Reducing 30-Day Readmissions in Diabetic Patients

Predictive Insights into Smarter Discharge Decisions

Mohamed Thalha Ahamed Ali

301418071

Centennial College

Business Analytics and Insights

David Parent, Bilal Hasanzadah

August 13, 2025

Table of Contents

Executive Summary

0.1 Executive Introduction

This project addresses the challenge of predicting 30-day hospital readmissions among diabetic patients by leveraging a cleaned dataset of 66,222 encounters from 130 U.S. hospitals spanning 1999–2008. Early readmissions represent a significant concern for healthcare providers, as they negatively impact patient recovery, increase operational costs, and influence hospital quality ratings.

The project applies a data-driven approach to identify patients at high risk of readmission before discharge, enabling proactive interventions that can improve outcomes and reduce avoidable costs.

0.2 Executive Objective

The primary objective was to develop a predictive model that accurately flags patients at risk of 30-day readmission, with a strong emphasis on maximizing recall to ensure high-risk patients are not missed.

Given the need for clinical adoption, interpretability was also prioritized so that the reasoning behind predictions could be easily understood by healthcare professionals and embedded into discharge workflows.

0.3 Executive Model Description

Several predictive modelling approaches were tested, including Decision Trees, Random Forests (standard and balanced), XGBoost, Logistic Regression, Neural Networks, and AdaBoost.

After evaluating performance on balanced datasets, the Decision Tree model with Random Undersampling emerged as the optimal choice, achieving:

Recall: ~68% (highest among all tested models)

ROC AUC: ~0.688

The Decision Tree's strength lies in its balance between predictive performance and transparency. It's simple, rule-based logic enables direct interpretation by clinical teams, making it suitable for integration into real-world hospital discharge processes.

Feature importance analysis using SHAP values confirmed that the model leverages clinically relevant variables, including:

- Number of prior inpatient visits (strongest predictor)
- Discharge disposition
- Interaction between age and number of diagnoses
- Diabetes medication status
- Length of stay

These drivers directly inform targeted clinical protocols for at-risk patients.

0.4 Executive Recommendations

It is recommended to Implement the Decision Tree model within the hospital EMR to provide real-time readmission risk scores during discharge planning.

Adopt four targeted clinical protocols derived from model insights:

- Enhanced Discharge Plan
- Structured Handover for Facility Transfers
- Diabetes Care Optimization
- Early Instability Intervention

Support clinical protocols with operational actions, including EMR-integrated risk alerts, automatic follow-up scheduling, and compliance dashboards.

Validate the model prospectively using current hospital data to confirm performance within the local patient population.

Enhance the model over time by incorporating social determinants of health, integrating post-discharge outcomes, and expanding the framework to other chronic conditions such as heart failure and COPD.

Develop user-friendly visualization tools to explain model outputs and support clinician engagement.

Introduction

1.0 Background

The dataset used in this project originates from a large-scale collection of diabetic patient encounters recorded between 1999 and 2008 across 130 hospitals and integrated delivery networks in the United States.

Following extensive cleaning and preparation, the working dataset comprises 66,222 encounters with 45 features capturing demographic, clinical, and administrative details relevant to hospital readmissions.

These features include patient demographics, admission and discharge details, prior inpatient and outpatient visits, laboratory tests, procedures, diagnoses, medications, and key lab results such as glucose serum and A1C levels.

2.0 Problem Statement

Hospital readmission within 30 days remains a critical challenge in managing diabetic patients, resulting in increased healthcare utilization, elevated costs, and potential penalties linked to quality metrics.

Existing predictive tools often underperform in two key areas:

Recall – failing to identify a significant proportion of patients who will actually be readmitted.

Interpretability – producing predictions without clear reasoning, limiting their practical use in clinical workflows.

This project addresses these gaps by developing a predictive model that balances high recall with transparency, enabling clinicians to proactively identify high-risk patients and implement targeted interventions before discharge.

3.0 Objectives & Measurement

The primary objective was to develop a predictive model capable of identifying patients at risk of 30-day readmission, with recall as the primary performance metric. Maximizing recall ensures that the majority of truly high-risk patients are flagged, reducing the likelihood of missed intervention opportunities.

Secondary evaluation metrics included:

- ROC AUC (discrimination ability)
- Accuracy
- Precision
- F1-score (balance of precision and recall)

A **stratified 70/30 train–test split** was used to maintain representative proportions of the readmitted and non-readmitted classes across both sets, ensuring fair and consistent evaluation.

4.0 Data Set Introduction

The dataset for this project is the *"Diabetes 130-US hospitals for years 1999–2008"* dataset from the UCI Machine Learning Repository. It consists of electronic health records for diabetic patients collected over a 10-year period from 130 U.S. hospitals and integrated delivery networks.

The raw dataset contained 101,766 hospital encounter records, with variables covering:

- Patient demographics
- Admission and discharge information
- Prior inpatient, outpatient, and emergency visits

- Laboratory tests and clinical procedures performed
- Diagnoses (ICD-9 codes)
- Medications prescribed
- Clinical lab results such as glucose serum and A1C levels

Three potential target variables relating to patient readmissions were initially considered:

- Readmitted within 30 days (<30)
- Readmitted after 30 days (>30)
- No readmission (No)

For this project, the focus was readmission within 30 days versus no readmission, as this is the most relevant measure for hospital quality reporting and intervention planning.

Encounters classified as readmitted after 30 days (>30) were excluded from the modelling dataset.

Following extensive exploratory data analysis (EDA), data cleaning, grouping of diagnosis and discharge variables, feature engineering, and imputation of missing values, the final dataset contained 66,222 patient encounters with 45 variables.

This dataset formed the basis for all predictive modelling and interpretability analysis in this project.

5.0 Initial Exclusions

To accurately align the dataset with the project's selected target (readmission within 30 days vs. no readmission/late readmission), specific exclusions were applied:

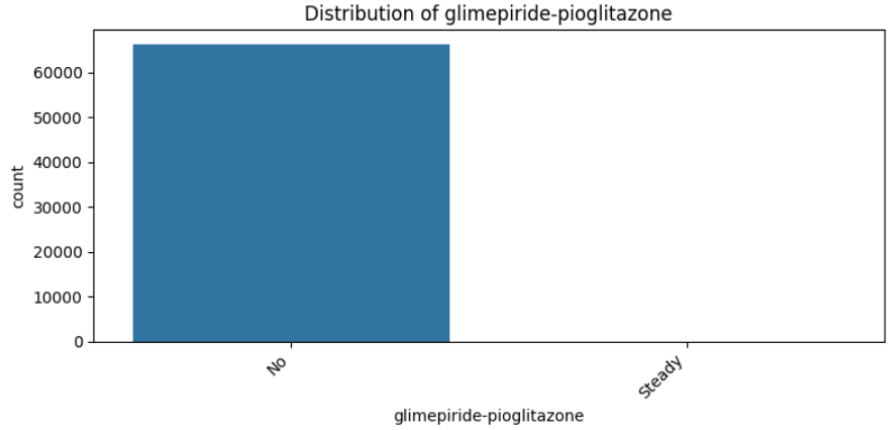
- **Excluded records with Readmission >30 days**

  Records indicating patient readmissions explicitly after 30 days (">30") were removed, as the analysis specifically focuses on identifying and predicting early readmissions within the critical 30-day period post-discharge.

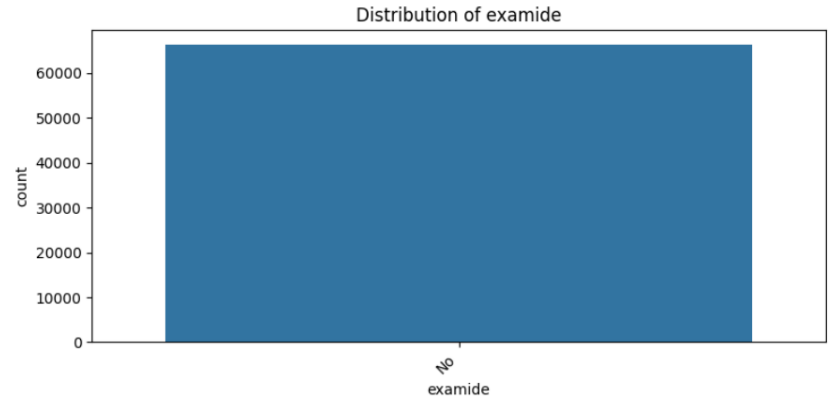
- **Dropped Identifier and Irrelevant Columns**

  Removed columns encounter_id and patient_nbr, as they serve as unique identifiers without predictive significance. Additionally, constant-value columns ('examide', 'acetohexamide', 'glimepiride-pioglitazone') were dropped because they provided no variance or predictive value.

```
Distribution for 'glimepiride-pioglitazone':
                        Count  Percent
glimepiride-pioglitazone
No                      66221    100.0
Steady                      0      0.0
```

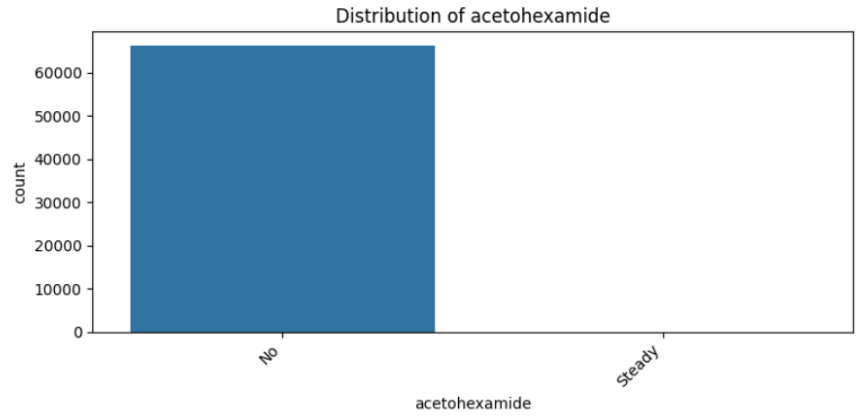



```
Distribution for 'examide':
         Count  Percent
examide
No       66221    100.0
```



Distribution for 'acetohexamide':

acetohexamide	Count	Percent
No	66221	100.0
Steady	0	0.0



6.0 Initial Data Cleansing or Preparation

After applying the initial exclusions to align with the chosen readmission target, several additional data cleansing steps were conducted to enhance dataset quality and predictive power:

6.1 Handling Missing Values:

	column	missing_count	missing_percent
0	encounter_id	0	0.00
1	patient_nbr	0	0.00
2	race	1735	2.62
3	gender	0	0.00
4	age	0	0.00
5	weight	64534	97.45
6	admission_type_id	0	0.00
7	discharge_disposition_id	0	0.00
8	admission_source_id	0	0.00
9	time_in_hospital	0	0.00
10	payer_code	26428	39.91
11	medical_specialty	31733	47.92
12	num_lab_procedures	0	0.00
13	num_procedures	0	0.00
14	num_medications	0	0.00
15	number_outpatient	0	0.00
16	number_emergency	0	0.00
17	number_inpatient	0	0.00
18	diag_1	17	0.03
19	diag_2	284	0.43
20	diag_3	1085	1.64
21	number_diagnoses	0	0.00
22	max_glu_serum	0	0.00
23	A1Cresult	0	0.00

- Variables with significant missing values such as payer_code (~40% missing) and medical_specialty (~50% missing) were retained and imputed missing with Missing, Race with Mode and weight which has missing percent of 97% was removed.
- Diagnosis-related variables (diag_1, diag_2, diag_3) had missing values imputed with "Unknown".

