within each of the 22 major disease categories, the ICD-10-CM coding system employs distinct values, which typically begin with 0, 1, 2, and so forth, to denote the frequency of specific disease-related diagnoses. The distribution of frequency counts in each chapter can be seen below in Figure 8.

Focusing on each chapter, our objective was to determine whether an individual "Has a disease" or "Does not Have a disease." This essential differentiation was effectively translated into binary values: 1 and 0. In this conversion, a new column, such as target_IX, was generated. It holds values of 1 if the original column IX (Circulatory Disease) is greater than zero, signifying "has a disease," and 0 otherwise, indicating "Does not have a disease." Similar binary columns target_X, target_VI were created for other categories such as X (Respiratory System) and VI (Nervous System).

After the transformation of the diagnostic chapter codes into binary values of target columns, the data of chapter codes looks as mentioned in figure 9:

	enrolid	IX	X	VI
0	3685447502	0	0	0
1	29761716502	0	0	0
2	3369570801	3	0	2
3	2181667301	0	0	1
4	3127372701	5	1	2

	enrolid	target_IX	target_X	target_VI
0	3685447502	0	0	0
1	29761716502	0	0	0
2	3369570801	1	0	1
3	2181667301	0	0	1
4	3127372701	1	1	1

Figure 9: The sample data of chapter codes after binary transformation

This transformation is specifically done for the data of target years 2018 and 2020, which plays a significant role in predicting a specific disease risk.

Data Exploration:

We have embarked on our data analysis journey with a focus on two consecutive pairs of years: 2017-2018 and 2019-2020. Our primary objective was to explore the data for the first pair of years, specifically 2017 and 2018. After completing our analysis for this year's range, we plan to shift our attention to the data for 2019 and 2020.

S.No	Column Name	Brief description
1	enrolid	A unique three-to-eleven-digit number identifying each enrollee.
2	year	The calendar year during which the service was rendered, the admission began, or the population was eligible
3	age	Patient age in years at the time of service
4	sex	Gender of the patient
5	agegrp	A value identifying the patient or members age group. 1: 0-17, 2: 18-34, 3: 35-44, 4: 45-54, 5: 55-64, 6: 65 and older
6	dxcodes	An array containing all the diagnosis codes for which the enrollee has undergone procedures.
7	shortstay	It represents the count of short hospital stays, indicating the number of times a particular enrollee has been admitted to the hospital for 1 to 3 days.
8	mediumstay	It represents the count of medium hospital stays, indicating the number of times a particular enrollee has been admitted to the hospital for 4 to 7 days.
9	longstay	It represents the count of long hospital stays, indicating the number of times a particular enrollee has been admitted to the hospital for a duration of more than 7 days.
10	imp_mdc	It is the frequency of times an enrollee has received a diagnosis falling within the important major diagnostic category list.
11	non_mdc	It is the frequency of times an enrollee has received a diagnosis falling within the non-important major diagnostic category list.
12	emer_visits	It is the number of times enrollees visited emergency rooms.
13	non_emer_visits	It is the number of times enrollees have not visited emergency rooms.
14	thergrp_count	It represents the count of instances when an enrollee used a drug from a list of significant therapeutic group codes.

Table 1: The above table provides a list of column names along with their respective descriptions for the dataset.

Here is a glimpse of some of our initial data exploration findings for 2017 and 2018 data.

- ***Distribution of Age and Sex:***

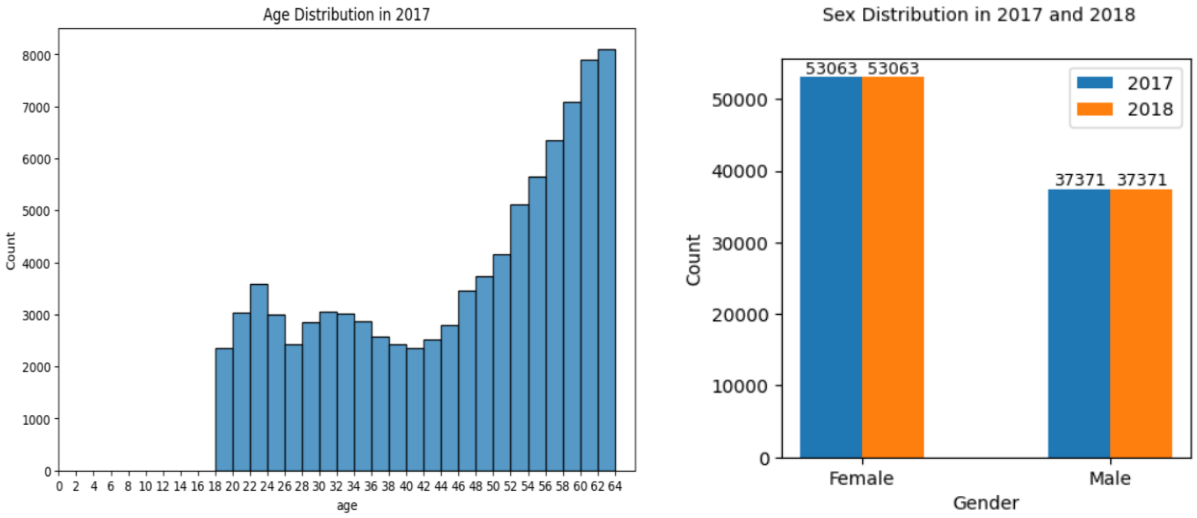


Figure 10: Age and Sex distribution in 2017 and 2018

The left graph in figure 10 reveals the absence of records for individuals aged 0 to 17 years, as they typically do not qualify for disease prediction or pilot eligibility. In contrast, the highest record count falls in the 62 to 64-year age range. From the graph on the right, it is evident that in both 2017 and 2018, there are more records for female patients compared to male patients. Since we have already considered the common enrolids between 2017 and 2018, the age distribution graph for 2018 shifts one position to the right, while the sex distribution for both 2017 and 2018 remains unchanged.

- ***Distribution of Age group ('agegrp') by Sex:***

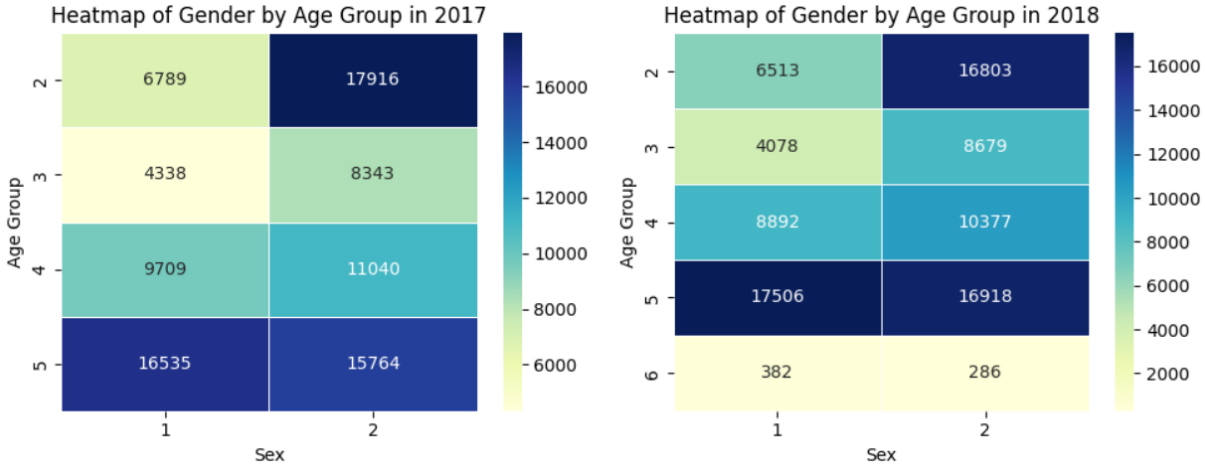


Figure 11: Distribution of Age Group by Sex in 2017 and 2018

In figure 11, on the x-axis, 1 and 2 correspond to Male and Female, respectively. On the y-axis, 2, 3, 4, 5, and 6 represent the age groups 18-34, 35-44, 45-54, 55-64, and 65 and older, respectively. The heatmaps provide interesting insights. In 2017, we observe the highest number of female members in the age group 18-34, while the lowest count is for males in the 35-44 age group. In 2018, the highest count is for males in the 45-54 age group, and the lowest count is for females in age groups older than 65.

- ***Distribution of short stay ('shortstay'), medium stay ('mediumstay'), long stay('longstay') by Age group:***

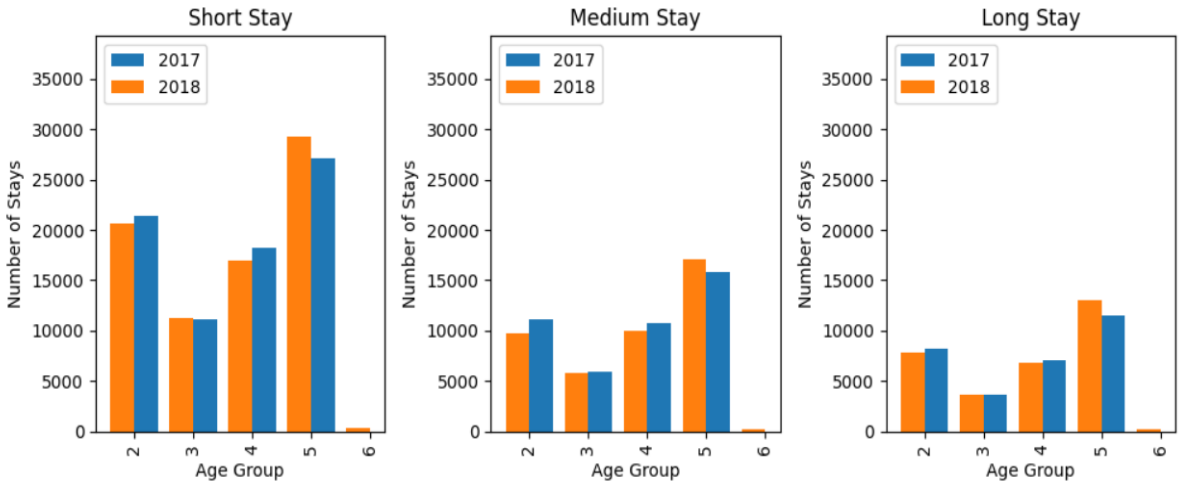


Figure 12: Distribution of short stay, medium stay, long stay by Age group in 2017 and 2018

In figure 12, on the x-axis, 2, 3, 4, 5, and 6 represent the age groups 18-34, 35-44, 45-54, 55-64, and 65 and older, respectively. The graph above reveals several insights. In both 2017 and 2018, the age group with the highest number of short, medium, and long stays is age group 5 (55-64). Notably, 2018 has a significantly higher count of

stays compared to 2017. Conversely, the age group with the lowest number of stays, whether short, medium, or long, is age group 3 (45-54), and this trend remains almost consistent in both years.

What is particularly interesting is that individuals in age group 2 (18-34) have a higher hospital stay count compared to those in age group 3 (35-44) and age group 4 (45-54). This observation is significant because, as indicated by the heat map of distribution of age group by sex in Figure 3, a substantial portion of the individuals in age group 2 are female. Many of them have sought hospitalization for reasons related to pregnancy or childbirth.

- ***Distribution of short stay ('shortstay'), medium stay ('mediumstay'), long stay ('longstay') by Sex:***

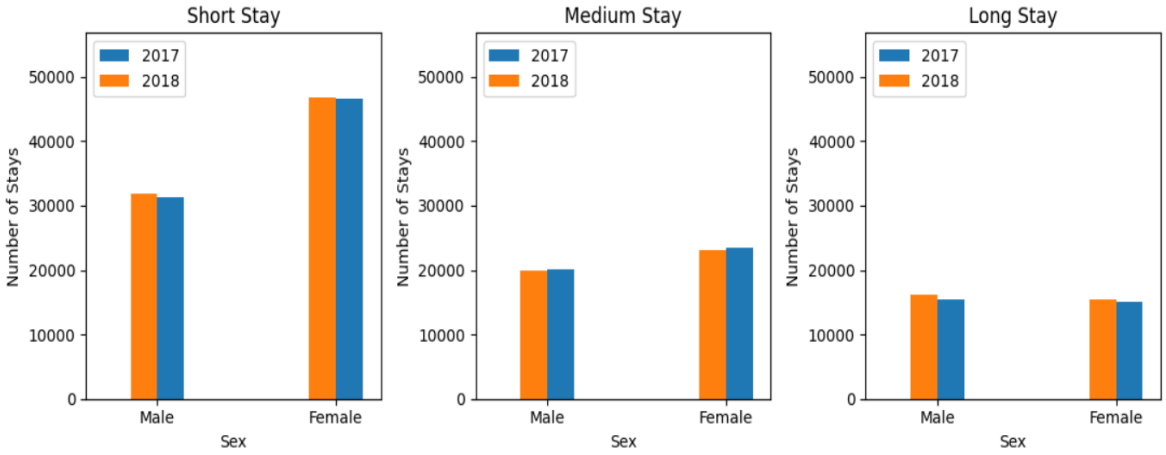


Figure 13: Distribution of short stay, medium stay, long stay in 2017 and 2018

The graph in figure 13 above indicates that, in both 2017 and 2018, female patients have had more short and medium stays compared to male patients, with a similar pattern across both years. However, for long stays, male patients have a higher frequency of hospitalization, and this difference is particularly pronounced in 2018. This is due to the higher number of females, often admitted for pregnancy or childbirth, which typically involves shorter stays.

- ***Distribution of Emergency Visits ('emer_visits') by Age Group and Sex:***

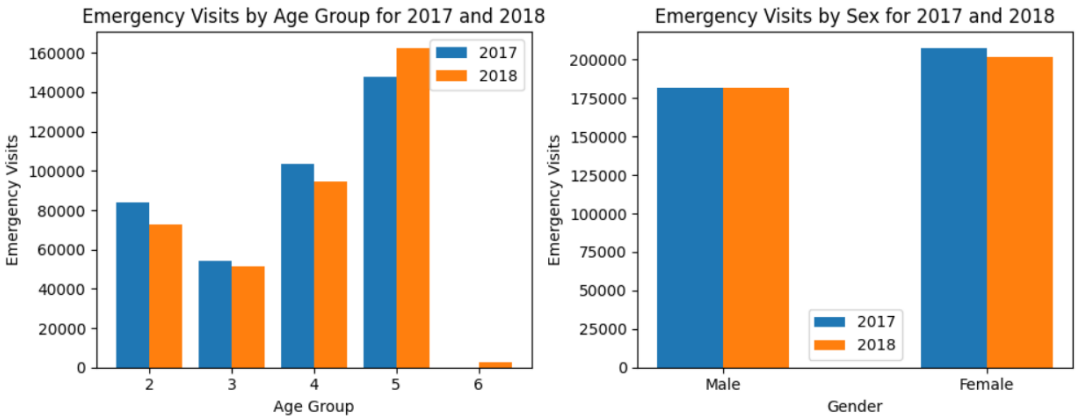


Figure 14: Distribution of Emergency visits by Age Group and Sex in 2017 and 2018

The left graph in figure 14 depicts the distribution of emergency room visits across different age groups, where age groups 2, 3, 4, 5, and 6 correspond to 18-34, 35-44, 45-54, 55-64, and 65 and older, respectively. It is evident that individuals in the age group 55-64 exhibited the highest frequency of visits to the emergency room, surpassing all other age groups. In contrast, the age group 35-44 recorded the lowest number of emergency room visits.