6.2 Creating Missing Indicators

- For variables where missing values were prevalent and potentially predictive (payer_code, medical_specialty, weight and diagnosis groups), binary missing indicators were created.

6.3 Variable Grouping to Reduce Dimensionality:

- Admission Type (originally 9 levels): grouped into meaningful categories (Emergency, Elective, Newborn, Other).

Before:

admission_type_id	Description
1	Emergency
2	Urgent
3	Elective
4	Newborn
5	Not Available
6	NULL
7	Trauma Center
8	Not Mapped

After:

```
admission_map = {
    1: 'Emergency_Urgent',
    2: 'Emergency_Urgent',
    3: 'Elective',
    4: 'Others',
    5: 'Unknown',
    6: 'Unknown',
    7: 'Others',
    8: 'Unknown'
}
```

- Discharge Disposition (originally 29 levels): grouped into categories such as Home, Transfer/Facility, Expired, and Others.

Before:

discharge_disposition_id	description
1	Discharged to home
2	Discharged/transferred to another short term hospital
3	Discharged/transferred to SNF
4	Discharged/transferred to ICF
5	Discharged/transferred to another type of inpatient care institution
6	Discharged/transferred to home with home health service
7	Left AMA
8	Discharged/transferred to home under care of Home IV provider
9	Admitted as an inpatient to this hospital
10	Neonate discharged to another hospital for neonatal aftercare
11	Expired
12	Still patient or expected to return for outpatient services
13	Hospice / home
14	Hospice / medical facility
15	Discharged/transferred within this institution to Medicare approved swing bed
16	Discharged/transferred/referred another institution for outpatient services
17	Discharged/transferred/referred to this institution for outpatient services
18	NULL
19	Expired at home. Medicaid only, hospice.
20	Expired in a medical facility. Medicaid only, hospice.
21	Expired, place unknown. Medicaid only, hospice.
22	Discharged/transferred to another rehab fac including rehab units of a hospital .
23	Discharged/transferred to a long term care hospital.
24	Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.
25	Not Mapped
26	Unknown/Invalid
30	Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere
27	Discharged/transferred to a federal health care facility.
28	Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
29	Discharged/transferred to a Critical Access Hospital (CAH).

After:

```
discharge_group_map = {
    1: 'Home', 6: 'Home', 8: 'Home',
    2: 'Transfer/Facility', 3: 'Transfer/Facility', 4: 'Transfer/Facility',
    5: 'Transfer/Facility', 15: 'Transfer/Facility', 22: 'Transfer/Facility',
    23: 'Transfer/Facility', 24: 'Transfer/Facility', 27: 'Transfer/Facility',
    11: 'Expired', 13: 'Hospice', 14: 'Hospice',
    19: 'Expired', 20: 'Expired',
    12: 'Outpatient Followup', 16: 'Outpatient Followup',
    17: 'Outpatient Followup', 25: 'Unknown',
    7: 'Left AMA',
    9: 'Other', 10: 'Other', 18: 'Unknown',
    26: 'Unknown', 28: 'Transfer/Facility', 29: 'Transfer/Facility'
}
```

- Admission Source (originally 21 levels): grouped into broader categories (Emergency Room, Referral, Transfer, etc.).

Before:

admission_source_id	description
1	Physician Referral
2	Clinic Referral
3	HMO Referral
4	Transfer from a hospital
5	Transfer from a Skilled Nursing Facility (SNF)
6	Transfer from another health care facility
7	Emergency Room
8	Court/Law Enforcement
9	Not Available
10	Transfer from critial access hospital
11	Normal Delivery
12	Premature Delivery
13	Sick Baby
14	Extramural Birth
15	Not Available
17	NULL
18	Transfer From Another Home Health Agency
19	Readmission to Same Home Health Agency
20	Not Mapped
21	Unknown/Invalid
22	Transfer from hospital inpt/same fac reslt in a sep claim
23	Born inside this hospital
24	Born outside this hospital
25	Transfer from Ambulatory Surgery Center
26	Transfer from Hospice

After:

```
[ ]  admission_source_map = {
        1: 'Referral',
        2: 'Referral',
        3: 'Referral',
        4: 'Transfer',
        5: 'Transfer',
        6: 'Transfer',
        7: 'Emergency',
        8: 'Others',
        9: 'Others',
        10: 'Transfer',
        11: 'Newborn/Birth',
        13: 'Newborn/Birth',
        14: 'Newborn/Birth',
        17: 'Others',
        20: 'Others',
        22: 'Transfer',
        25: 'Transfer',
    }
```

- Medical Specialty (originally 68 levels): grouped into broader clinical categories including Medicine Subspecialty, Surgery, General Practice, and Missing.

Before:

```
📊 Value counts of medical_specialty:
medical_specialty
Unknown                          31733
InternalMedicine                  9912
Family/GeneralPractice            4777
Emergency/Trauma                  4559
Cardiology                        3499
Surgery-General                   2053
Orthopedics                       1073
Nephrology                         947
Orthopedics-Reconstructive         940
Radiologist                        757
Psychiatry                         592
ObstetricsandGynecology            561
Pulmonology                        543
Surgery-Cardiovascular/Thoracic    522
Urology                            500
Surgery-Neuro                      392
Gastroenterology                   350
Surgery-Vascular                   333
PhysicalMedicineandRehabilitation  297
Oncology                           248
Pediatrics                         175
Neurology                          157
Hematology/Oncology                143
Pediatrics-Endocrinology           136
Otolaryngology                      96
Endocrinology                       81
Surgery-Thoracic                    79
Surgery-Cardiovascular              74
Psychology                          66
Pediatrics-CriticalCare             59
Hematology                          58
Podiatry                            52
Gynecology                          48
Radiology                           39
Hospitalist                         38
Surgeon                             37
Surgery-Plastic                     31
Ophthalmology                       27
InfectiousDiseases                  25
SurgicalSpecialty                   24
Osteopath                           21
Obsterics&Gynecology-GynecologicOnco  20
Obstetrics                          17
Anesthesiology-Pediatric            14
Rheumatology                        12
Surgery-Maxillofacial                9
Anesthesiology                       9
Pediatrics-Pulmonology               8
Surgery-Colon&Rectal                 8
PhysicianNotFound                    7
Pathology                            7
```

After:

Group	Specialties Included
Unknown	Unknown, (empty string), (missing/NaN)
General Practice	InternalMedicine, Family/GeneralPractice, Hospitalist, Osteopath, PhysicianNotFound, Resident
Emergency/Trauma	Emergency/Trauma, Pediatrics-EmergencyMedicine
Orthopedics	Orthopedics, Orthopedics-Reconstructive
Surgery	Surgery-General, Surgery-Cardiovascular/Thoracic, Surgery-Vascular, Surgery-Neuro, Surgery-Thoracic, Surgeon, Surgery-Plastic, Surgery-Colon&Rectal, Surgery-Maxillofacial, Surgery-Pediatric, SurgicalSpecialty, Surgery-PlasticwithinHeadandNeck, Proctology
Medicine Subspecialty	Cardiology, Cardiology-Pediatric, Gastroenterology, Nephrology, Pulmonology, Neurology, Hematology, Hematology/Oncology, Oncology, Endocrinology, Endocrinology-Metabolism, InfectiousDiseases, Rheumatology, AllergyandImmunology, Dermatology, Ophthalmology, Otolaryngology, SportsMedicine, Neurophysiology, Perinatology
Obstetrics & Gynecology	ObstetricsandGynecology, Gynecology, Obstetrics, Obsterics&Gynecology-GynecologicOnco
Pediatrics	Pediatrics, Pediatrics-Endocrinology, Pediatrics-CriticalCare, Pediatrics-Pulmonology, Pediatrics-Neurology, Pediatrics-Hematology-Oncology, Pediatrics-AllergyandImmunology, Pediatrics-InfectiousDiseases

Psychiatry/Psychology	Psychiatry, Psychiatry-Child/Adolescent, Psychiatry-Addictive, Psychology
Anesthesiology/Pain	Anesthesiology, Anesthesiology-Pediatric
Radiology/Pathology	Radiologist, Radiology, Pathology, DCPTEAM
Rehabilitation	PhysicalMedicineandRehabilitation
Other	Dentistry, Podiatry, Speech, OutreachServices, and any other specialty not explicitly listed above

- Diagnosis Codes (diag_1, diag_2, diag_3): grouped into clinically relevant diagnosis groups (e.g., Diabetes, Circulatory, Respiratory).

Before:

This had unique codes for each diagnosis.

After:

Grouped based on Diagnosis code.

- List of ICD-9 codes 001–139: infectious and parasitic diseases
- List of ICD-9 codes 140–239: neoplasms
- List of ICD-9 codes 240–279: endocrine, nutritional and metabolic diseases, and immunity disorders
- List of ICD-9 codes 280–289: diseases of the blood and blood-forming organs
- List of ICD-9 codes 290–319: mental disorders
- List of ICD-9 codes 320–389: diseases of the nervous system and sense organs
- List of ICD-9 codes 390–459: diseases of the circulatory system
- List of ICD-9 codes 460–519: diseases of the respiratory system
- List of ICD-9 codes 520–579: diseases of the digestive system
- List of ICD-9 codes 580–629: diseases of the genitourinary system
- List of ICD-9 codes 630–679: complications of pregnancy, childbirth, and the puerperium
- List of ICD-9 codes 680–709: diseases of the skin and subcutaneous tissue
- List of ICD-9 codes 710–739: diseases of the musculoskeletal system and connective tissue
- List of ICD-9 codes 740–759: congenital anomalies
- List of ICD-9 codes 760–779: certain conditions originating in the perinatal period
- List of ICD-9 codes 780–799: symptoms, signs, and ill-defined conditions
- List of ICD-9 codes 800–999: injury and poisoning
- List of ICD-9 codes E and V codes: external causes of injury and supplemental classification

https://en.wikipedia.org/wiki/List_of_ICD-9_codes

7.0 Data Dictionary

1. Demographics:

- race: Patient race (e.g., Caucasian, African American)
- gender: Patient gender (Male, Female, Unknown/Invalid)
- age: Patient age in 10-year intervals (e.g., "[60-70)")

2. Admission and Discharge Information:

- admission_type_grouped: Type of admission (Emergency, Elective, Newborn, Other)
- discharge_group: Discharge disposition (Home, Transfer/Facility, Expired, Other)
- admission_source_grouped: Source of admission (ER, Referral, Transfer, Other)
- time_in_hospital: Length of stay (days)

3. Visit History:

- number_inpatient: Prior inpatient visits
- number_outpatient: Prior outpatient visits
- number_emergency: Prior emergency visits

4. Clinical and Medical Information:

- medical_specialty_grouped: Physician medical specialty (grouped categories including Medicine Subspecialties, Surgery, General Practice, Missing)
- diag_1_group, diag_2_group, diag_3_group: Primary, secondary, and tertiary diagnosis groups (clinically relevant categories: Diabetes, Circulatory, Respiratory, etc.)
- number_diagnoses: Total number of diagnoses recorded

5. Laboratory and Procedures:

- num_lab_procedures: Number of lab tests performed
- num_procedures: Number of medical procedures performed

- num_medications: Number of distinct medications administered during hospital stay

6. Medication Information:

- Diabetes medications (e.g., metformin, insulin, repaglinide, glipizide, glyburide, etc.): Medication status categorized as "Up," "Down," "Steady," or "No."

7. Laboratory Results:

- max_glu_serum: Results of glucose serum tests (">300," ">200," "Norm," "None")
- A1Cresult: Results of A1C tests (">8," ">7," "Norm," "None")

8. Other Binary and Indicator Variables:

- change: Indicates if there was a change in diabetes medications ("Yes," "No")
- diabetesMed: Indicates if any diabetes medication was prescribed ("Yes," "No")
- Missing indicators: payer_code_missing, medical_specialty_missing, diag_1_missing, diag_2_missing, diag_3_missing, Weight_missing (1 if missing , 0 otherwise)

9. Engineered Interaction Features:

- stay_medication_load: Interaction of length of hospital stay and number of medications
- acute_instability: Interaction of prior inpatient and emergency visits
- age_num_diagnoses: Interaction of numeric age and number of diagnoses
- proc_stay_intensity: Interaction of number of procedures and length of stay

10. Target Variable:

- target (Derived from readmitted): Binary classification identifying readmission within 30 days:
    - 1 = Readmitted within 30 days
    - 0 = Not readmitted or readmitted

8.0 Data Exploration Techniques

An extensive exploratory data analysis (EDA) was conducted on the raw dataset to understand data structure, identify potential predictive variables, and ensure data quality. Key EDA techniques applied included:

8.1. Descriptive Statistics:

- Analyzed numeric variables for central tendency, dispersion, and range.

		No	Yes
time_in_hospital	count	54864.000000	11357.000000
	mean	4.254429	4.768249
	std	2.964964	3.028165
	min	1.000000	1.000000
	25%	2.000000	2.000000
	50%	3.000000	4.000000
	75%	6.000000	6.000000
	max	14.000000	14.000000
num_lab_procedures	count	54864.000000	11357.000000
	mean	42.381598	44.226028

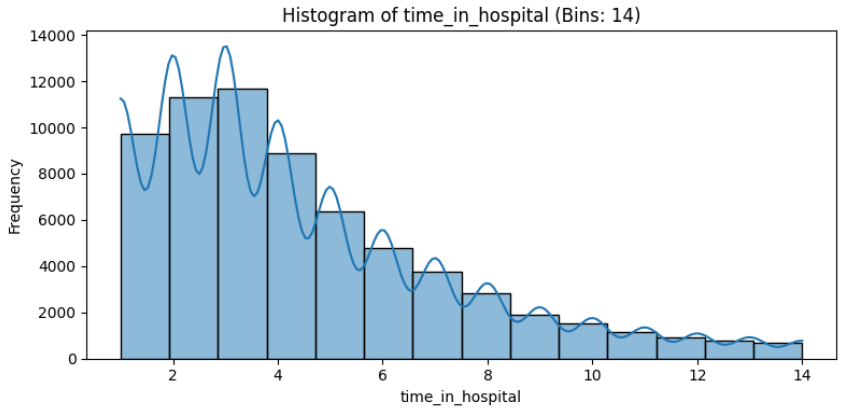
	std	19.796262	19.276087
	min	1.000000	1.000000
	25%	30.000000	33.000000
	50%	44.000000	45.000000
	75%	56.000000	58.000000
	max	126.000000	132.000000
num_procedures	count	54864.000000	11357.000000
	mean	1.410305	1.280884
	std	1.739693	1.635992
	min	0.000000	0.000000
	25%	0.000000	0.000000
	50%	1.000000	1.000000
	75%	2.000000	2.000000
	max	6.000000	6.000000

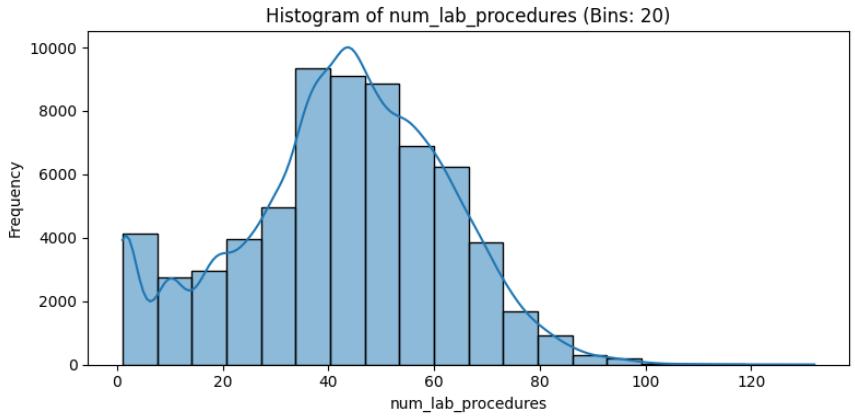
num_medications	count	54864.000000	11357.000000
	mean	15.670367	16.903143
	std	8.427628	8.096696
	min	1.000000	1.000000
	25%	10.000000	11.000000
	50%	14.000000	16.000000
	75%	20.000000	21.000000
	max	79.000000	81.000000
number_outpatient	count	54864.000000	11357.000000
	mean	0.273112	0.436911
	std	1.030704	1.302788
	min	0.000000	0.000000
	25%	0.000000	0.000000
	50%	0.000000	0.000000

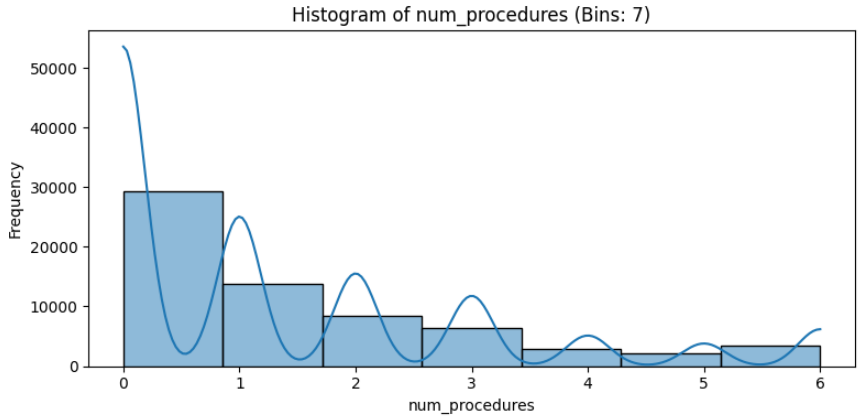
	75%	0.000000	0.000000
	max	36.000000	40.000000
number_emergency	count	54864.000000	11357.000000
	mean	0.109216	0.357313
	std	0.523609	1.370384
	min	0.000000	0.000000
	25%	0.000000	0.000000
	50%	0.000000	0.000000
	75%	0.000000	0.000000
	max	37.000000	64.000000
number_inpatient	count	54864.000000	11357.000000
	mean	0.381963	1.224003
	std	0.864301	1.954577
	min	0.000000	0.000000

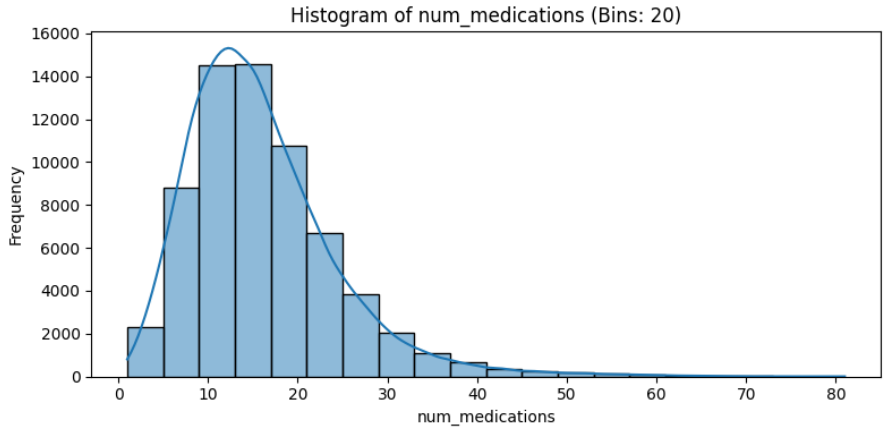
	25%	0.000000	0.000000
	50%	0.000000	0.000000
	75%	0.000000	2.000000
	max	16.000000	21.000000
number_diagnoses	count	54864.000000	11357.000000
	mean	7.221366	7.692789
	std	2.017054	1.773477
	min	1.000000	1.000000
	25%	6.000000	6.000000
	50%	8.000000	9.000000
	75%	9.000000	9.000000
	max	16.000000	16.000000