On the right, the second graph provides a clear distinction: female patients have a notably higher rate of emergency room visits compared to their male counterparts.

- ***Distribution of Therapeutic Group ('thergrp_count') by Age Group and Sex:***

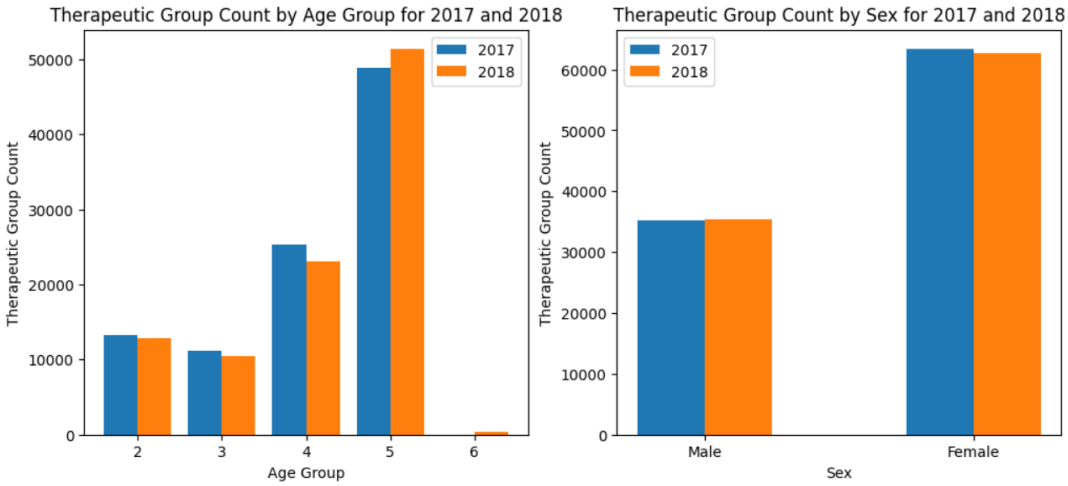


Figure 15: Distribution of Therapeutic Group by Age Group and Sex in 2017 and 2018

The left graph in figure 15 illustrates the distribution of emergency room visits across various age groups. Age groups 2, 3, 4, 5, and 6 represent 18-34, 35-44, 45-54, 55-64, and 65 and older, respectively. Notably, individuals in age group 5 (55-64) have a higher usage of important drugs compared to other age groups. On the right, the graph reveals that female patients have a higher consumption of important drugs compared to their male counterparts.

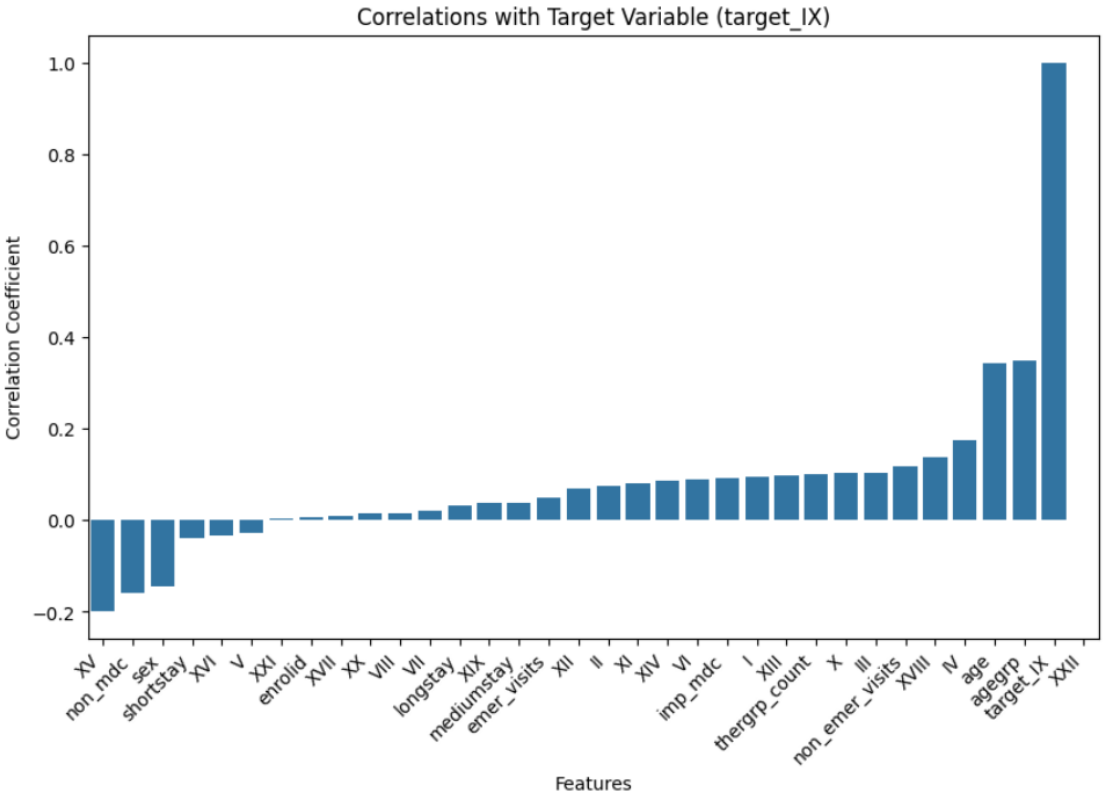
Methodology:

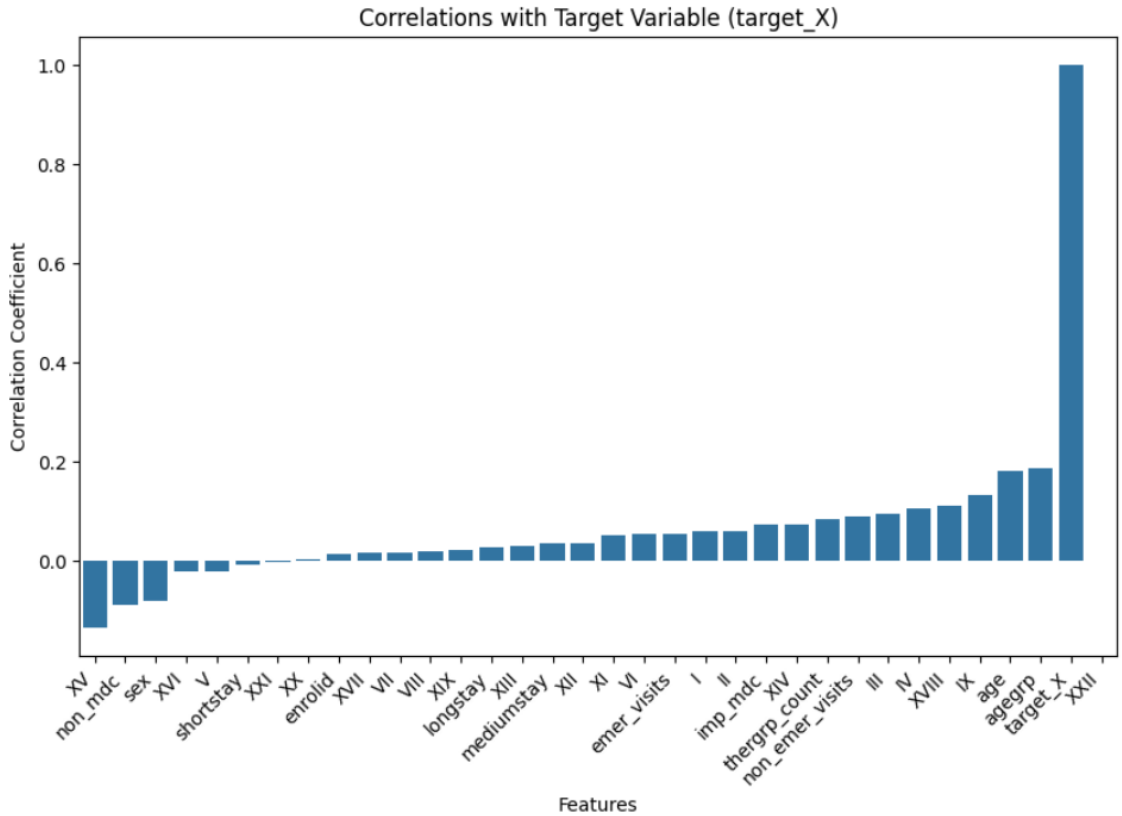
Techniques:

The FAA has furnished a comprehensive list of diseases that are of paramount importance when scrutinizing the medical data of pilots. The specific details of this list can be found in the Appendix section of our report. We undertook a meticulous search within the ICD-10 library to identify and categorize these medical conditions effectively. Our search efforts yielded the relevant chapter codes in the ICD-10 classification system associated with each listed disease.

The FAA has organized these medical conditions into distinct chapters within the ICD-10-CM coding system. The FAA classifies these medical conditions into distinct chapters, including Chapter IX for Diseases of the Circulatory System, Chapter X for Diseases of the Respiratory System, and Chapter VI for the nervous system. For our modeling, we have designated these chapters as our target columns. This strategic choice aligns with our objective to develop a robust and effective model that can evaluate and predict the presence or absence of these specific medical conditions.

The correlation between all the features w.r.correspomding target variables is depicted in Figure 16.


Correlations with Target Variable (target_IX)
Correlation Coefficient
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
XV
non_mdc
sex
shortstay
XVI
V
XXI
enrolid
XVII
XX
VIII
VII
longstay
XIX
mediumstay
emer_visits
XII
II
XI
XIV
VI
imp_mdc
I
XIII
thergrp_count
X
III
non_emer_visits
XVIII
IV
age
agegrp
target_IX
XXII
Features


Correlations with Target Variable (target_X)
Correlation Coefficient
1.0
0.8
0.6
0.4
0.2
0.0
XV
non_mdc
sex
XVI
V
shortstay
XXI
XX
enrolid
XVII
VII
VIII
XIX
longstay
XIII
mediumstay
XII
XI
VI
emer_visits
I
II
imp_mdc
XIV
thergrp_count
non_emer_visits
III
IV
XVIII
IX
age
agegrp
target_X
XXII
Features

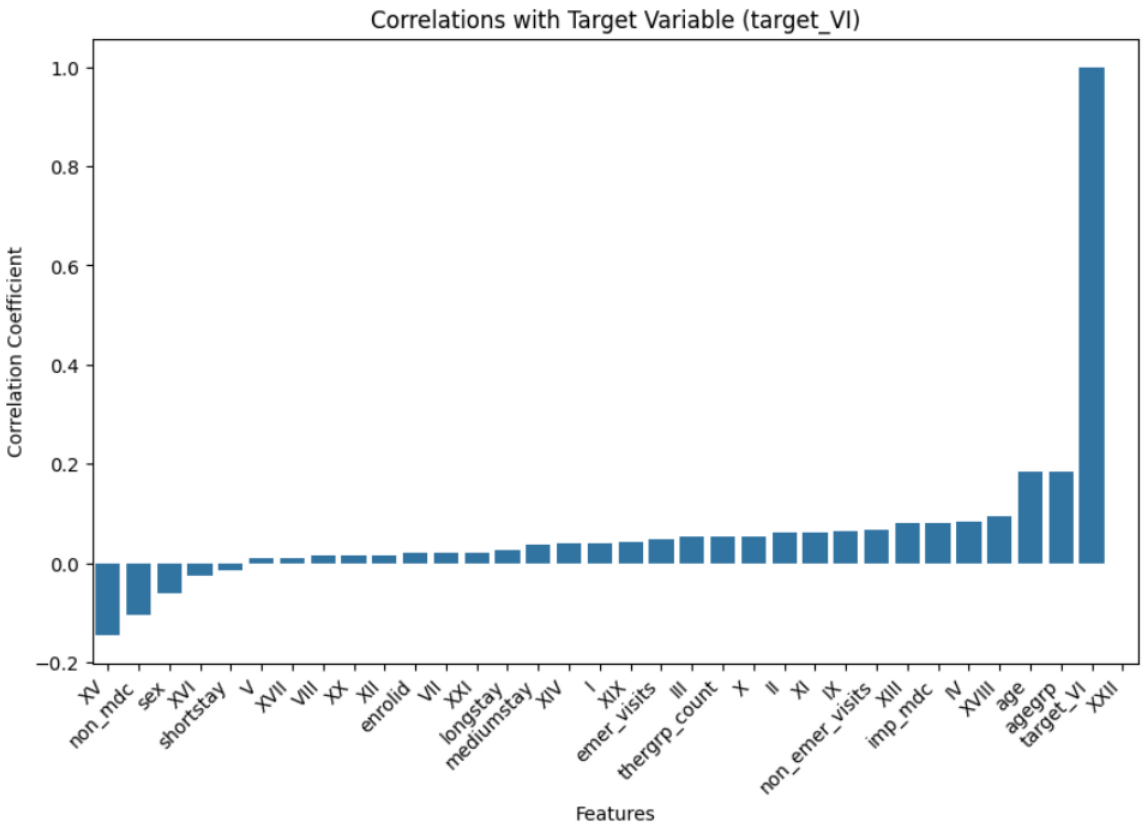


Figure 16: Correlation between all the features w.r.t target variables

Our primary research objective is to predict the likelihood of individuals developing specific medical conditions outlined by the FAA within the next year, assuming they are disease-free in the current year. To achieve this goal, we are conducting a contemporary simulation using data from the year 2017, treating 2018 as our point of reference for the future. In addition to the data from 2017 and 2018, we have extended the dataset by incorporating information from 2019 into the 2017 dataset and integrating 2020 data into the 2018 dataset. This approach allows us to scrutinize potential trends and make predictions based on the available data, which is depicted in Figure 17.

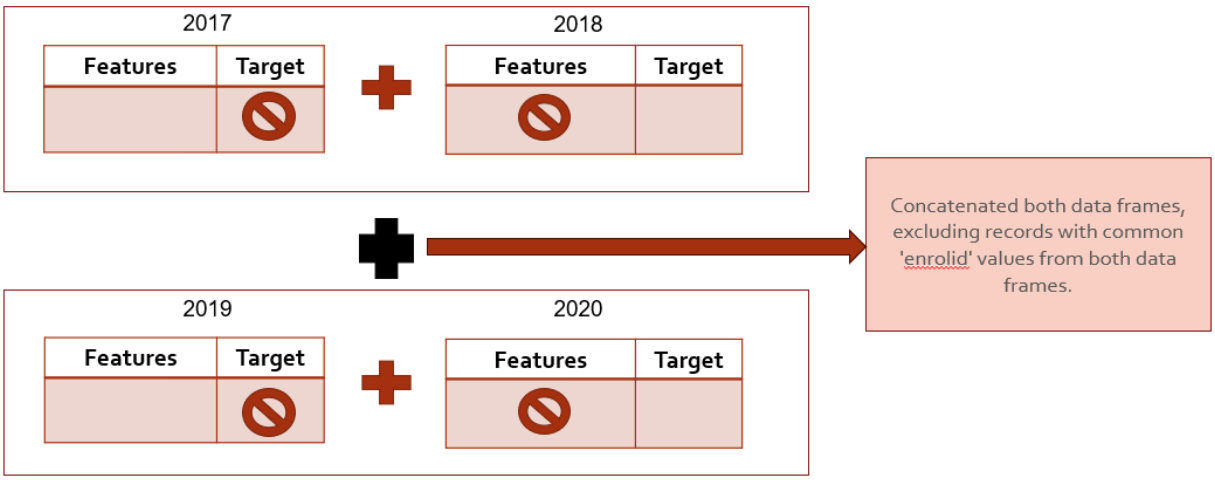


Figure 17: Contemporary simulation of Data Approach

To enrich our analysis, we refined the dataset by aligning features from 2017 with corresponding labels from 2018, providing a holistic perspective. Similarly, we conducted a simulation for the years 2019 and 2020, aligning their respective features and labels. Following this, we merged the simulated data from 2017-2018 and 2019-2020 into a consolidated data frame, excluding records with common enrolid values from both datasets. This process ensures that our dataset comprises unique enrolid values across the entire dataset. By adopting this methodology, we have crafted a more refined dataset, contributing to a deeper and more insightful analysis.
After converting the target variables into binary class labels (0 for absence, 1 for presence), we observed an imbalance in our dataset as shown in the Figure 18. This imbalance pertains to the distribution of various chapters, specifically Chapters IX, X, and VI.