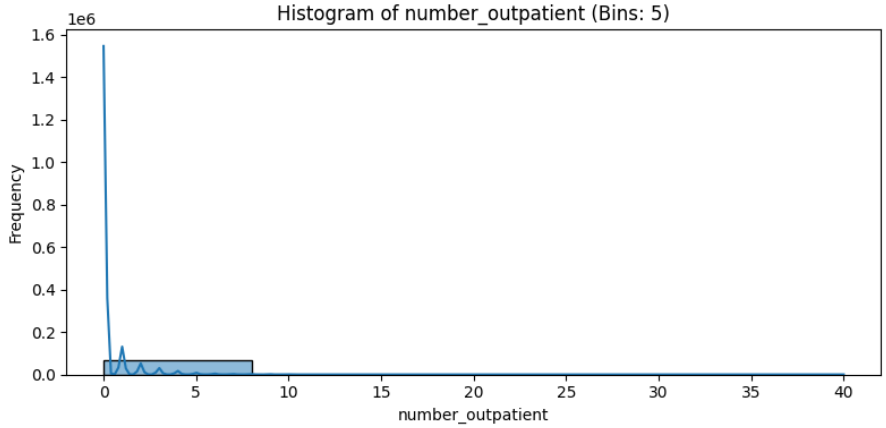
8.2 Distribution of Numeric variables

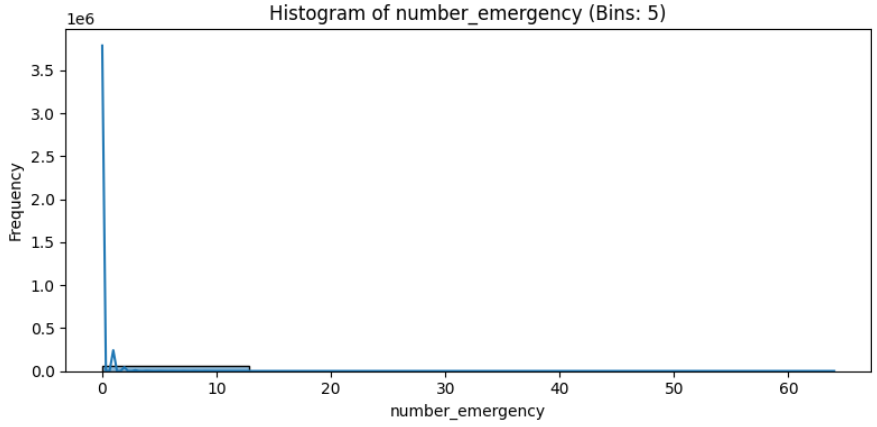


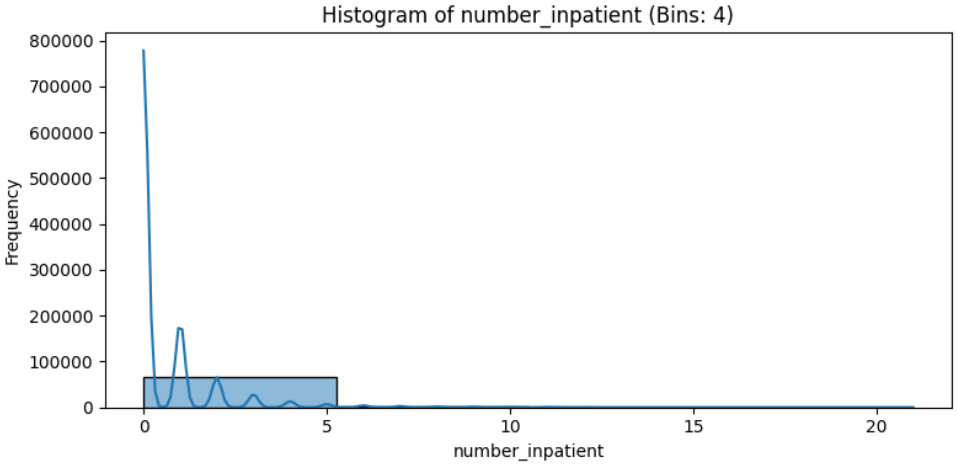


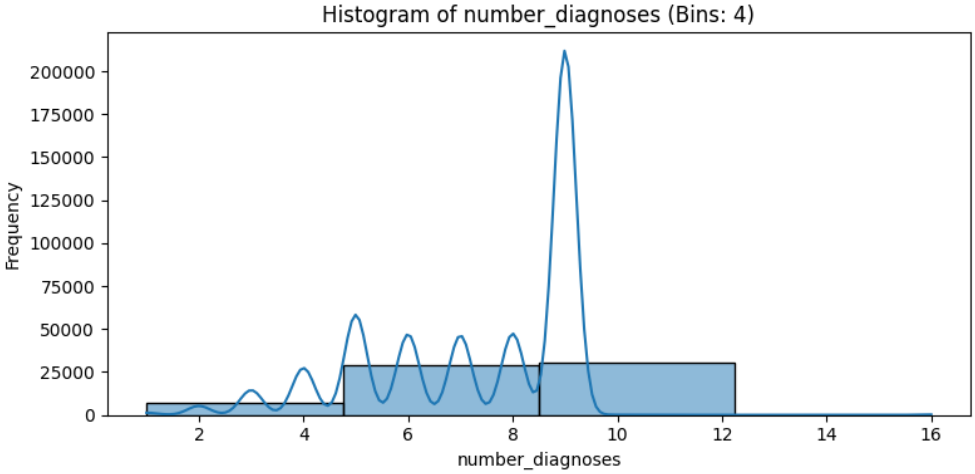



Histogram of num_medications (Bins: 20)
Frequency
num_medications
0
2000
4000
6000
8000
10000
12000
14000
16000
0
10
20
30
40
50
60
70
80

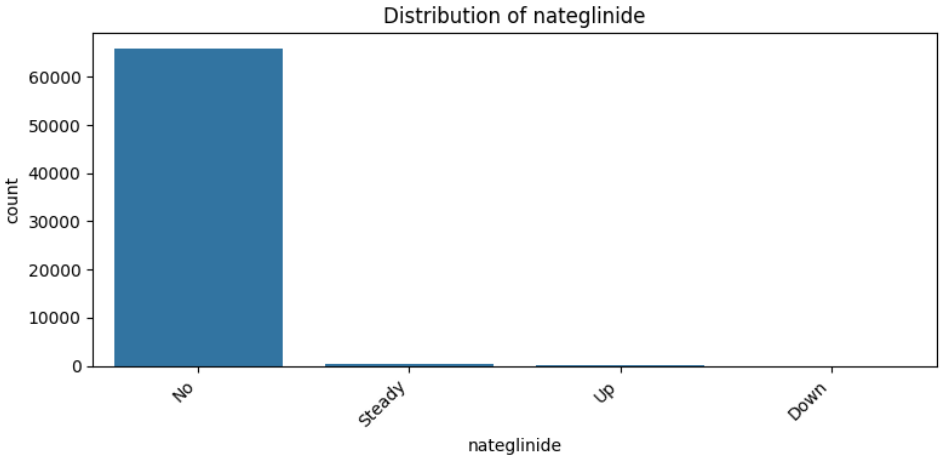

Histogram of number_outpatient (Bins: 5)
1e6
Frequency
number_outpatient
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
5
10
15
20
25
30
35
40

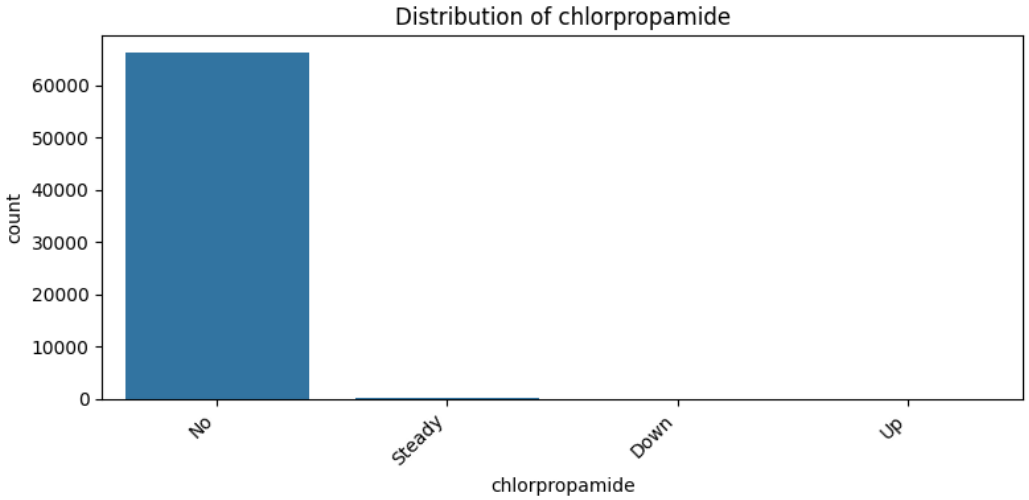

Histogram of number_emergency (Bins: 5)
1e6
Frequency
number_emergency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
10
20
30
40
50
60