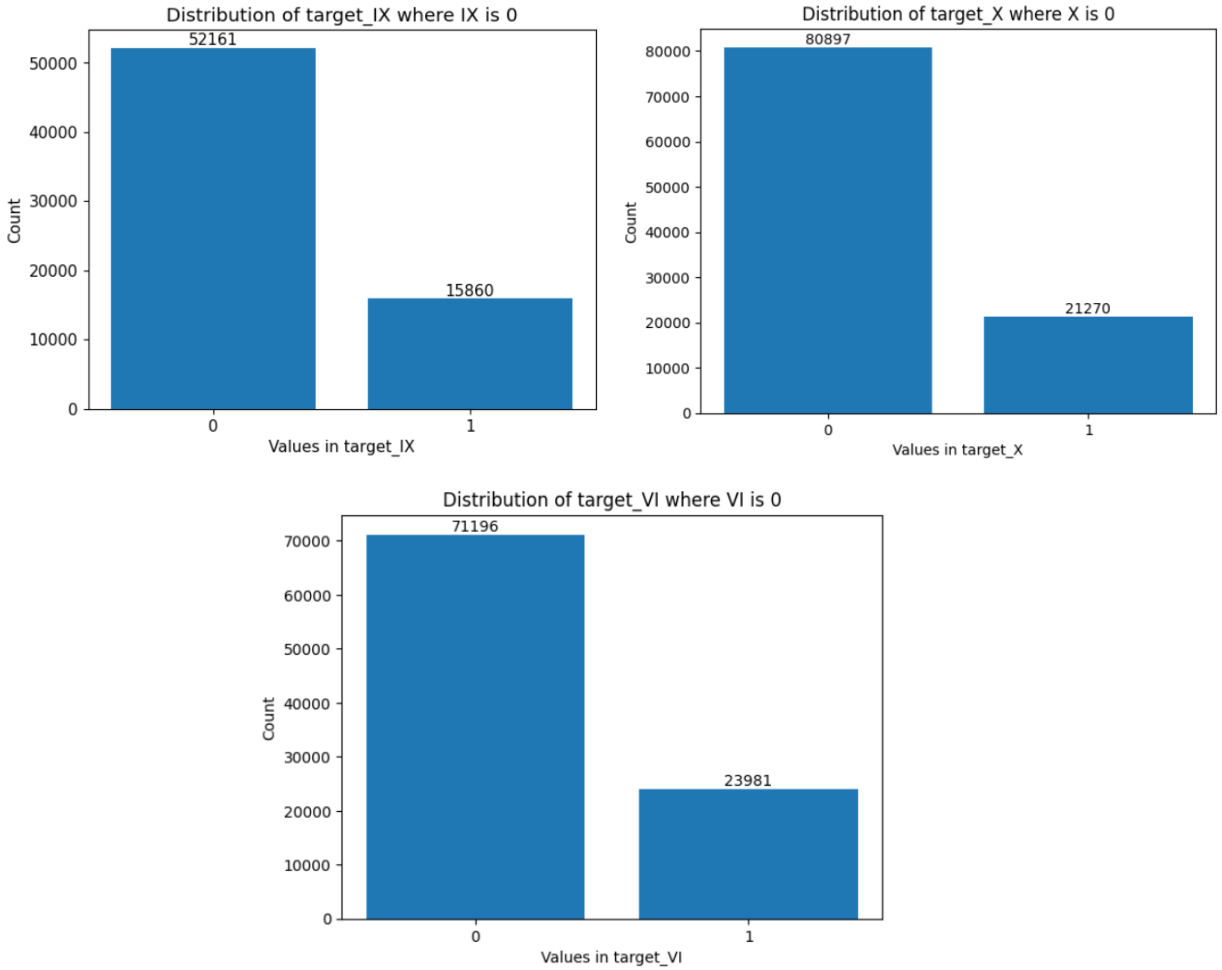


Figure 18: Distribution of data in target columns

After plotting the class distribution, it becomes evident that a class imbalance issue exists, as depicted in the figure. There is a noticeable disparity between the number of instances in the positive class and the number of instances in the negative class. This imbalance poses a significant challenge for modeling and predictive tasks.

Addressing Class Imbalance:
From figure 19, To tackle the issue of class imbalance, we implemented two essential techniques: Synthetic Minority Over-sampling Technique (SMOTE) and RandomUnderSampler. SMOTE plays a pivotal role in addressing class imbalance by generating synthetic instances of the minority class, ensuring a more balanced dataset. On the other hand, Random Under Sampler randomly removes instances from the majority class until a more equitable class distribution is achieved.

In our comprehensive analysis, we explored four distinct approaches, each involving the application of both SMOTE and RandomUnderSampler, with and without the integration of Principal Component Analysis (PCA). The

incorporation of PCA allowed us to investigate how different methods of handling imbalanced datasets influenced model performance and generalization.

PCA, or Principal Component Analysis, is a dimensionality reduction technique that transforms a dataset into a new coordinate system with uncorrelated variables known as principal components. In addressing the class imbalance, PCA offers benefits such as dimensionality reduction, decorrelation for feature independence, and noise reduction, enhancing model efficiency and robustness. In employing PCA, we determined the number of components by assessing the cumulative explained variance, selecting a threshold of 95%.

Furthermore, for each of the four approaches, we conducted experiments both with and without hyperparameter tuning. This thorough exploration aimed to assess the impact of hyperparameter optimization on the effectiveness of our techniques in handling class imbalance.

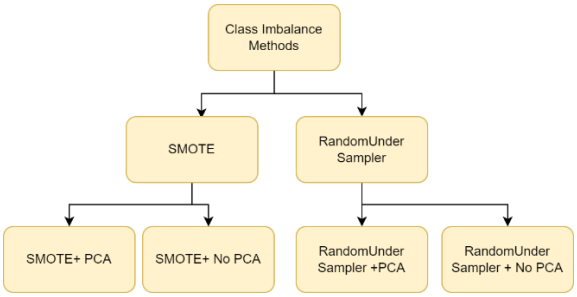


Figure 19: Class Imbalance and Feature Reduction Techniques

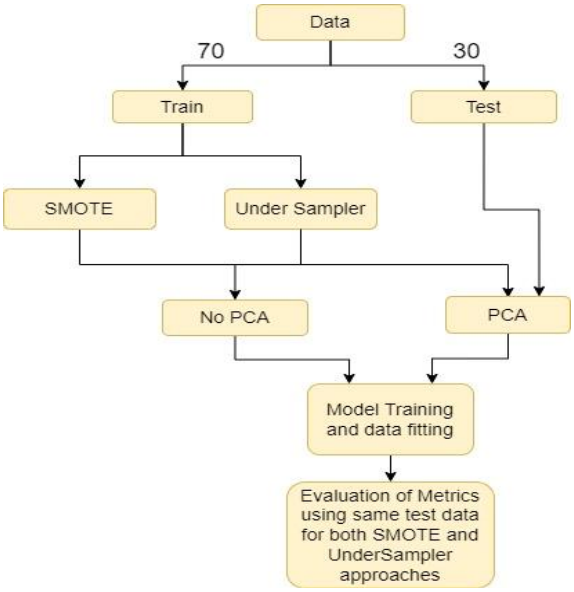


Figure 20: Flowchart for modeling procedure

From the figure 20, We divided our dataset into 70% training and 30% testing sets. SMOTE and Under Sampling were exclusively applied to the training set to address class imbalance. Both approaches were conducted with and without PCA. If PCA was applied to the training dataset, the same transformation was performed on the test data. Following these preprocessing steps, we trained the models and assessed performance metrics using identical test data for both SMOTE and Under sampler techniques, enabling comprehensive comparison across all four approaches.

Our project is centered around a binary classification task, categorizing pilots' licenses into two primary classes: "Have a disease" or "Do Not have Disease" (YES/1 or NO/0). The modeling technique in this project is supervised machine learning since the models learn from labeled training data to predict the class labels of new, unseen data. Our main aim is to construct a classification model capable of accurately predicting an individual's risk of developing a particular ailment outlined in the FAA list. We intend to map Chapters IX, VI, and X to a target variable (label) for model training and testing.

In this project, a diverse set of classification models, including Support Vector Machines (SVM), MLP (Multi-layer Perceptron Classifier), Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Gradient Boosting, XGBoost, Gaussian Naïve Bayes were employed. To assess the efficiency of these models, their performance was systematically evaluated using various metrics. Furthermore, hyperparameters for each model were fine-tuned utilizing techniques such as GridSearchCV or RandomizedSearchCV to attain optimal results.

GridSearchCV operates by systematically exploring a predefined set of hyperparameters, evaluating the model's performance for each set. It then selects the hyperparameter configuration that yields the best performance based on a specified evaluation metric, such as roc_auc, accuracy, precision, recall, or F1 score.

For a few of the models within the project, we opted for RandomizedSearchCV as an alternative hyperparameter tuning technique. In contrast to GridSearchCV, RandomizedSearchCV does not exhaustively search through all potential hyperparameter combinations. Instead, it samples a fixed number of hyperparameter settings from specified probability distributions. This random sampling approach facilitates a more efficient exploration of the hyperparameter space, particularly when dealing with large search spaces. Although it does not guarantee an exhaustive search, RandomizedSearchCV often identifies effective hyperparameter configurations with reduced computational cost compared to GridSearchCV.

Results and Analysis:

Our primary goal is to develop a model that accurately predicts the risk of circulatory (IX), respiratory (X), and nervous system (VI) diseases. Given the imbalanced nature of the data, our performance assessment will rely on the following metrics:

- Area under ROC curve
- F1 Score
- Area under Precision-Recall curve
- Confusion matrix

In the confusion matrix, we focused more on False negatives (Type II error) compared to false positives (Type I error) because in disease prediction scenarios, minimizing False Negatives is paramount due to the potential harm and adverse consequences associated with failing to identify a true positive case. This oversight may result in delayed treatment, disease progression, and compromised patient outcomes. Emphasizing the reduction of False Negatives is essential in disease prediction models to ensure timely identification and intervention for individuals genuinely affected by the condition.

We have chosen not to focus much on accuracy as a primary metric for distinguishing due to its potential for deception in imbalanced datasets. High accuracy can be achieved by predominantly predicting the majority class, neglecting the minority class, and providing an inaccurate representation of overall model performance, especially for the minority class.

Chapter IX (Circulatory System):

The Random UnderSampling approach demonstrates superior performance in Chapter IX compared to other approaches.

AUC ROC Curve & Precision-Recall Curve:

The comparison of AUC-ROC values is depicted in Figure 21. Notably, the Random Forest, XGBoost, and Gradient Boosting models emerge as the leading performers, with the highest result of ROC-AUC at 0.77, while Gaussian Naïve Bayes lags with the lowest score AUC of 0.70. AUC-ROC values for all models span from 0.77 to 0.70. Additional insights into the simulation results for alternative strategies are available in Appendix Section 1.1.

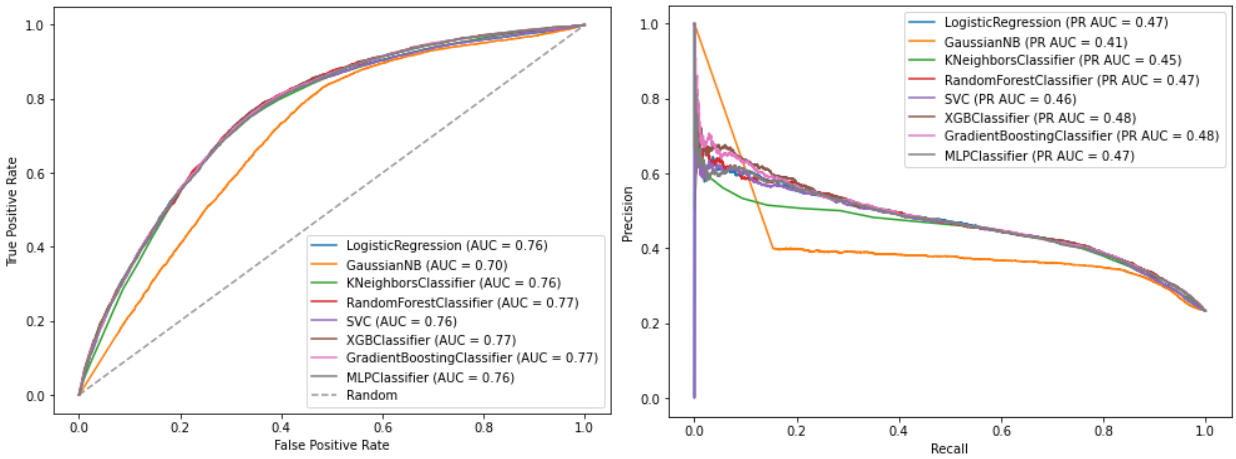


Figure 21: AUC-ROC & PR-AUC graphs for the Undersampling technique

The graphical representation in Figure 21 illustrates the comparison of PRAUC values. Remarkably, the XGBoost and Gradient Boosting models exhibit the highest PRAUC, achieving a notable score of 0.48, while Gaussian Naïve Bayes falls behind with the lowest recorded score of 0.41. Across all models, PRAUC values range from 0.48 to 0.41. For further details on simulation results involving alternative strategies, refer to Appendix Section 1.1.

Model Name	**Confusion Matrix** TN FP FN TP	**AUC ROC**	**Max F1-Score**	**PRAUC**	**Precision**	**Recall**	**Accuracy**
Logistic Regression	11100 4549 1416 3342	0.76	0.53	0.47	0.42	0.72	0.70
Gaussian Naïve Bayes	9930 5719 1512 3246	0.70	0.53	0.41	0.36	0.68	0.65
KNN	10321 5328 1177 3581	0.76	0.49	0.45	0.41	0.72	0.69
Random Forest	10573 5076 1207 3551	0.77	0.53	0.47	0.40	0.78	0.67
SVM	10849 4800 1321 3437	0.76	0.53	0.46	0.42	0.71	0.70
XGBoost	10094 5555 1057 3701	0.77	0.52	0.48	0.40	0.78	0.68
Gradient Boosting	10000 5649 1046 3712	0.77	0.53	0.48	0.40	0.78	0.67
MLP	10506 5143 1234 3524	0.76	0.53	0.47	0.39	0.77	0.67

Table 2.1: The above table provides the results of all 8 models for the Undersampling Approach related to target_IX

Table 2.1 presents a comprehensive overview of performance metrics, including the Confusion Matrix, AUC ROC, Max F1-Score, PRAUC, Precision, Recall, and Accuracy for each model employing the Undersampling technique on target_IX are meticulously documented in this table. This detailed analysis provides insights into the efficacy of the Undersampling approach in enhancing the models' predictive capabilities for target_IX.

Max-F1 Scores:

In Figure 22, the maximum F1 scores obtained from all four distinct approaches for each model are presented. It is noteworthy that the Undersampling technique consistently yields the highest maximum F1 scores across all approaches.

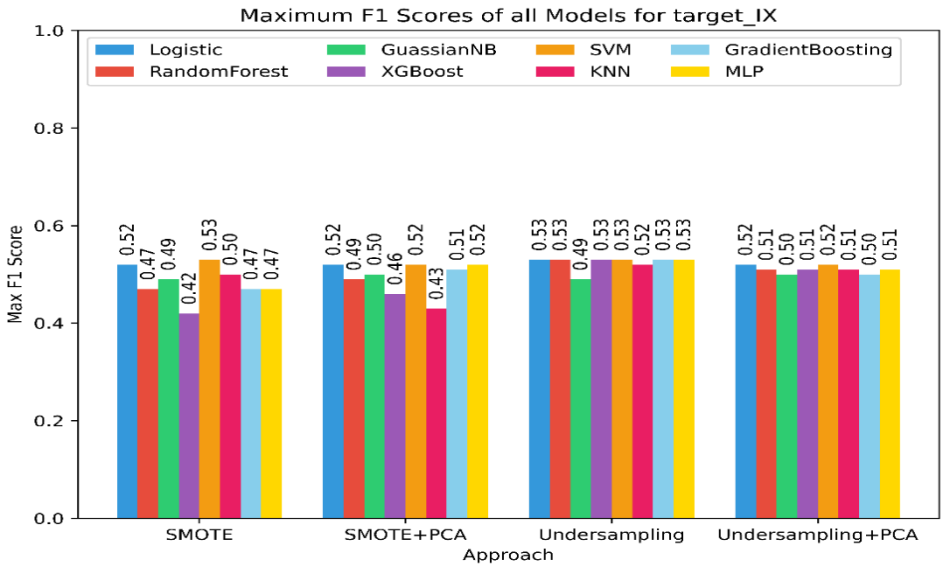


Figure 22: Max-F1 Scores for the all the 4 approaches

Chapter X (Respiratory System):

The Random UnderSampling approach demonstrates superior performance in Chapter X compared to other approaches.

AUC ROC Curve & Precision-Recall Curve:

The comparison of AUC-ROC values is depicted in Figure 23. Notably, the Random Forest, Gradient Boosting, and MLP models emerge as the leading performers, with the highest result at 0.68, while SVM, KNN, and Gaussian Naïve Bayes lags with the lowest score of 0.66. AUC-ROC values for all models span from 0.68 to 0.66. Additional insights into the simulation results for alternative strategies are available in Appendix Section 1.2.

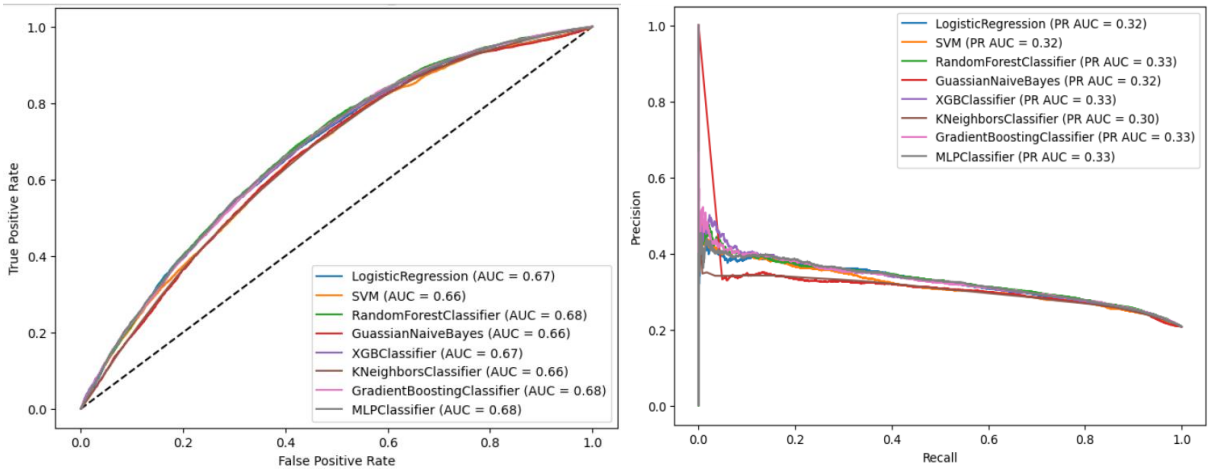


Figure 23: AUC-ROC & PR-AUC graphs for the Undersampling technique

The graphical representation in Figure 23 illustrates the comparison of PRAUC values. Remarkably, the Random Forest, XGBoost, Gradient Boosting, MLP models exhibit the highest PRAUC, achieving a notable score of 0.38, while KNN falls behind with the lowest recorded score of 0.30. Across all models, PRAUC values range from 0.38 to 0.30. For further details on simulation results involving alternative strategies, refer to Appendix Section 1.2.

Model Name	**Confusion Matrix** **TN FP** **FN TP**	**AUC ROC**	**Max F1-Score**	**PRAUC**	**Precision**	**Recall**	**Accuracy**
Logistic Regression	14065 10205 / 2062 4319	0.67	0.41	0.32	0.29	0.70	0.58
Gaussian Naïve Bayes	12488 11782 / 1744 4637	0.66	0.41	0.32	0.30	0.55	0.65
KNN	13223 11047 / 1977 4404	0.66	0.40	0.30	0.29	0.63	0.60
Random Forest	14033 10237 / 1984 4397	0.68	0.42	0.33	0.29	0.76	0.56
SVM	10786 13484 / 1322 5059	0.66	0.41	0.32	0.28	0.73	0.55
XGBoost	13386 10884 / 1860 4521	0.67	0.42	0.33	0.30	0.65	0.61
Gradient Boosting	13320 10950 / 1832 4549	0.68	0.42	0.33	0.29	0.75	0.56
MLP	12211 12059 / 1553 4828	0.68	0.41	0.33	0.29	0.70	0.58

Table 2.2: The above table provides the results of all 8 models for the Undersampling Approach related to target_X

Table 2.2 presents a comprehensive overview of performance metrics, including Confusion Matrix, AUC ROC, Max F1-Score, PRAUC, Precision, Recall, and Accuracy associated with the application of the Undersampling technique on target_X across various models are meticulously detailed in the table. This tabulated representation serves as a valuable reference for assessing the efficacy of each model in the context of undersampling, providing a clear and organized summary of their respective performance across these critical metrics.

Max-F1 Scores:

In Figure 24, the maximum F1 scores obtained from all four distinct approaches for each model are presented. It is noteworthy that the Undersampling technique consistently yields the highest maximum F1 scores across all approaches.

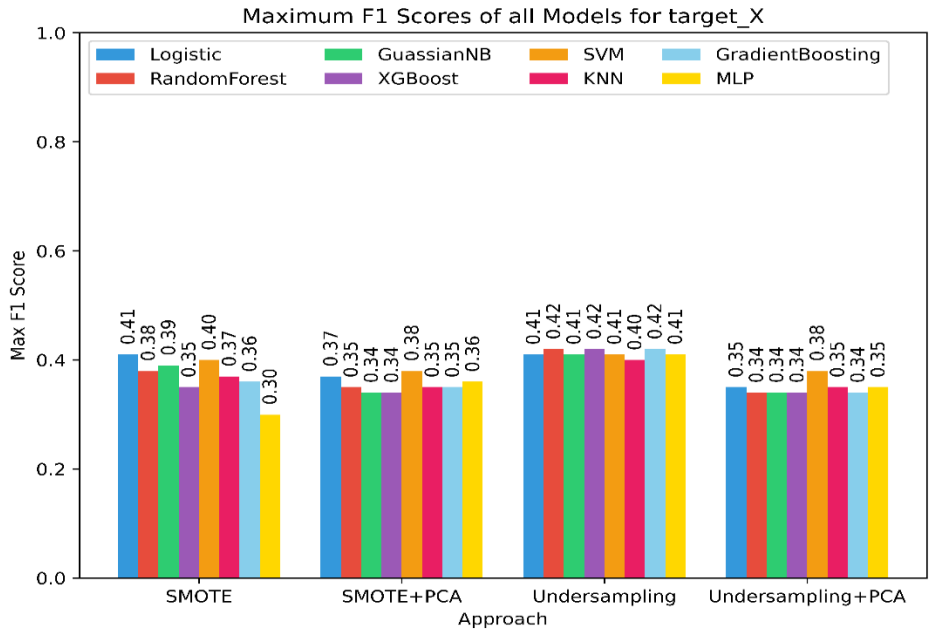


Figure 24: Max-F1 Scores for the all the 4 approaches

Chapter VI (Nervous System):

The Random UnderSampling approach demonstrates superior performance in Chapter VI compared to other approaches.

AUC ROC Curve & Precision-Recall Curve:

The comparison of AUC-ROC values is depicted in Figure 25. Notably, the Random Forest, XGBoost, Gradient Boosting, and MLP models emerge as the leading performers, with the highest result at 0.65, while KNN, and Gaussian Naïve Bayes lag with the lowest score of 0.62. AUC-ROC values for all models span from 0.65 to 0.62. Additional insights into the simulation results for alternative strategies are available in Appendix Section 1.3.

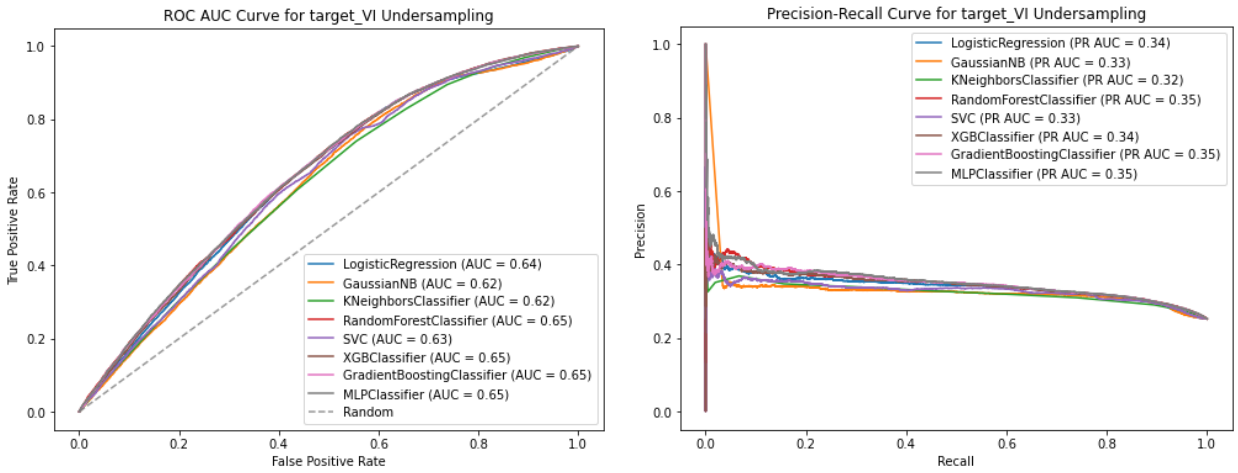


Figure 25: AUC-ROC & PR-AUC graphs for the Undersampling technique

The graphical representation in Figure 25 illustrates the comparison of PRAUC values. Remarkably, the Random Forest, Gradient Boosting, MLP models exhibit the highest PRAUC, achieving a notable score of 0.35, while KNN falls behind with the lowest recorded score of 0.32. Across all models, PRAUC values range from 0.35 to 0.32. For further details on simulation results involving alternative strategies, refer to Appendix Section 1.3.

Model Name	**Confusion Matrix** **TN FP** **FN TP**	**AUC ROC**	**Max F1-Score**	**PRAUC**	**Precision**	**Recall**	**Accuracy**
Logistic Regression	11057 10302 2106 5089	0.64	0.46	0.34	0.33	0.71	0.57
Gaussian Naïve Bayes	14874 6485 3996 3199	0.62	0.45	0.33	0.33	0.44	0.63
KNN	12166 9193 2791 4404	0.62	0.44	0.32	0.32	0.61	0.58
Random Forest	9287 12072 1460 5735	0.65	0.46	0.35	0.32	0.80	0.53
SVM	11557 9802 2415 4780	0.63	0.45	0.33	0.33	0.66	0.57
XGBoost	10196 11163 1815 5380	0.65	0.46	0.34	0.33	0.75	0.55
Gradient Boosting	9711 11648 1615 5580	0.65	0.46	0.35	0.32	0.78	0.54
MLP	9725 11634 1632 5563	0.65	0.46	0.35	0.32	0.33	0.54

Table 2.3: The above table provides the results of all 8 models for the Undersampling Approach related to target_VI

Table 2.3 presents a comprehensive overview of performance metrics corresponding to the Confusion Matrix, AUC ROC, Max F1-Score, PRAUC, Precision, Recall, and Accuracy associated with various models employing the Undersampling technique on target_VI. The values are meticulously documented, providing a detailed assessment of each model's efficacy in handling undersampled data. This tabulated representation serves as a valuable reference for evaluating the performance and effectiveness of the implemented techniques across the specified metrics.

Max-F1 Scores:

In Figure 26, the maximum F1 scores obtained from all four distinct approaches for each model are presented. It is noteworthy that the Undersampling technique consistently yields the highest maximum F1 scores across all approaches.

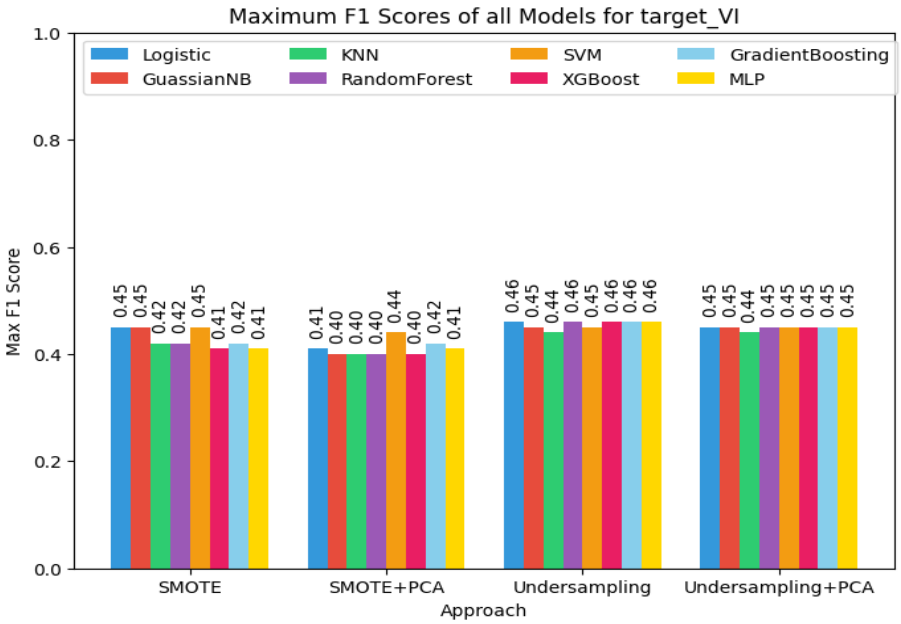


Figure 26: Max-F1 Scores for the all the 4 approaches

Process Validation:

Kovida and Bhavani (2 DSA Students) attended two weekly meetings on Tuesdays with Dr. Gopichand (DISC project supervisor), and another meeting on Thursdays with Dr. Gopichand and Dr. Beattie (practicum project supervisor). These meetings aimed to validate the process and review every step we undertook.