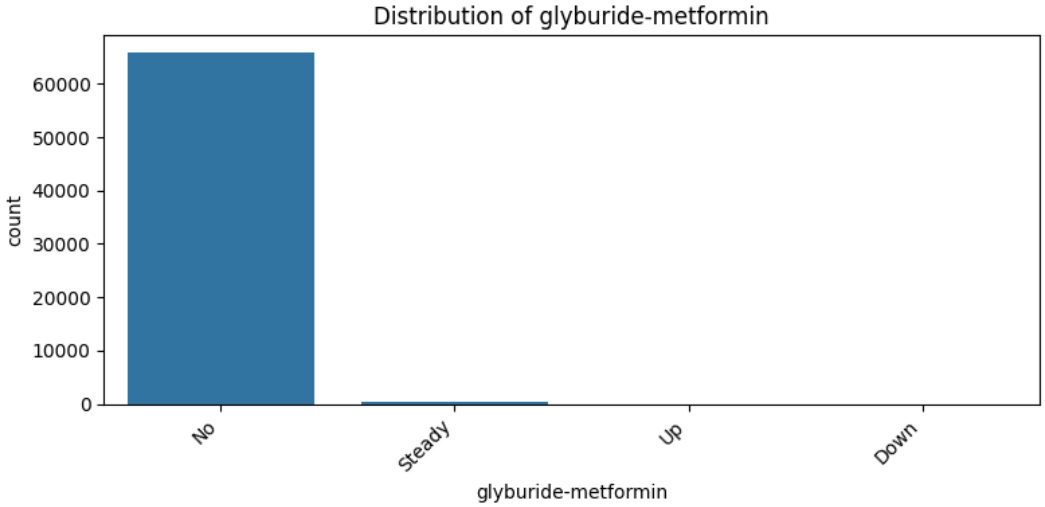


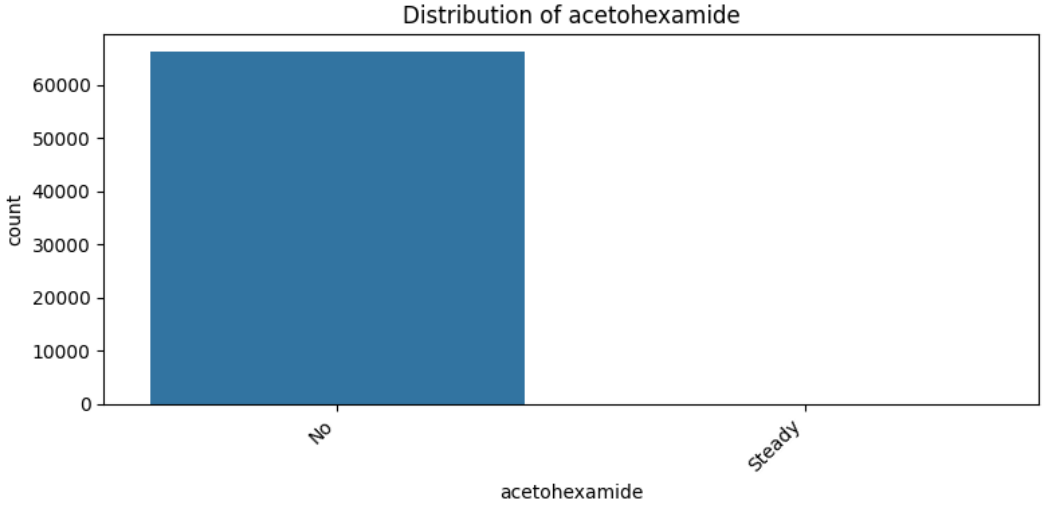


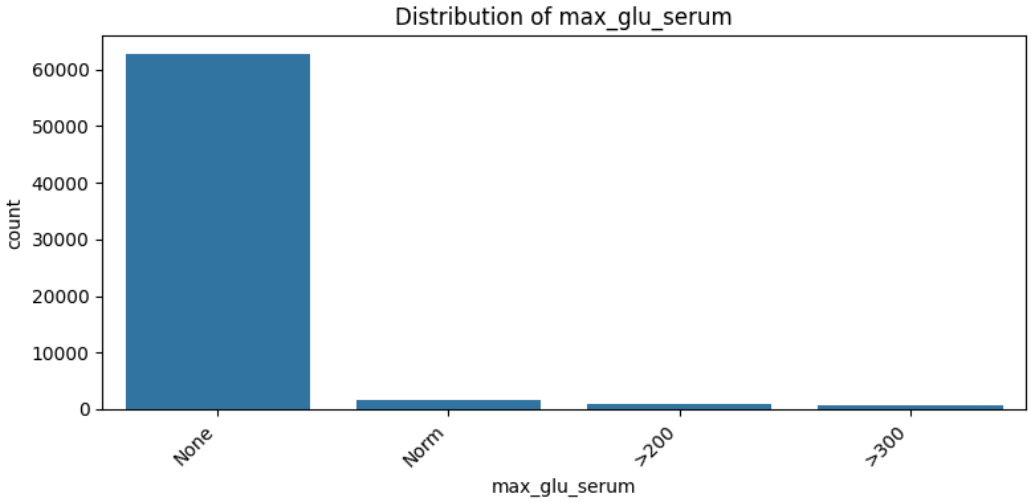
8.3 Distribution of categorical variables

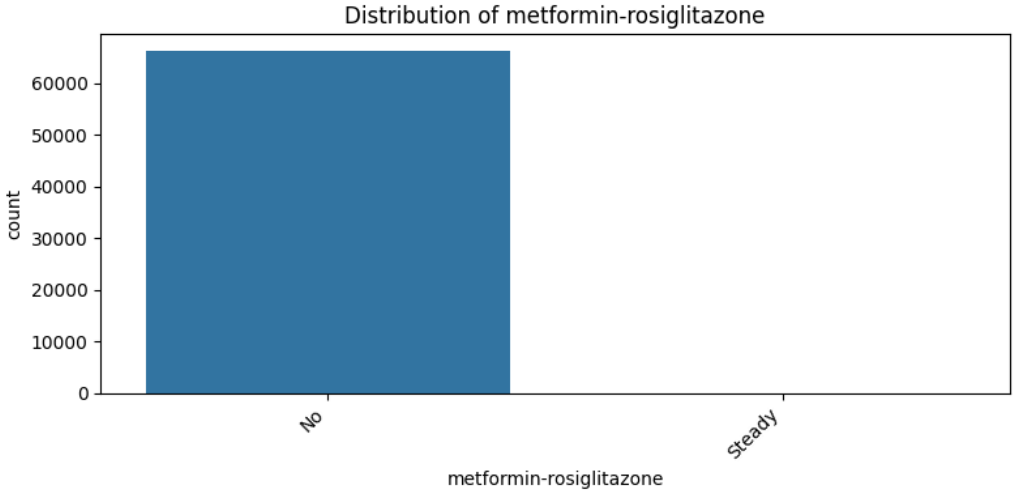


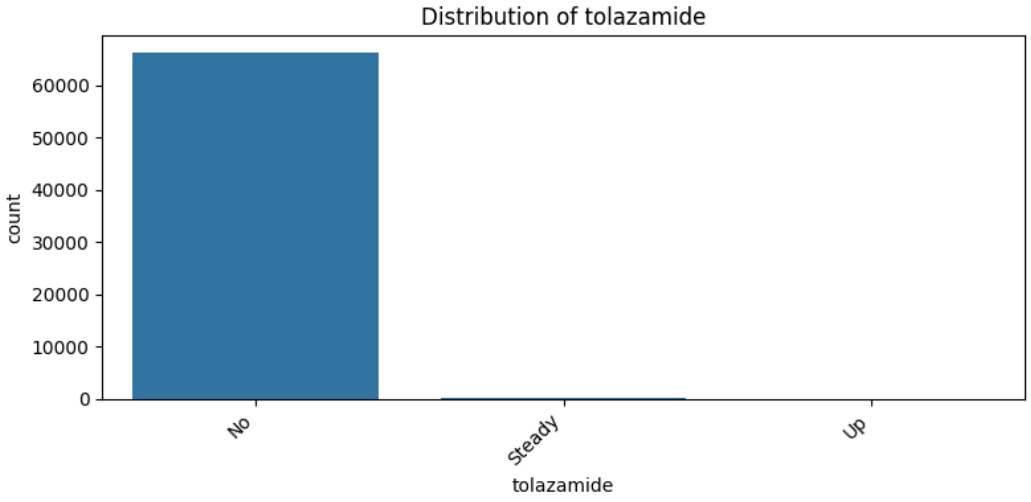

Distribution of chlorpropamide
count
60000
50000
40000
30000
20000
10000
0
No
Steady
Down
Up
chlorpropamide

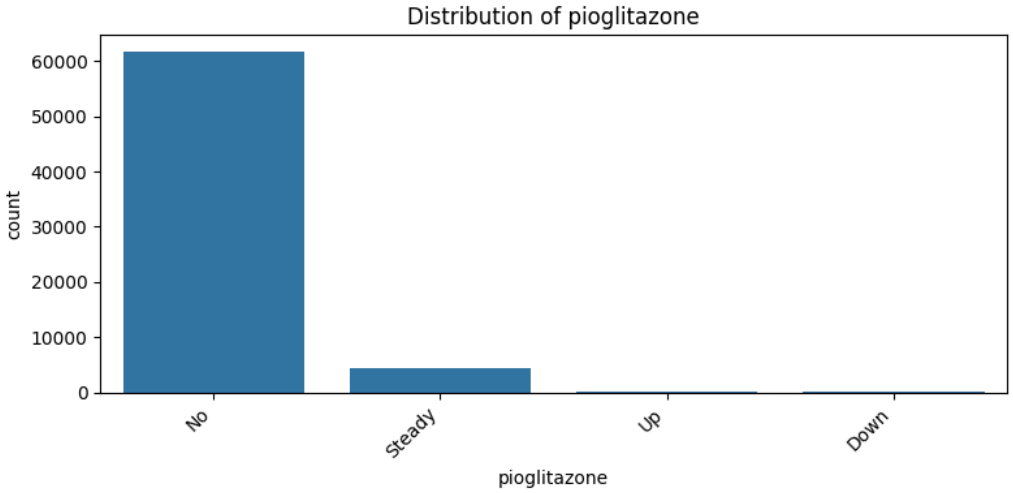

Distribution of glyburide-metformin
count
60000
50000
40000
30000
20000
10000
0
No
Steady
Up
Down
glyburide-metformin

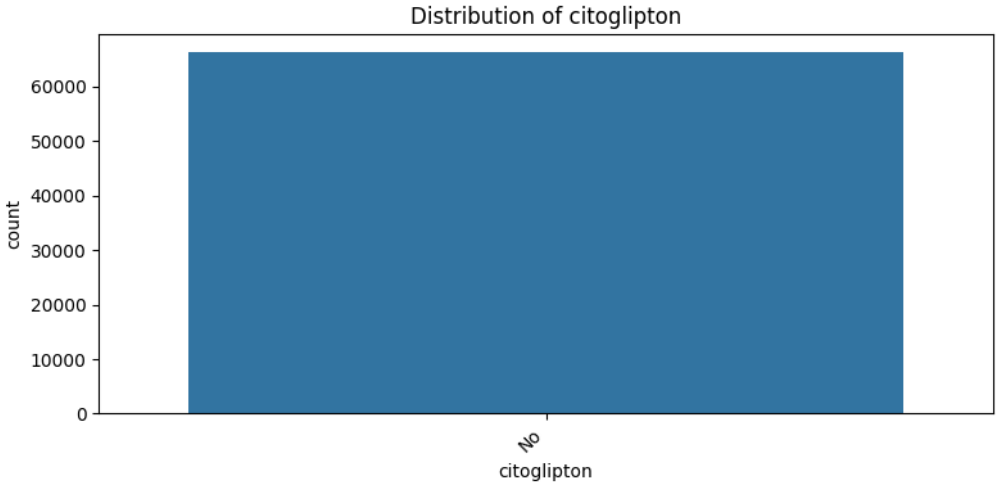

Distribution of acetohexamide
count
60000
50000
40000
30000
20000
10000
0
No
Steady
acetohexamide

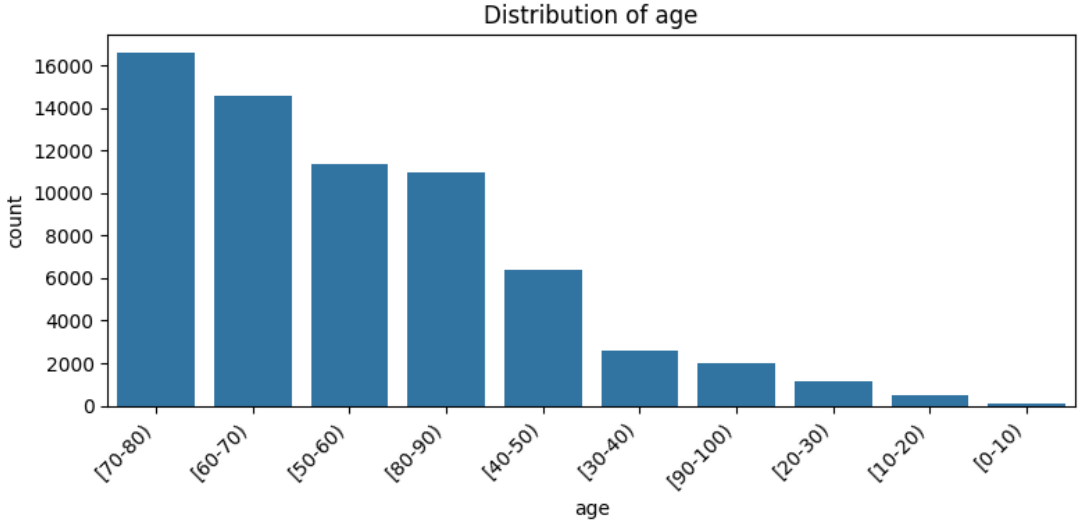

Distribution of max_glu_serum
count
60000
50000
40000
30000
20000
10000
0
None
Norm
>200
>300
max_glu_serum