Deliverables:

The deliverables of the project are as listed below in Table 3:

Task	**Description**	**Due Dates**
Weekly Meetings	Project update Meeting with Dr. Gopichand for Review.	Every Tuesday of this semester
	Consultation with Dr. Beattie to assess the advancement of project tasks and seek recommendations and evaluations.	Every Thursday of this semester
Project Proposal	Submit a Project Proposal to the committee	Sep 1
Mid semester progress report	Submit a Midterm Progress Report to the committee	Oct 20
Practicum Sponsor Evaluation	Present and submit the Final report to the Practical Sponsor team	Nov 17
Peer Presentation	Present a project draft presentation to the student peers for review	Nov 17

Final Report	Submit Final Project Report to the committee	Dec 1
Presentation Slides	Submit Final Project Presentation to the committee	Dec 1
Oral Presentation	Present the Final Project Presentation to the committee	Dec 8

Table 3: Deliverables of the Project

References:

- IBM Market-Scan User Guide: https://theclearcenter.org/wp-content/uploads/2020/01/IBM-MarketScan-User-Guide.pdf
- Market-Scan Data Dictionary: https://studylib.net/doc/25537904/2017-marketscan-ccae-mdcr-data-dictionary
- ICD-10-CM python package: https://pypi.org/project/icd10-cm/
- Previous Student Work: https://github.com/JUmaMaheshwarReddy/FAA-Pliot-License-Renewal-Forecasting
- PCA: https://statisticsbyjim.com/basics/principal-component-analysis/
- Logistic Regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- RandomForest:https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- Naïve Bayes:https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
- SVM:https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- Gradient Boosting: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
- KNN:https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- MLP:https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
- SMOTE:https://www.geeksforgeeks.org/imbalanced-learn-module-in-python/
- Undersampler:https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/

Self-Assessment:

Individual Learning Objectives:

- Gain familiarity working with remote database (PostgreSQL) on a server to understand, process, access via programming (python) and perform required analysis.
- Comprehending the dataset, including the significance of each feature. Through this understanding, we aim to identify and prioritize the essential features that will play a pivotal role in the subsequent modeling process.
- Enhance programming skills, focusing on data understanding, and proficiently implementing preprocessing steps. This includes data cleaning, feature extraction, standardization techniques, and implementing robust ML models and their evaluation procedures. Seek to acquire new methods or techniques to optimize results.
- Gain hands-on experience by working with real-time data, understanding its challenges, and opportunities for improving model performance.

Throughout this project, we successfully achieved the outlined objectives, demonstrating a comprehensive analysis of medical records. Our efforts included expanding the dataset by identifying critical features in various tables, conducting rigorous data cleaning, and preprocessing in SQL, and transforming raw data into a meaningful and usable format for modeling. We validated previous work and developed a robust predictive model with accurate predictions.

This project provided valuable opportunities to apply diverse Data Science and Analytics (DSA) skills. We gained hands-on experience using PostgreSQL, becoming adept at handling large datasets, and honing our SQL proficiency. Notably, data cleaning and preprocessing emerged as crucial DSA skills, encompassing feature engineering, feature

reduction, handling class imbalances, model selection, training, and hyperparameter tuning using techniques like GridSearchCV and RandomizedSearchCV.

An essential learning point was the significance of considering metrics beyond accuracy when dealing with imbalanced data. AUC ROC, F1 score, and PRAUC became pivotal in assessing model performance. Additionally, our visualization skills played a key role in effectively communicating project results to stakeholders. We experimented with various charts and graphs to determine the most impactful visual representations.

The complexity of our project, involving eight different models across four approaches (SMOTE, SMOTE+PCA, UNDERSAMPLER, UNDERSAMPLER+PCA) for three target variables, required a strategic presentation of findings. We developed skills to convey results in reports and presentations comprehensively without overwhelming the audience.

Furthermore, we acquired proficiency in collaborative tools such as Jupyter Notebook and GitHub, enabling effective teamwork and codebase management. Independently learning these tools showcased our adaptability to new challenges and our capability to solve complex problems.

This project was undertaken as part of a 4-hour credit practicum within the research project of the OU Data Institute for Societal Challenges (DISC).

Faculty Supervisors: Dr. Beattie Matt J and Dr. Danala Gopichandh
Company and Sponsor: Dr. David Ebert, Data Institute for Societal Challenges (DISC).

Appendix:

Abbreviations:

Description	**Abbreviation**
Federal Aviation Administration	FAA
Commercial Claims and Encounters Database	CCAE
Inpatient Admissions Table	CCAEI
Inpatient Services Table	CCAES
Outpatient Pharmaceutical Claims Table	CCAED
Diagnosis Version	dxver
Major Diagnosis Category	mdc
Therapeutic Group	thergrp
Synthetic Minority Over-sampling Technique	SMOTE

In the CCAEI table, we transformed the MDC column to align with our project's objectives. We identified certain MDC codes as highlighted in the below figure as important and created a new column, "mdc_flag," where a value of 1 indicates a match with the important codes and 0 otherwise. Similarly, a new column, "non_mdc_flag," was introduced, with a value of 1 for MDC codes deemed unimportant and 0 for all others. This can be seen in the "*Data Cleaning and Transformation of CCAEI table*" *figure:3*.

ATTACHMENT D - MDC

Value	Label
00	Missing/Invalid Diagnosis
01	Nervous
02	Eye
03	Ear, Nose, Mouth & Throat
04	Respiratory
05	Circulatory
06	Digestive
07	Liver, Pancreas
08	Musculoskeletal
9	Skin, Breast
10	Metabolic
11	Kidney
12	Male Reproductive
13	Female Reproductive
14	Pregnancy, Childbirth
15	Newborns
16	Blood
17	Myeloproliferative Diseases
18	Infections
19	Mental
20	Alcohol/Drug Use
21	Injuries, Poisonings
22	Burns
23	Health Status
24	Multiple Trauma
25	HIV Infections

In the CCAED table, we considered specific "thergrp" values as important (highlighted in the figure) and removed records with values outside this list to focus our analysis on key information. This can be seen in the *"Data Cleaning and Transformation of CCAED table" figure:4.*

ATTACHMENT L THERGRP

Value	Label	Value	Label
01	Antihistamines & Comb. (Class 1)	17	Gastrointestinal Drugs (Classes 147-162, 273)
02	Anti-infective Agents (Classes 2-20)	18	Gold Compounds (Class 163)
03	Antineoplastic Agents (Classes 21-22, 260-265)	19	Heavy Metal Antagonists (Class 164)
04	Autonomic Drugs (Classes 23-33)	20	Hormones & Synthetic Substitutes (Classes 165-180 246 252-253 256 266-268)
05	Blood Derivatives (Class 34)	21	Immunosuppressants (Class 181)
06	Blood Form/Coagul Agents (Classes 35-45, 259)	22	Anesthetics, Local (Class 122)
07	Cardiovascular Agents (Classes 46-56, 245, 250, 271)	23	Oxytoxics (Class 183)
08	Central Nervous System (Classes 57-77, 272)	24	Radioactive Agents (Class 184)
09	Contraceptive Cream/Foam/Devices (Classes 78)	25	Serums, Toxoids, Vaccines (Classes 185-189)
10	Dental Agents (Classes 79-83)	26	Skin & Mucous Membrane (Classes 190-213, 242)
11	Diagnostic Agents (Classes 84-98, 239, 243-244, 247)	27	Smooth Muscles Relaxants (Classes 214-216)
12	Disinfectants (Class 99)	28	Vitamins & Comb (Classes 217-233)
13	Electrolytic, Caloric, Water (Classes 100-126, 241, 292)	29	Unclassified Agents (Classes 234-236, 251, 254, 257-258, 270)
14	Enzymes (Class 127)	30	Devices and Non-drug Items (Class 237)
15	Antituss/Expector/Mucolytic (Classes 128-131, 248, 255)	31	Pharmaceutical Aids/Adjuvants (Class 238)
16	Eye, Ear, Nose Throat (Classes 132-146, 240, 290)	99	Other/unavailable

The list of diseases in the figure represents the conditions of interest to the FAA. We specifically focused on the highlighted diseases, using the ICD-10-CM library to categorize them into chapters. These chapters include IX, X, XIX, and VII, which serve as our target variables in the modeling process.

Finalized Incapacitating Conditions List
Acute glaucoma (narrow angle)
Acute hemorrhage (intracranial hemorrhage, gastrointestinal, unspecified)
Anaphylactic shock, unspecified (initial encounter)
Aneurysms and dissections non-site specific
Dissection of aorta
Cardiac conduction abnormalities (A-fib, bradycardia/high grade AV block, unstable tachycardia)
Cardiac tamponade
Headache (migraine)
Hypoglycemia, unspecified (Includes hypoglycemia syndrome and hypoglycemia disorder)
Myocardial infarction/cardiac arrest
Nephrolithiasis (acute renal colic)
Pulmonary embolism (pulmonary, mesenteric, retinal, etc.)
Seizure, unspecified
Stroke (CVA) - Cerebal infarction, unspecified
Tension pneumothorax (unspecified)
Vertigo

In the CCAEI table, we selected the columns highlighted in yellow for consideration.

COMMERCIAL CLAIMS AND ENCOUNTERS
MEDICARE SUPPLEMENTAL AND COORDINATION OF BENEFITS
INPATIENT ADMISSIONS TABLE

Name	Long Name	Data Type	Name	Long Name	Data Type	Name	Long Name	Data Type
ADMDATE	Date of Admission	DT	EIDFLAG	Enrollee ID Derivation Flag	C	POADX9	Present On Admission Diagnosis 9	C
ADMTYP	Admission Type	C	EMPREL	Relation to Employee	C	POAPDX	Present On Admission Diagnosis Principal	C
AGE	Age of Patient	N	ENRFLAG	Enrollment Flag	C	PPROC	Procedure Principal	C
AGEGRP	Age Group	C	ENROLID	Enrollee ID	N	PROC1	Procedure 1	C
CASEID	Case and Services Link	N	HLTHPLAN	Health Plan Indicator	C	PROC2	Procedure 2	C
DATATYP	Data Type	N	HOSPNET	Net Payments: Hospital	N	PROC3	Procedure 3	C
DAYS	Length of Stay	N	HOSPPAY	Payments Hospital	N	PROC4	Procedure 4	C
DISDATE	Date of Discharge	DT	INDSTRY	Industry	C	PROC5	Procedure 5	C
DOBYR	Patient Birth Year	N	MDC	Major Diagnostic Category	C	PROC6	Procedure 6	C
DRG	Diagnosis Related Group	N	MHSACOVG	Coverage Indicator MHSA	C	PROC7	Procedure 7	C
DSTATUS	Discharge Status	C	MSA	Metropolitan Statistical Area	N	PROC8	Procedure 8	C
DX1	Diagnosis 1	C	PDX	Diagnosis Principal	C	PROC9	Procedure 9	C
DX2	Diagnosis 2	C	PHYFLAG	Physician Specialty Coding Flag	C	PROC10	Procedure 10	C
DX3	Diagnosis 3	C	PHYSID	Physician ID	N	PROC11	Procedure 11	C
DX4	Diagnosis 4	C	PHYSNET	Net Payments Physician	N	PROC12	Procedure 12	C
DX5	Diagnosis 5	C	PHYSPAY	Payments Physician	N	PROC13	Procedure 13	C
DX6	Diagnosis 6	C	PLANTYP	Plan Indicator	N	PROC14	Procedure 14	C
DX7	Diagnosis 7	C	POADX1	Present On Admission Diagnosis 1	C	PROC15	Procedure 15	C
DX8	Diagnosis 8	C	POADX10	Present On Admission Diagnosis 10	C	REGION	Region	C
DX9	Diagnosis 9	C	POADX11	Present On Admission Diagnosis 11	C	RX	Cohort Drug Indicator	C
DX10	Diagnosis 10	C	POADX12	Present On Admission Diagnosis 12	C	SEQNUM	Sequence Number	N
DX11	Diagnosis 11	C	POADX13	Present On Admission Diagnosis 13	C	SEX	Gender of Patient	C
DX12	Diagnosis 12	C	POADX14	Present On Admission Diagnosis 14	C	STATE	State Hospital	C
DX13	Diagnosis 13	C	POADX15	Present On Admission Diagnosis 15	C	TOTCOB	COB and Other Savings: Total (Case)	N
DX14	Diagnosis 14	C	POADX2	Present On Admission Diagnosis 2	C	TOTCOINS	Coinsurance: Total (Case)	N
DX15	Diagnosis 15	C	POADX3	Present On Admission Diagnosis 3	C	TOTCOPAY	Copayment: Total (Case)	N
DXVER	Diagnosis Version	C	POADX4	Present On Admission Diagnosis 4	C	TOTDED	Deductible: Total (Case)	N
EECLASS	Employee Classification	C	POADX5	Present On Admission Diagnosis 5	C	TOTNET	Payments Net Case	N
EESTATU	Employment Status	C	POADX6	Present On Admission Diagnosis 6	C	TOTPAY	Payments Total Case	N
EFAMID	Family ID	N	POADX7	Present On Admission Diagnosis 7	C	VERSION	Version	C
EGEOLOC	Geographic Location Employee	C	POADX8	Present On Admission Diagnosis 8	C	YEAR	Date Year Incurred	N

In the CCAES table, we selected the columns highlighted in yellow for consideration.

COMMERCIAL CLAIMS AND ENCOUNTERS
MEDICARE SUPPLEMENTAL AND COORDINATION OF BENEFITS
INPATIENT SERVICES TABLE

Name	Long Name	Data Type	Name	Long Name	Data Type	Name	Long Name	Data Type
ADMDATE	Date of Admission	DT	EFAMID	Family ID	N	PHYFLAG	Physician Specialty Coding Flag	C
ADMTYP	Admission Type	C	EGEOLOC	Geographic Location Employee	C	PLANTYP	Plan Indicator	N
AGE	Age of Patient	N	EIDFLAG	Enrollee ID Derivation Flag	C	PPROC	Procedure Principal	C
AGEGRP	Age Group	C	EMPREL	Relation to Employee	C	PROC1	Procedure Code 1	C
CAP_SVC	Capitated Service-Claim Indicator	C	ENRFLAG	Enrollment Flag	C	PROCMOD	Procedure Code Modifier	C
CASEID	Case and Services Link	N	ENROLID	Enrollee ID	N	PROCTYP	Procedure Code Type	C
COB	COB and Other Savings	N	FACHDID	Facility Header Record ID	N	PROVID	Provider ID	N
COINS	Coinsurance	N	FACPROF	Facility-Professional Claim Indicator	C	QTY	Quantity of Services	N
COPAY	Copayment	N	HLTHPLAN	Health Plan Indicator	C	REGION	Region	C
DATATYP	Data Type	N	INDSTRY	Industry	C	REVCODE	Revenue Code	C
DEDUCT	Deductible	N	MDC	Major Diagnostic Category	C	RX	Cohort Drug Indicator	C
DISDATE	Date of Discharge	DT	MHSACOVG	Coverage Indicator MHSA	C	SEQNUM	Sequence Number	N
DOBYR	Patient Birth Year	N	MSA	Metropolitan Statistical Area	N	SEX	Gender of Patient	C
DRG	Diagnosis Related Group	N	MSCLMID	MarketScan Claim ID	N	STDPLAC	Place of Service	N
DSTATUS	Discharge Status	C	NETPAY	Payments Net	N	STDPROV	Provider Type	N
DX1	Diagnosis Code 1	C	NPI	National Provider Identifier	C	SVCDATE	Date Service Incurred	DT
DX2	Diagnosis Code 2	C	NTWKPROV	Network Provider Indicator	C	SVCSCAT	Service Sub-Category Code	C
DX3	Diagnosis Code 3	C	PAIDNTWK	Network Paid Indicator	C	TSVCDAT	Date Service Ending	DT
DX4	Diagnosis Code 4	C	PAY	Payment	N	UNITS	Units	N
DXVER	Diagnosis Version	C	PDDATE	Date Claim Paid	DT	VERSION	Version	C
EECLASS	Employee Classification	C	PDX	Diagnosis Principal	C	YEAR	Date Year Incurred	N
EESTATU	Employment Status	C	-	-	-	-	-	-

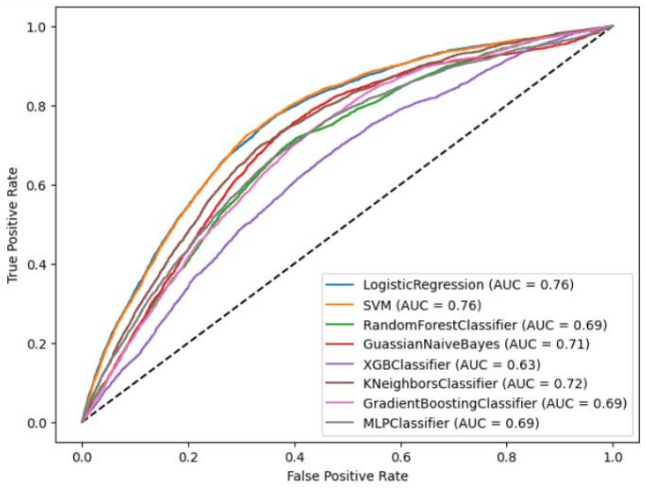
In the CCAED table, we selected the columns highlighted in yellow for consideration.

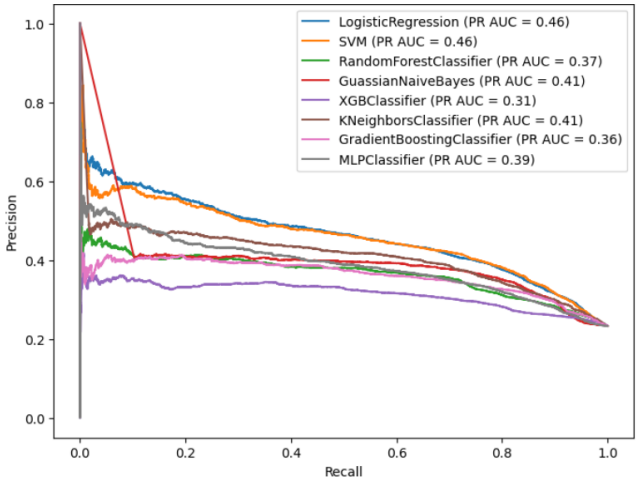
COMMERCIAL CLAIMS AND ENCOUNTERS
MEDICARE SUPPLEMENTAL AND COORDINATION OF BENEFITS
OUTPATIENT PHARMACEUTICAL CLAIMS TABLE

Name	Long Name	Data Type	Name	Long Name	Data Type	Name	Long Name	Data Type
AGE	Age of Patient	N	EIDFLAG	Enrollee ID Derivation Flag	C	PAY	Payment	N
AGEGRP	Age Group	C	EMPREL	Relation to Employee	C	PDDATE	Date Claim Paid	DT
AWP	Average Wholesale Price	N	ENRFLAG	Enrollment Flag	C	PHARMID	Pharmacy ID	N
CAP_SVC	Capitated Service-Claim Indicator	C	ENROLID	Enrollee ID	N	PHYFLAG	Physician Specialty Coding Flag	C
COB	COB and Other Savings	N	GENERID	Generic Product ID	N	PLANTYP	Plan Indicator	N
COINS	Coinsurance	N	GENIND	Generic Indicator	C	QTY	Quantity of Services	N
COPAY	Copayment	N	HLTHPLAN	Health Plan Indicator	C	REFILL	Refill Number	N
DATATYP	Data Type	N	INDSTRY	Industry	C	REGION	Region	C
DAWIND	Dispense as Written Indicator	C	INGCOST	Ingredient Cost	N	RXMR	Rx Mail Retail	C
DAYSUPP	Days Supply	N	MAINTIN	Maintenance Indicator	C	SALETAX	Sales Tax	N
DEACLAS	DEA Classification	C	METQTY	Metric Quantity	N	SEQNUM	Sequence Number	N
DEDUCT	Deductible	N	MHSACOVG	Coverage Indicator MHSA	C	SEX	Gender of Patient	C
DISPFEE	Dispensing Fee	N	MSA	Metropolitan Statistical Area	N	SVCDATE	Date Service Incurred	DT
DOBYR	Patient Birth Year	N	NDCNUM	National Drug Code	C	THERCLS	Therapeutic Class	N
EECLASS	Employee Classification	C	NETPAY	Payments Net	N	THERGRP	Therapeutic Group	C
EESTATU	Employment Status	C	NTWKPROV	Network Provider Indicator	C	VERSION	Version	C
EFAMID	Family ID	N	PAIDNTWK	Network Paid Indicator	C	YEAR	Date Year Incurred	N
EGEOLOC	Geographic Location Employee	C	-	-	-	-	-	-

Sections 1.1 (Target IX) additional results:

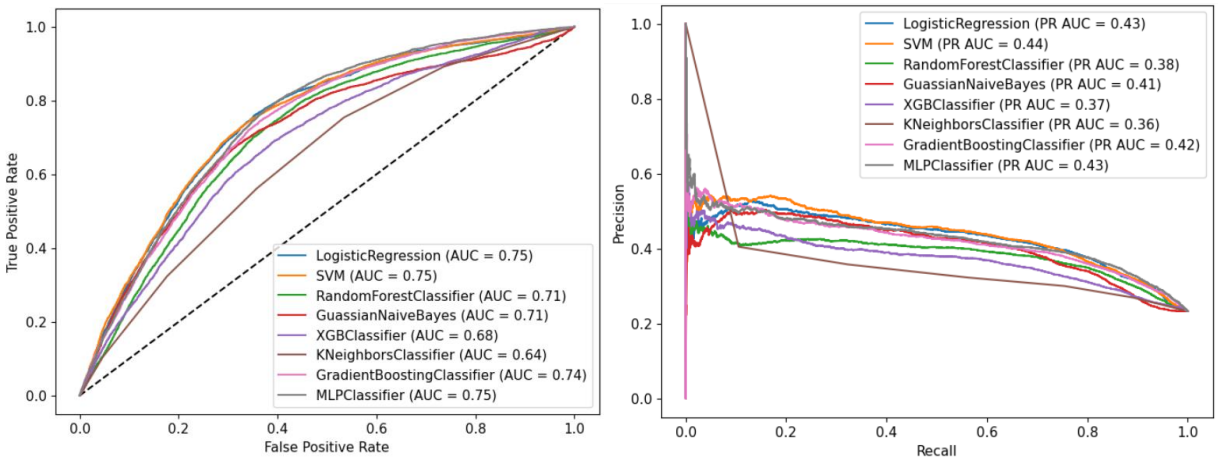
SMOTE:





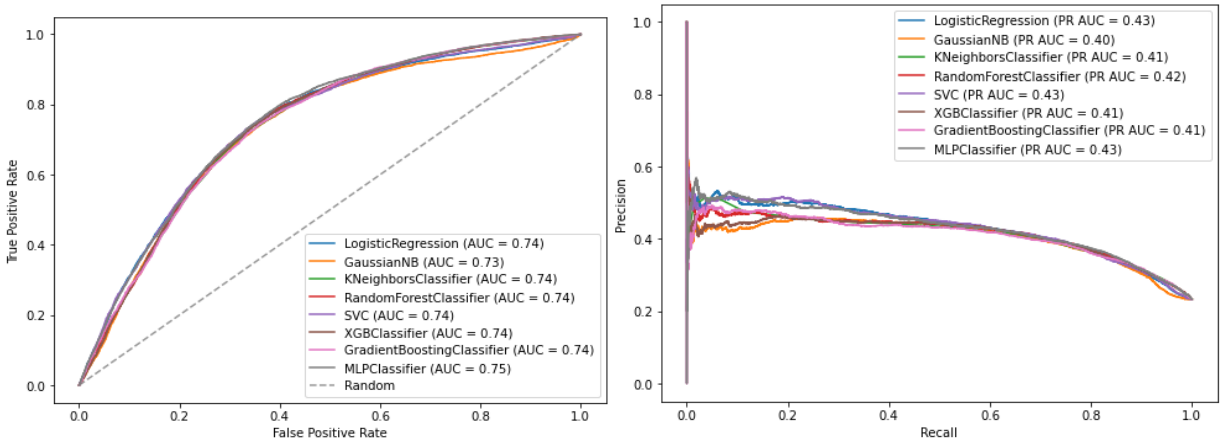
Model Name	Confusion Matrix TN FP FN TP	AUC ROC	Max F1-Score	PRAUC	Precision	Recall	Accuracy
Logistic Regression	12891 11379 1814 4567	0.76	0.52	0.46	0.41	0.72	0.69
Gaussian Naïve Bayes	13773 10497 2184 4197	0.71	0.49	0.41	0.37	0.75	0.64
KNN	9153 15117 1317 5064	0.72	0.5	0.41	0.38	0.71	0.66
Random Forest	7233 17037 925 5456	0.69	0.47	0.37	0.31	0.81	0.53
SVM	11327 12943 1465 4916	0.76	0.53	0.46	0.41	0.73	0.69
XGBoost	4626 19644 661 5720	0.63	0.42	0.31	0.24	0.98	0.29
Gradient Boosting	7880 16390 1183 5198	0.69	0.47	0.36	0.28	0.91	0.44
MLP	11249 13021 2042 4339	0.69	0.47	0.39	0.38	0.58	0.68

SMOTE + PCA:

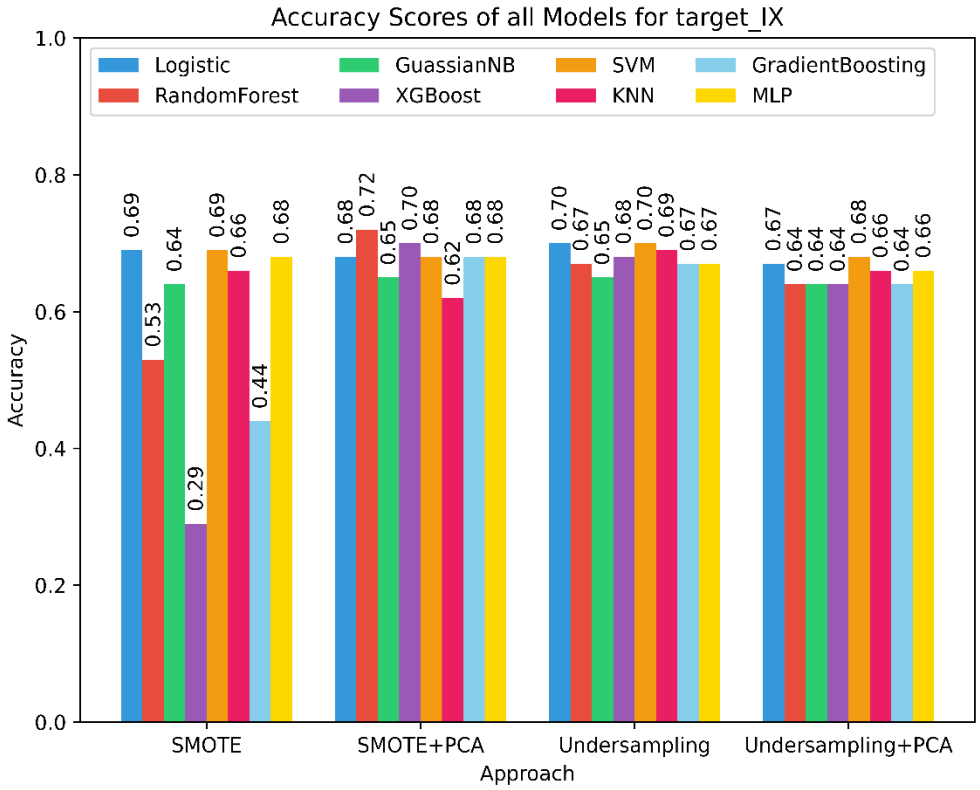


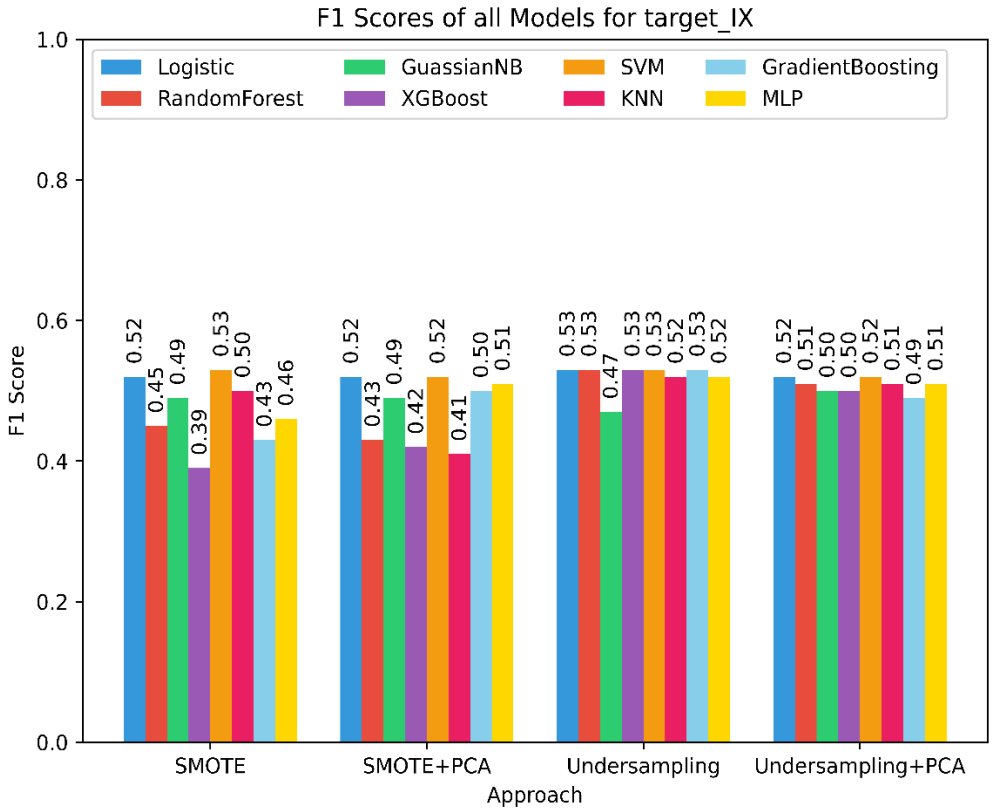
Model Name	Confusion Matrix TN FP FN TP	AUC ROC	Max F1-Score	PRAUC	Precision	Recall	Accuracy
Logistic Regression	10141 5508 1150 3608	0.75	0.52	0.43	0.40	0.74	0.68
Gaussian Naïve Bayes	10683 4966 1532 3226	0.71	0.50	0.41	0.37	0.73	0.65
KNN	7286 8363 1169 3589	0.64	0.43	0.36	0.32	0.56	0.62
Random Forest	10156 5493 1423 3335	0.71	0.49	0.38	0.40	0.46	0.72
SVM	11058 4591 1467 3291	0.75	0.52	0.44	0.40	0.73	0.68
XGBoost	9488 6161 1470 3288	0.68	0.46	0.37	0.38	0.48	0.70
Gradient Boosting	10312 5337 1334 3424	0.74	0.50	0.42	0.40	0.68	0.68
MLP	9770 5879 1062 3696	0.75	0.52	0.43	0.40	0.72	0.68