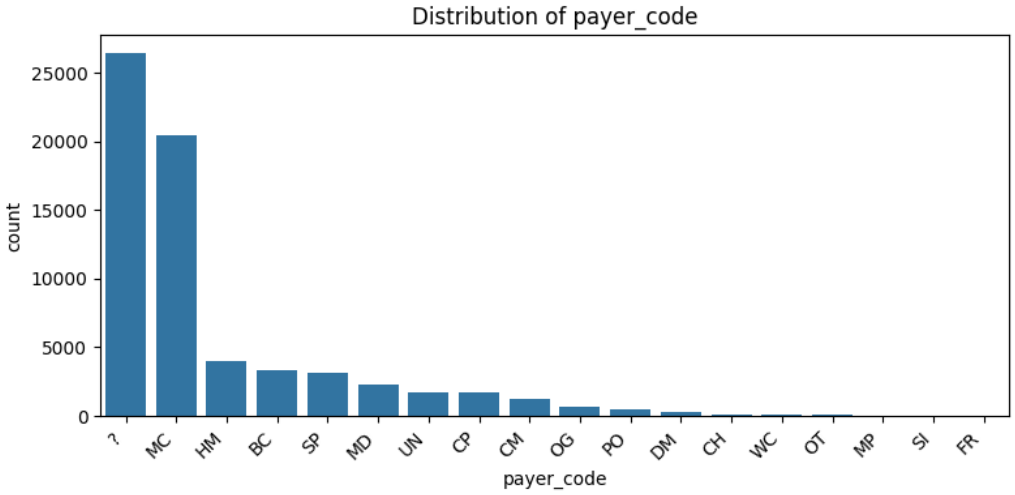

Distribution of metformin-rosiglitazone
count
60000
50000
40000
30000
20000
10000
0
No
Steady
metformin-rosiglitazone

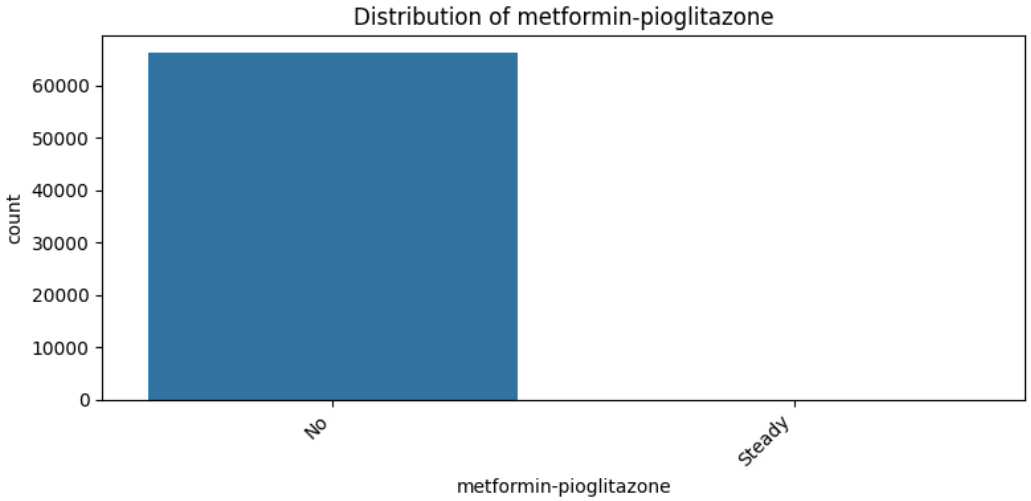

Distribution of tolazamide
count
60000
50000
40000
30000
20000
10000
0
No
Steady
Up
tolazamide

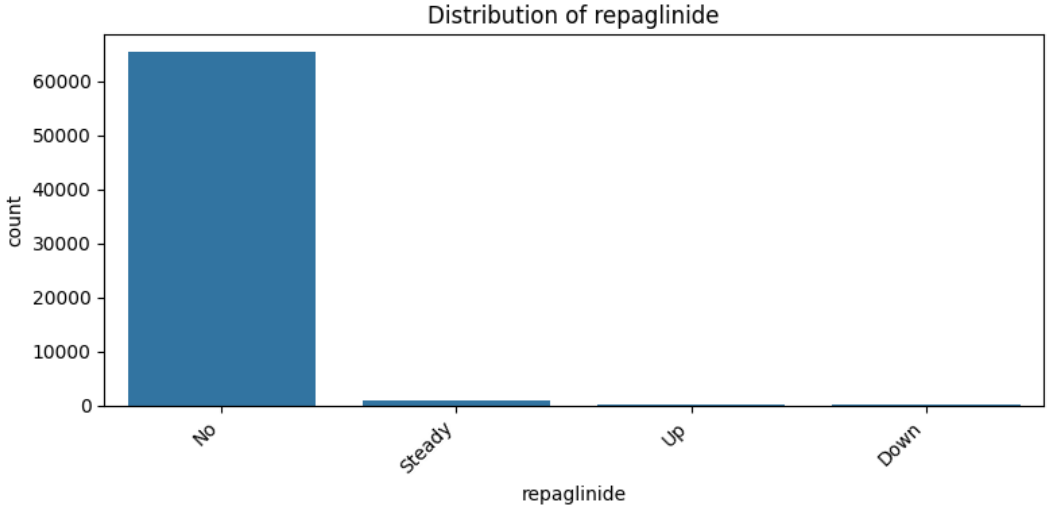

Distribution of pioglitazone
60000
50000
40000
30000
20000
10000
0
count
No
Steady
Up
Down
pioglitazone

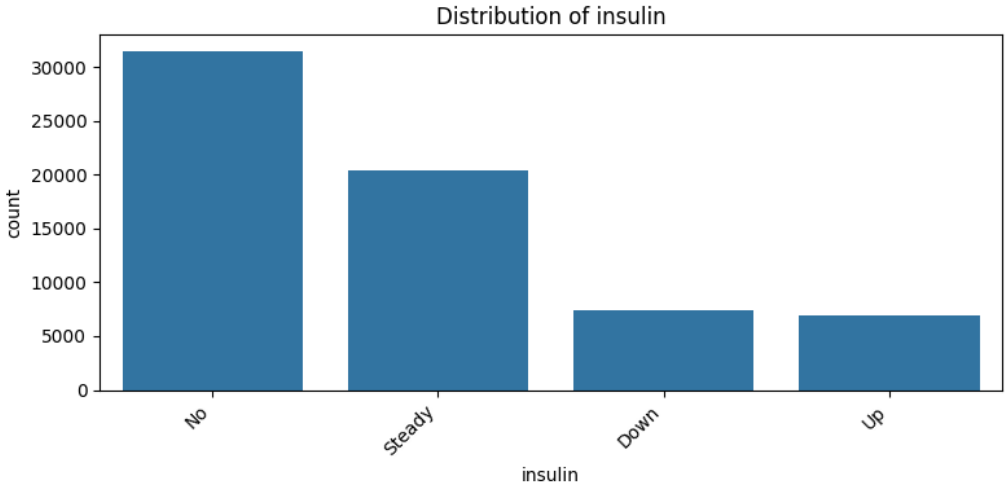

Distribution of citoglipton
60000
50000
40000
30000
20000
10000
0
count
No
citoglipton

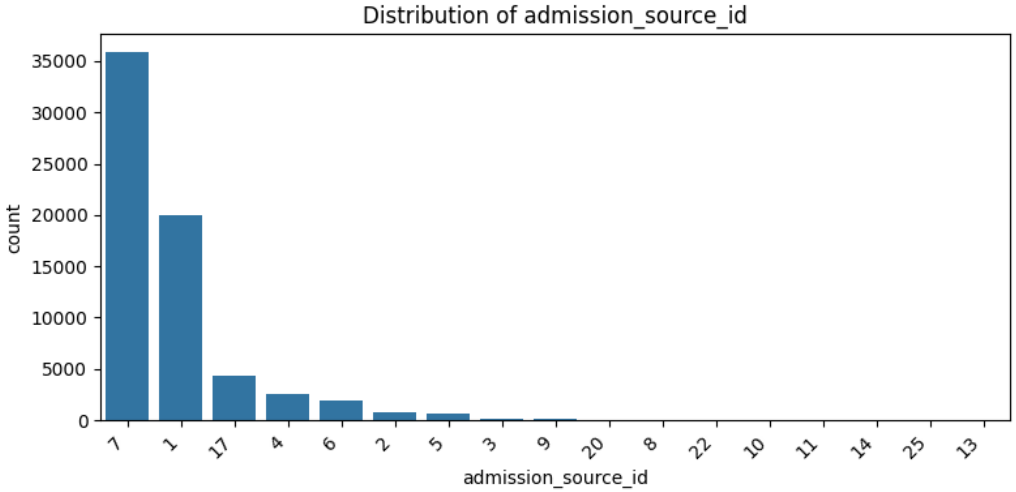

Distribution of age
16000
14000
12000
10000
8000
6000
4000
2000
0
count
[70-80)
[60-70)
[50-60)
[80-90)
[40-50)
[30-40)
[90-100)
[20-30)
[10-20)
[0-10)
age