UNDERSAMPLING + PCA:



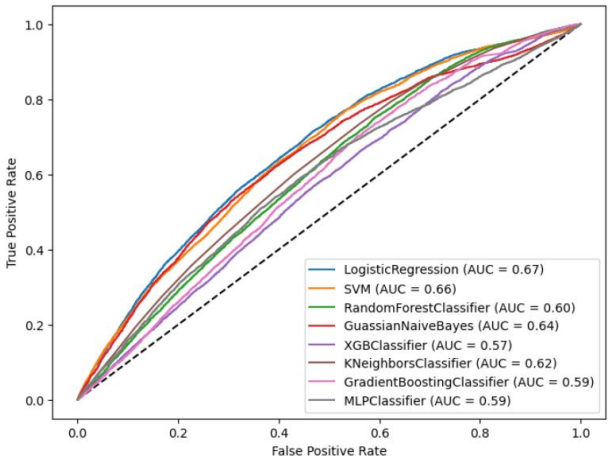
Model Name	**Confusion Matrix** **TN FP** **FN TP**	**AUC ROC**	**Max F1-Score**	**PRAUC**	**Precision**	**Recall**	**Accuracy**
Logistic Regression	10179 5470 1202 3556	0.74	0.52	0.43	0.39	0.75	0.67
Gaussian Naïve Bayes	9478 6171 1160 3598	0.73	0.50	0.40	0.37	0.76	0.64
KNN	10416 5233 1330 3428	0.74	0.51	0.41	0.39	0.75	0.66
Random Forest	10270 5379 1281 3477	0.74	0.52	0.42	0.37	0.79	0.64
SVM	10479 5170 1291 3467	0.74	0.52	0.43	0.40	0.73	0.68
XGBoost	10270 5379 1281 3477	0.74	0.51	0.41	0.37	0.79	0.64
Gradient Boosting	10322 5327 1436 3322	0.74	0.50	0.41	0.37	0.74	0.64
MLP	10771 4878 1435 3323	0.75	0.51	0.43	0.38	0.75	0.66

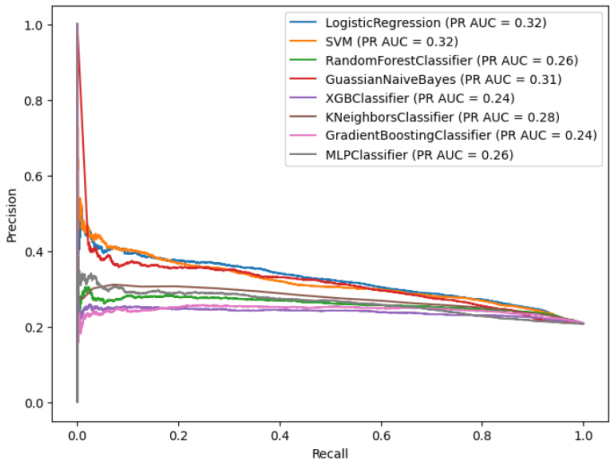

Accuracy Scores of all Models for target_IX
Logistic
RandomForest
GuassianNB
XGBoost
SVM
KNN
GradientBoosting
MLP
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.69 0.53 0.64 0.29 0.69 0.66 0.44 0.68
0.68 0.72 0.65 0.70 0.68 0.62 0.68 0.68
0.70 0.67 0.65 0.68 0.70 0.69 0.67 0.67
0.67 0.64 0.64 0.64 0.68 0.66 0.64 0.66
SMOTE
SMOTE+PCA
Undersampling
Undersampling+PCA
Approach


F1 Scores of all Models for target_IX
Logistic
RandomForest
GuassianNB
XGBoost
SVM
KNN
GradientBoosting
MLP
F1 Score
1.0
0.8
0.6
0.4
0.2
0.0
0.52 0.45 0.49 0.39 0.53 0.50 0.43 0.46
0.52 0.43 0.49 0.42 0.52 0.41 0.50 0.51
0.53 0.53 0.47 0.53 0.53 0.52 0.53 0.52
0.52 0.51 0.50 0.50 0.52 0.51 0.49 0.51
SMOTE
SMOTE+PCA
Undersampling
Undersampling+PCA
Approach

Sections 1.2 (Target X) additional results:

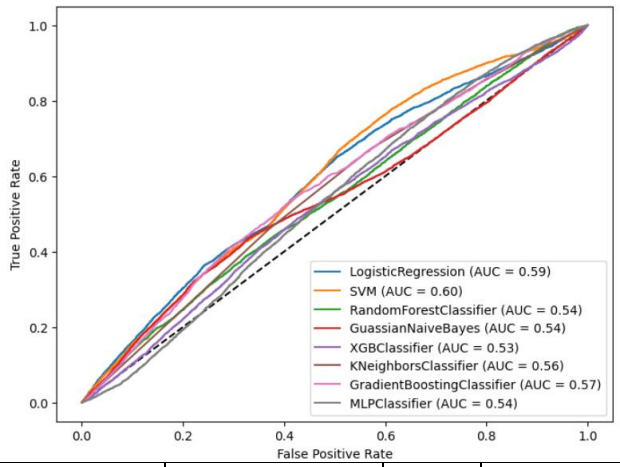
SMOTE:

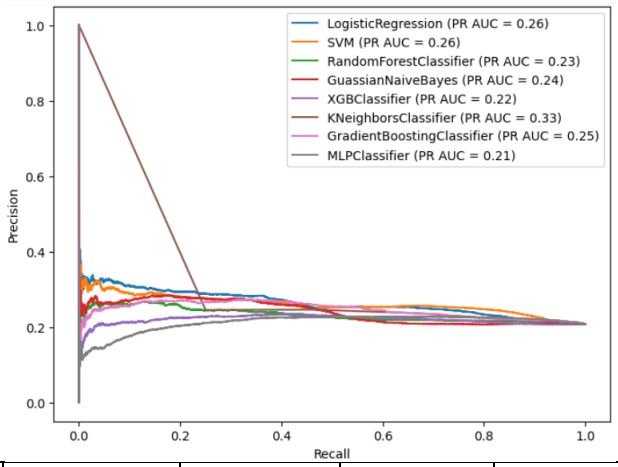




Model Name	Confusion Matrix TN FP FN TP	AUC ROC	Max F1-Score	PRAUC	Precision	Recall	Accuracy
Logistic Regression	12891 11379 1814 4567	0.67	0.41	0.32	0.29	0.68	0.59
Gaussian Naïve Bayes	13773 10497 2184 4197	0.64	0.39	0.31	0.27	0.70	0.55
KNN	9153 15117 1317 5064	0.62	0.38	0.28	0.27	0.60	0.58
Random Forest	7233 17037 925 5456	0.60	0.38	0.26	0.25	0.77	0.47
SVM	11327 12943 1465 4916	0.66	0.41	0.32	0.28	0.71	0.56
XGBoost	4626 19644 661 5720	0.57	0.36	0.24	0.21	0.97	0.25
Gradient Boosting	7880 16390 1183 5198	0.59	0.37	0.24	0.22	0.93	0.31
MLP	11249 13021 2042 4339	0.59	0.37	0.26	0.28	0.32	0.69

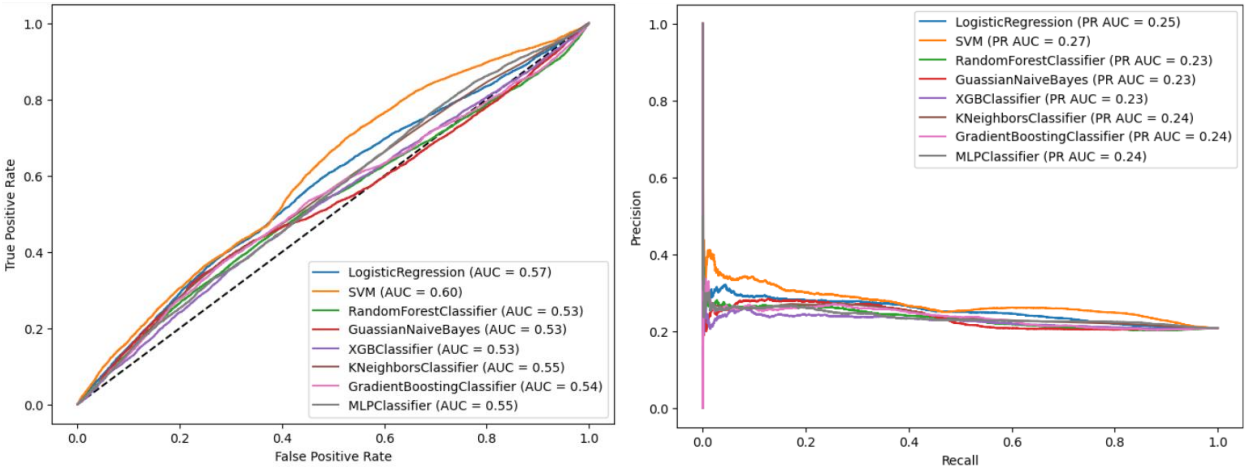
SMOTE+PCA:



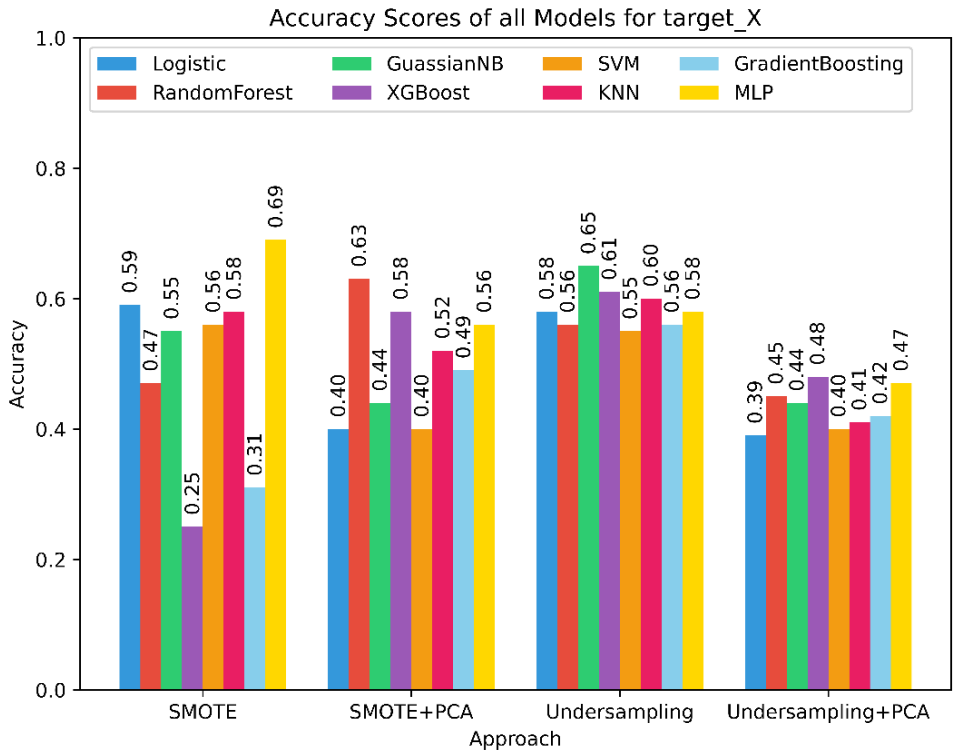


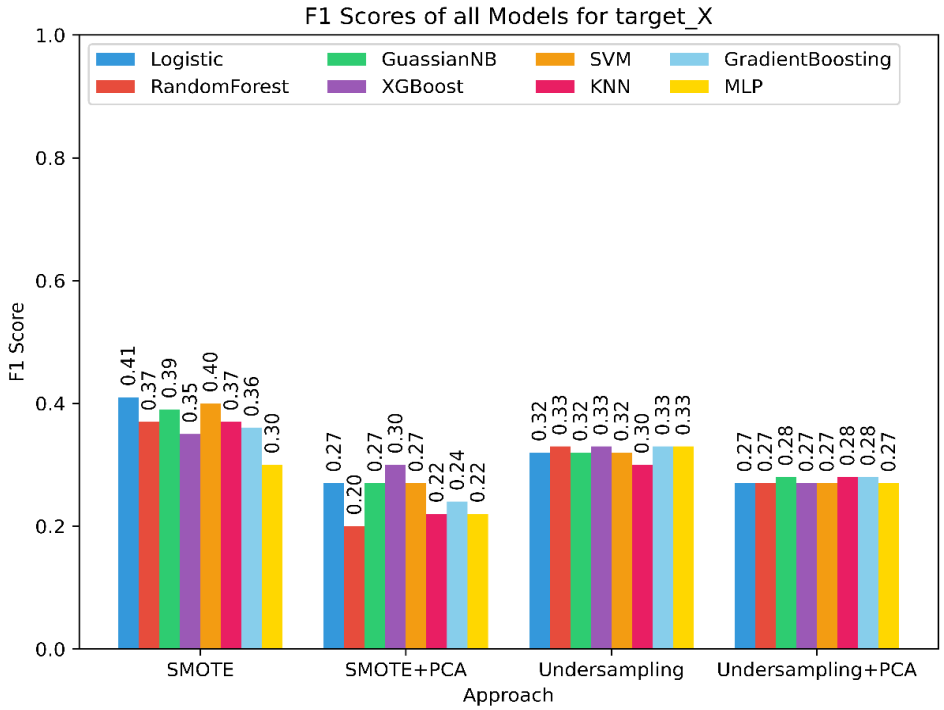
Model Name	Confusion Matrix TN FP FN TP	AUC ROC	Max F1-Score	PRAUC	Precision	Recall	Accuracy
Logistic Regression	9387 14883 1590 4791	0.59	0.37	0.26	0.18	0.53	0.40
Gaussian Naïve Bayes	10375 13895 3207 3174	0.54	0.34	0.24	0.19	0.49	0.44
KNN	5128 19142 970 5411	0.56	0.35	0.33	0.17	0.33	0.52
Random Forest	1715 22555 241 6140	0.54	0.35	0.23	0.18	0.22	0.63
SVM	9389 14881 1416 4965	0.60	0.38	0.26	0.18	0.53	0.40
XGBoost	15975 8295 4489 1892	0.53	0.34	0.22	0.19	0.30	0.58
Gradient Boosting	2984 21286 452 5929	0.57	0.35	0.25	0.17	0.37	0.49
MLP	4696 19574 771 5610	0.54	0.36	0.21	0.17	0.31	0.56

UNDERSAMPLING + PCA:



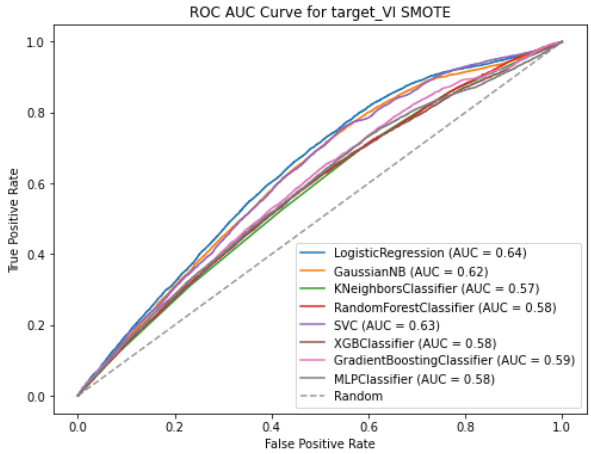
Model Name	Confusion Matrix TN FP FN TP	AUC ROC	Max F1-Score	PRAUC	Precision	Recall	Accuracy
Logistic Regression	9341 14929 1850 4531	0.57	0.35	0.25	0.18	0.54	0.39
Gaussian Naïve Bayes	10074 14196 3021 3360	0.53	0.34	0.23	0.19	0.53	0.44
KNN	4760 19510 967 5414	0.55	0.35	0.24	0.19	0.55	0.41
Random Forest	10631 13639 3193 3188	0.53	0.34	0.23	0.19	0.49	0.45
SVM	10848 13422 1749 4632	0.60	0.38	0.27	0.18	0.53	0.40
XGBoost	11633 12637 3389 2992	0.53	0.34	0.23	0.19	0.47	0.48
Gradient Boosting	9533 14737 2996 3385	0.54	0.34	0.24	0.19	0.53	0.42
MLP	5428 18842 984 5397	0.55	0.35	0.24	0.19	0.47	0.47

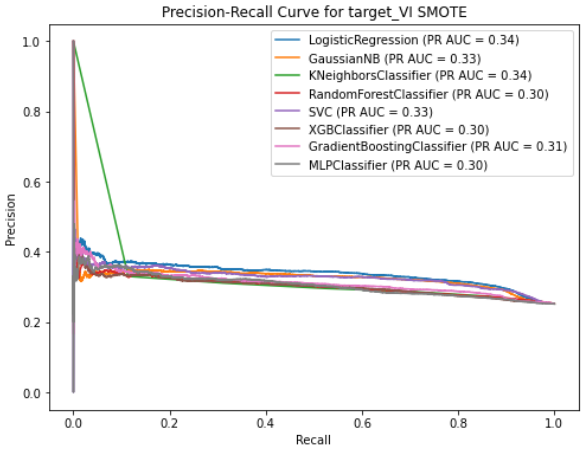

Accuracy Scores of all Models for target_X
Logistic
RandomForest
GuassianNB
XGBoost
SVM
KNN
GradientBoosting
MLP
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.59
0.47
0.55
0.25
0.56
0.58
0.31
0.69
0.40
0.63
0.44
0.58
0.40
0.52
0.49
0.56
0.58
0.56
0.65
0.61
0.55
0.60
0.56
0.58
0.39
0.45
0.44
0.48
0.40
0.41
0.42
0.47
SMOTE
SMOTE+PCA
Undersampling
Undersampling+PCA
Approach


F1 Scores of all Models for target_X
Logistic
RandomForest
GuassianNB
XGBoost
SVM
KNN
GradientBoosting
MLP
F1 Score
1.0
0.8
0.6
0.4
0.2
0.0
0.41
0.37
0.39
0.35
0.40
0.37
0.36
0.30
0.27
0.20
0.27
0.30
0.27
0.22
0.24
0.22
0.32
0.33
0.32
0.33
0.32
0.30
0.33
0.33
0.27
0.27
0.28
0.27
0.27
0.28
0.28
0.27
SMOTE
SMOTE+PCA
Undersampling
Undersampling+PCA
Approach

Sections 1.3 (Target VI) additional results:

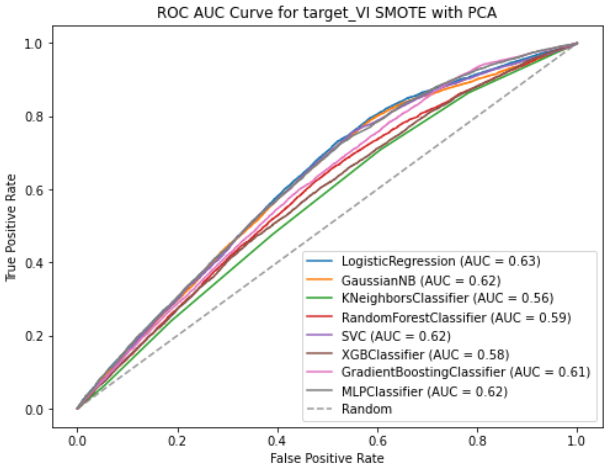
SMOTE:

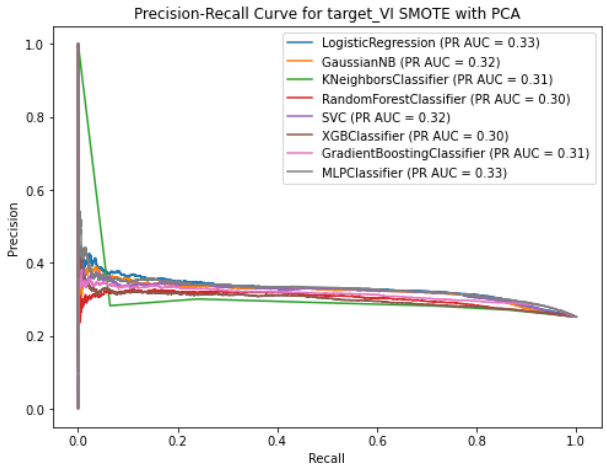




Model Name	Confusion Matrix TN FP FN TP	AUC ROC	Max F1-Score	PRAUC	Precision	Recall	Accuracy
Logistic Regression	11271 10088 2251 4944	0.64	0.45	0.34	0.33	0.69	0.57
Gaussian Naïve Bayes	10280 11079 2005 5190	0.62	0.45	0.33	0.32	0.72	0.54
KNN	12381 8978 3423 3772	0.57	0.42	0.34	0.30	0.52	0.57
Random Forest	5168 16191 1100 6095	0.58	0.42	0.30	0.27	0.85	0.39
SVM	11507 9852 2508 4687	0.63	0.45	0.33	0.32	0.65	0.57
XGBoost	4727 16632 1031 6164	0.58	0.41	0.30	0.25	0.99	0.26
Gradient Boosting	6580 14779 1275 5920	0.59	0.42	0.31	0.26	0.97	0.29
MLP	7513 13846 1719 5476	0.58	0.41	0.30	0.34	0.32	0.67

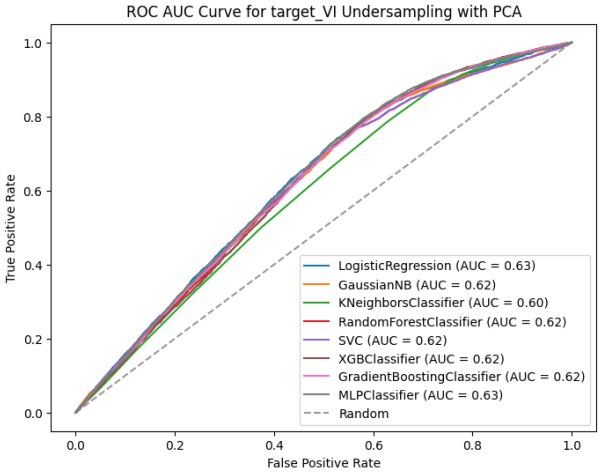
SMOTE + PCA:

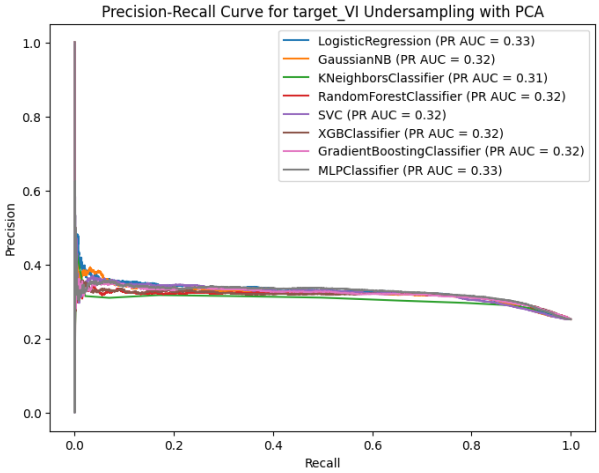




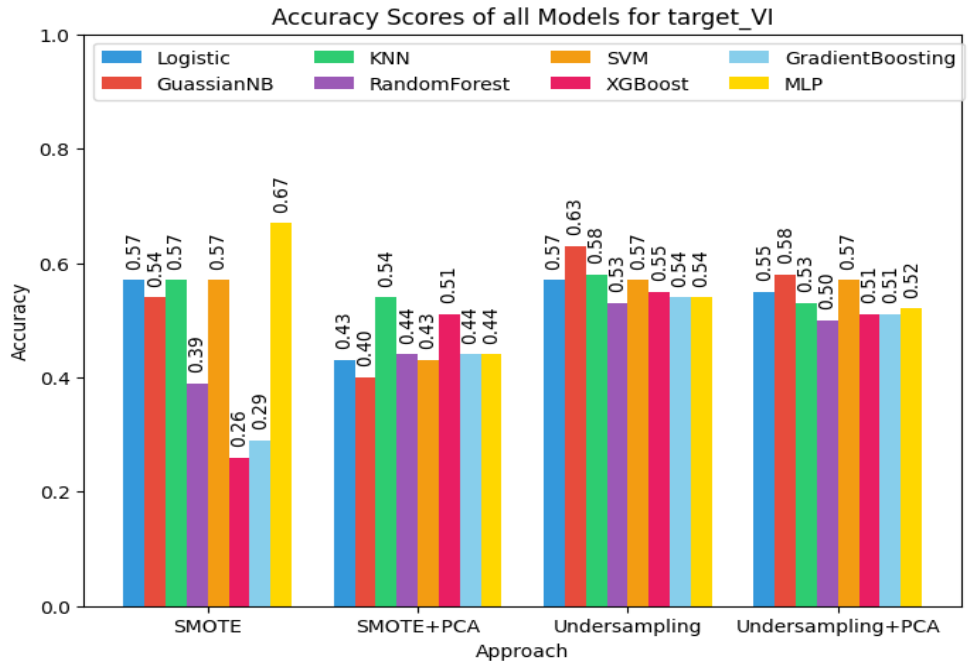
Model Name	Confusion Matrix TN FP FN TP	AUC ROC	Max F1-Score	PRAUC	Precision	Recall	Accuracy
Logistic Regression	8605 12754 3491 3704	0.63	0.41	0.30	0.23	0.51	0.43
Gaussian Naïve Bayes	7270 14089 3114 4081	0.62	0.40	0.30	0.22	0.57	0.40
KNN	12808 8551 4570 2625	0.56	0.40	0.35	0.23	0.36	0.54
Random Forest	8807 12552 3413 3782	0.59	0.40	0.28	0.23	0.53	0.44
SVM	7871 13488 1304 5891	0.62	0.44	0.30.	0.23	0.56	0.43
XGBoost	11937 9422 4448 2747	0.58	0.40	0.28	0.23	0.38	0.51
Gradient Boosting	8860 12499 3533 3662	0.61	0.42	0.29	0.23	0.51	0.44
MLP	8690 12399 3620 3575	0.62	0.41	0.29	0.22	0.50	0.44

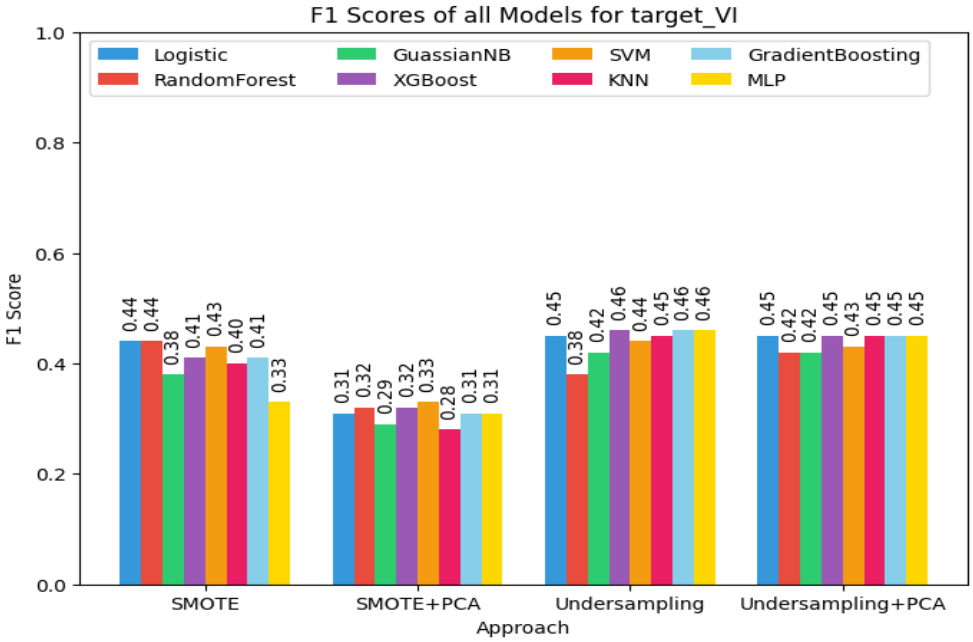
UNDERSAMPLING + PCA:





Model Name	**Confusion Matrix** **TN FP** **FN TP**	**AUC ROC**	**Max F1-Score**	**PRAUC**	**Precision**	**Recall**	**Accuracy**
Logistic Regression	10456 10903 2008 5187	0.63	0.45	0.33	0.32	0.72	0.55
Gaussian Naïve Bayes	12031 9328 2785 4410	0.62	0.45	0.32	0.32	0.61	0.58
KNN	10443 10916 2433 4762	0.60	0.44	0.31	0.30	0.66	0.53
Random Forest	8458 12901 1379 5816	0.62	0.45	0.32	0.31	0.81	0.50
SVM	11508 9851 2508 4687	0.62	0.45	0.32	0.32	0.65	0.57
XGBoost	8868 12491 1497 5698	0.62	0.45	0.32	0.31	0.79	0.51
Gradient Boosting	8707 12652 1439 5756	0.62	0.45	0.32	0.31	0.80	0.51
MLP	9043 12316 1511 5684	0.63	0.45	0.33	0.32	0.79	0.52


Accuracy Scores of all Models for target_VI
Logistic
GuassianNB
KNN
RandomForest
SVM
XGBoost
GradientBoosting
MLP
Accuracy
Approach
SMOTE
SMOTE+PCA
Undersampling
Undersampling+PCA
0.57 0.54 0.57 0.39 0.57 0.26 0.29 0.67
0.43 0.40 0.54 0.44 0.43 0.51 0.44 0.44
0.57 0.63 0.58 0.53 0.57 0.55 0.54 0.54
0.55 0.58 0.53 0.50 0.57 0.51 0.51 0.52


F1 Scores of all Models for target_VI
Logistic
RandomForest
GuassianNB
XGBoost
SVM
KNN
GradientBoosting
MLP
F1 Score
Approach
SMOTE
SMOTE+PCA
Undersampling
Undersampling+PCA
0.44 0.44 0.38 0.41 0.43 0.40 0.41 0.33
0.31 0.32 0.29 0.32 0.33 0.28 0.31 0.31
0.45 0.38 0.42 0.46 0.44 0.45 0.46 0.46
0.45 0.42 0.42 0.45 0.43 0.45 0.45 0.45