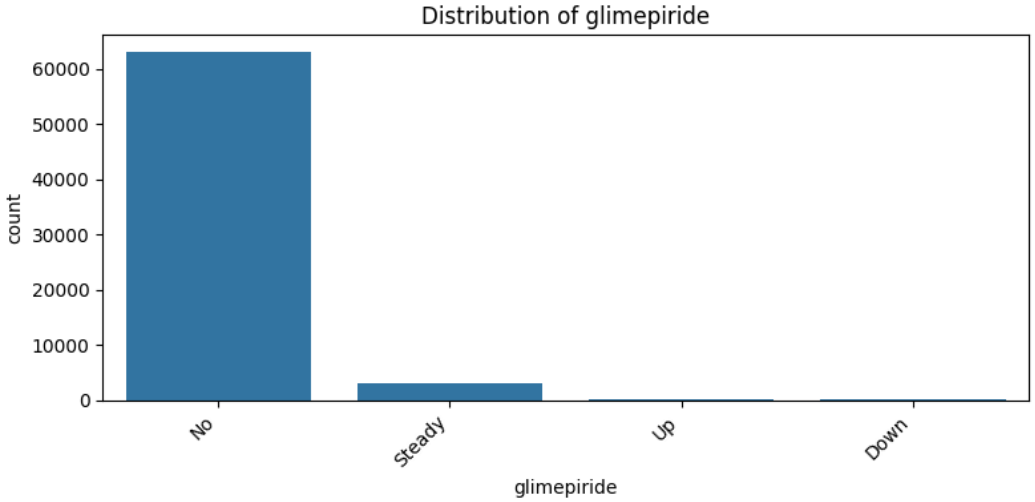

Distribution of payer_code
count
25000
20000
15000
10000
5000
0
?
MC
HM
BC
SP
MD
UN
CP
CM
OG
PO
DM
CH
WC
OT
MP
SI
FR
payer_code

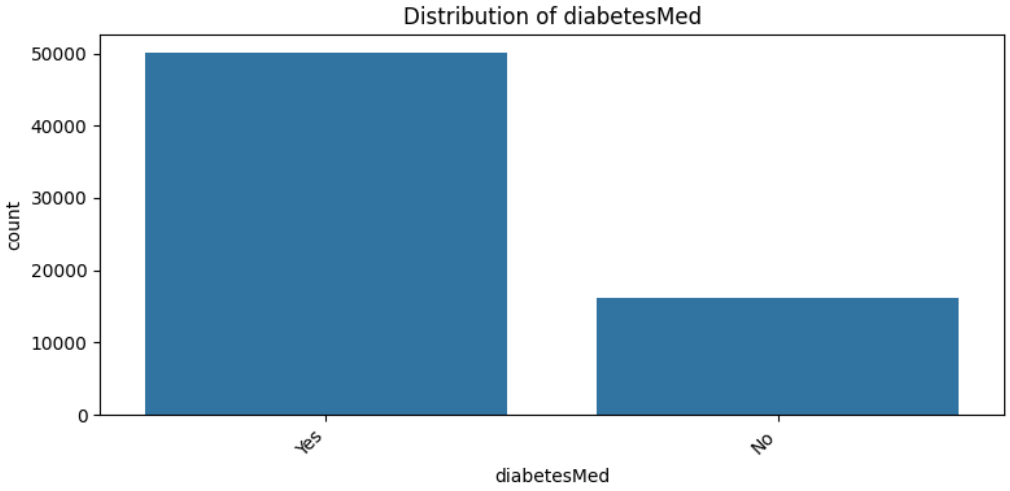

Distribution of metformin-pioglitazone
count
60000
50000
40000
30000
20000
10000
0
No
Steady
metformin-pioglitazone

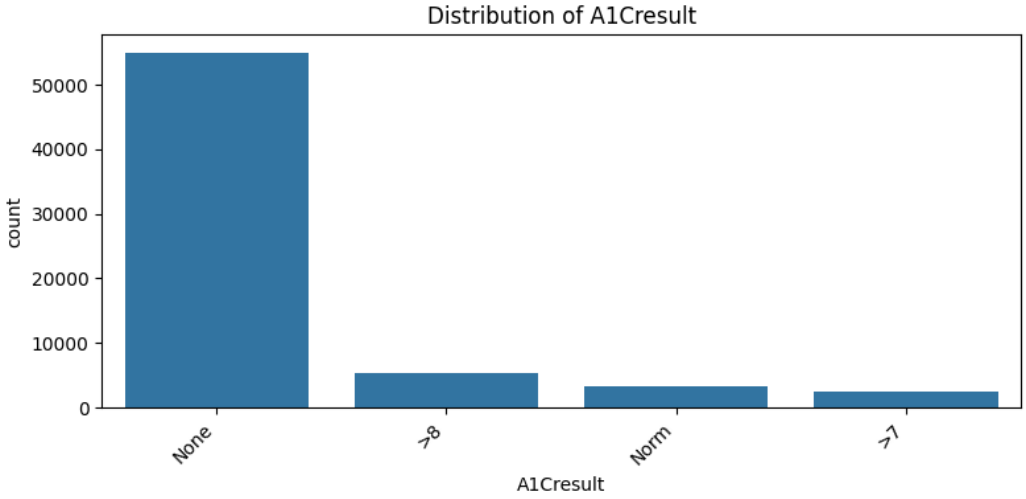

Distribution of repaglinide
count
60000
50000
40000
30000
20000
10000
0
No
Steady
Up
Down
repaglinide

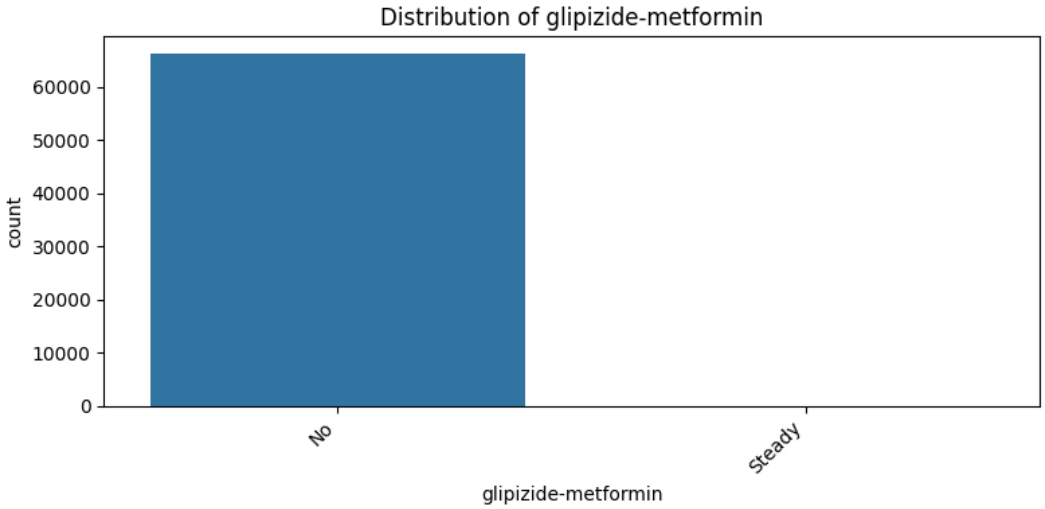

Distribution of insulin
count
30000
25000
20000
15000
10000
5000
0
No
Steady
Down
Up
insulin

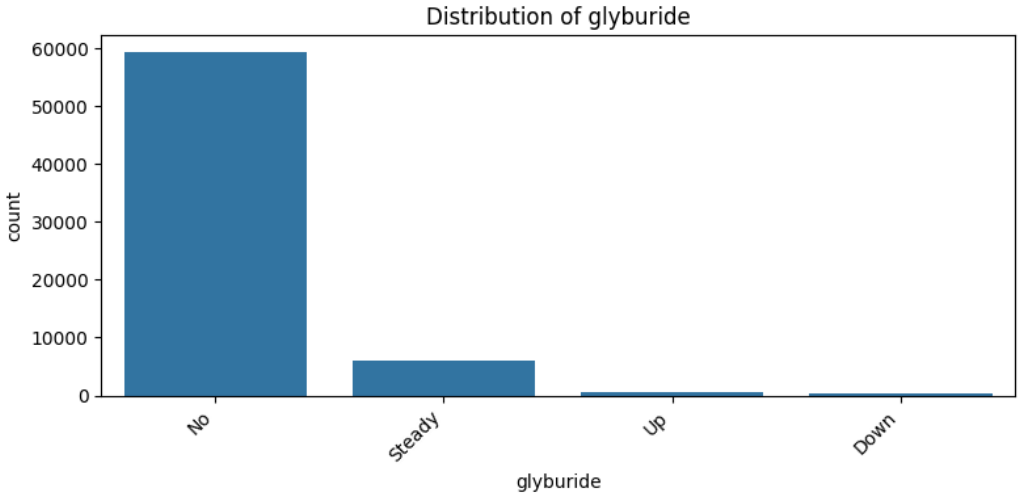

Distribution of admission_source_id
count
35000
30000
25000
20000
15000
10000
5000
0
7
1
17
4
6
2
5
3
9
20
8
22
10
11
14
25
13
admission_source_id

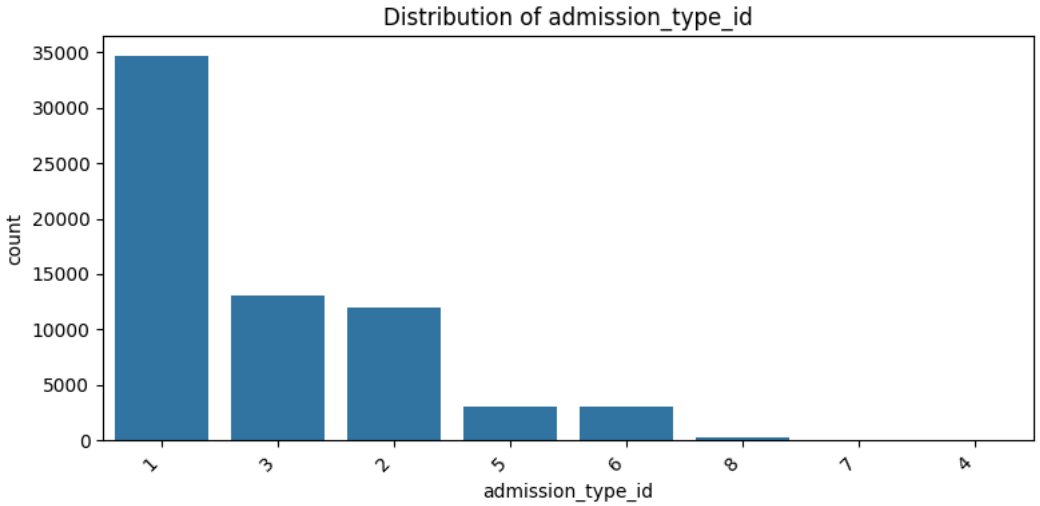

Distribution of glimepiride
count
60000
50000
40000
30000
20000
10000
0
No
Steady
Up
Down
glimepiride

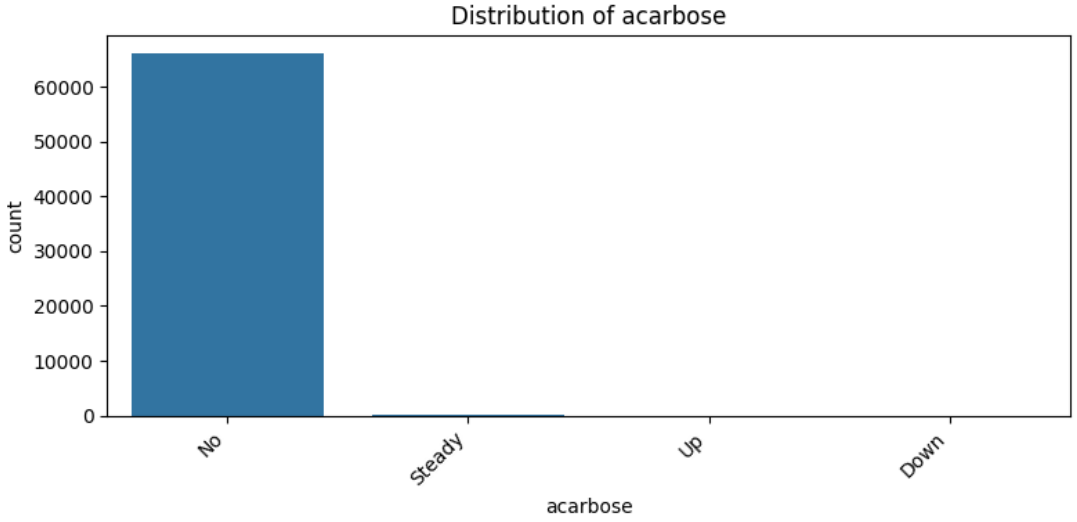

Distribution of diabetesMed
50000
40000
30000
20000
10000
0
count
Yes
No
diabetesMed

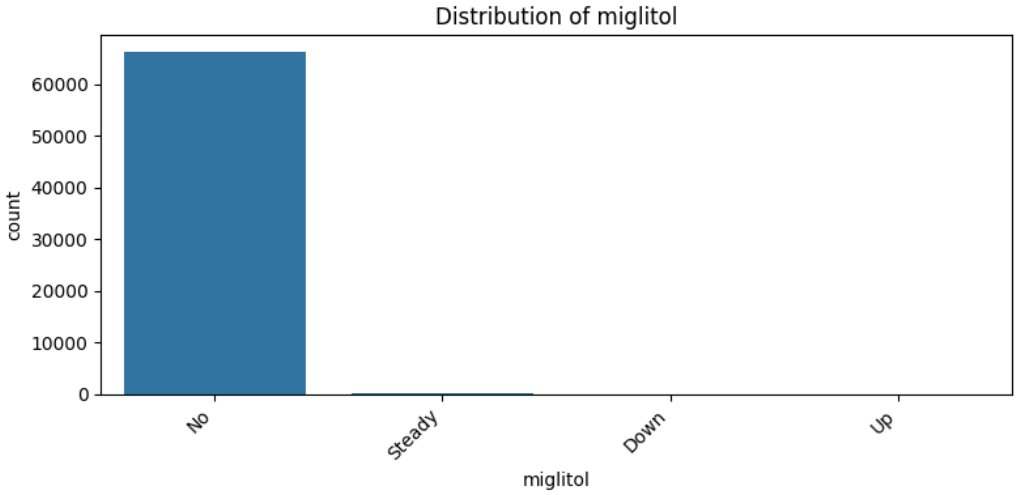

Distribution of A1Cresult
50000
40000
30000
20000
10000
0
count
None
>8
Norm
>7
A1Cresult

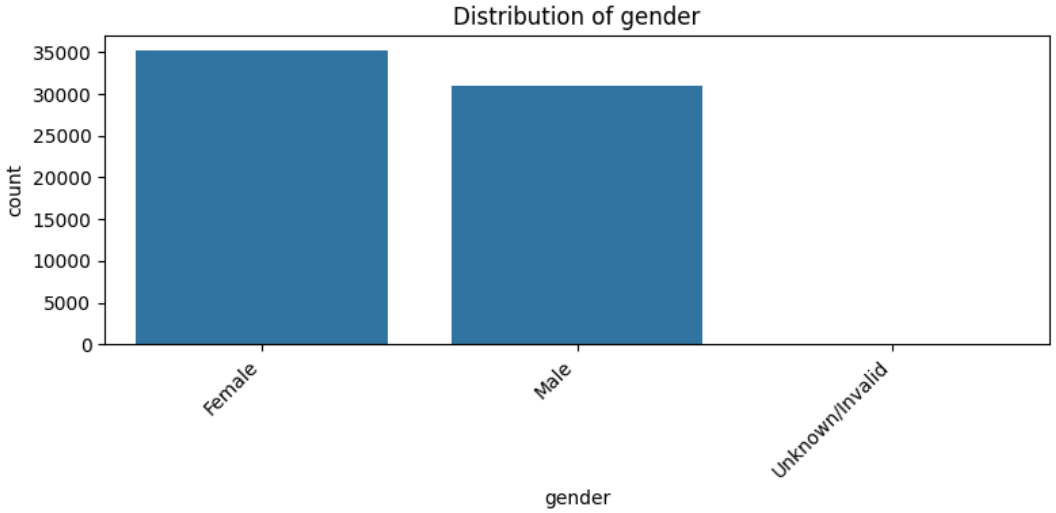

Distribution of glipizide-metformin
60000
50000
40000
30000
20000
10000
0
count
No
Steady
glipizide-metformin

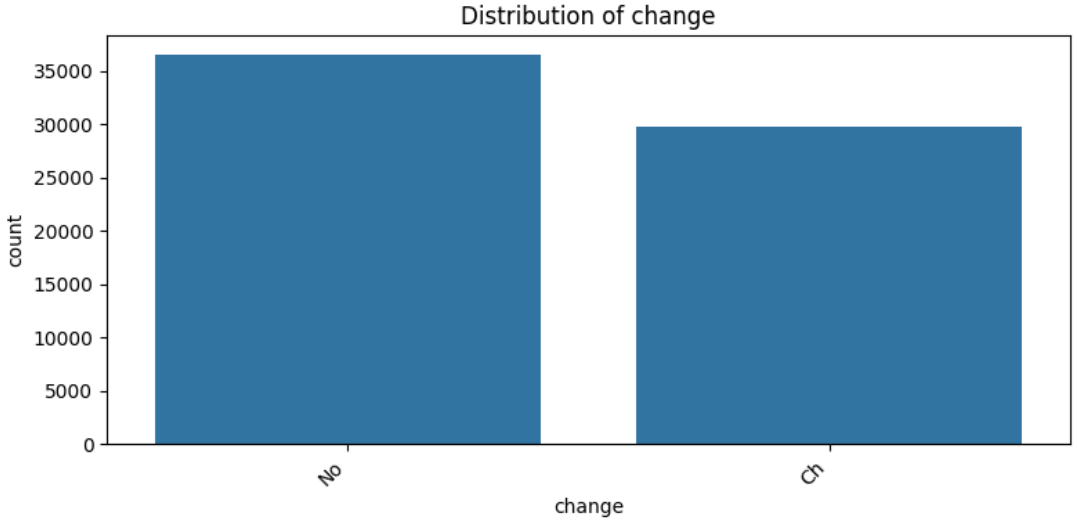

Distribution of glyburide
60000
50000
40000
30000
20000
10000
0
count
No
Steady
Up
Down
glyburide

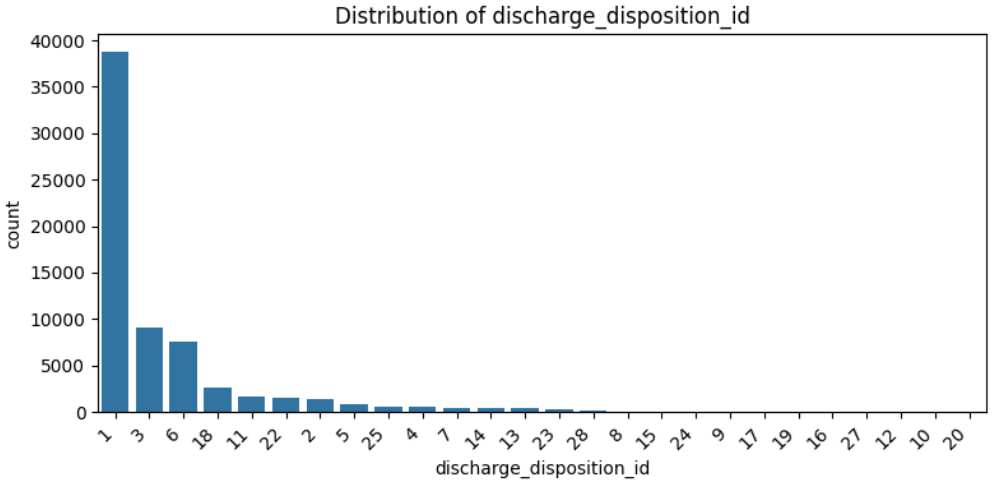

Distribution of admission_type_id
35000
30000
25000
20000
15000
10000
5000
0
count
1
3
2
5
6
8
7
4
admission_type_id

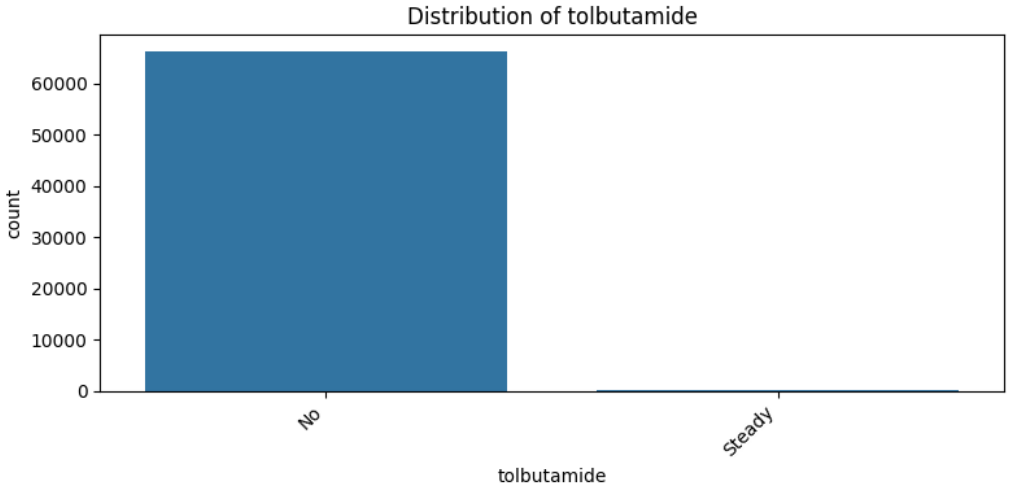

Distribution of acarbose
60000
50000
40000
30000
20000
10000
0
count
No
Steady
Up
Down
acarbose

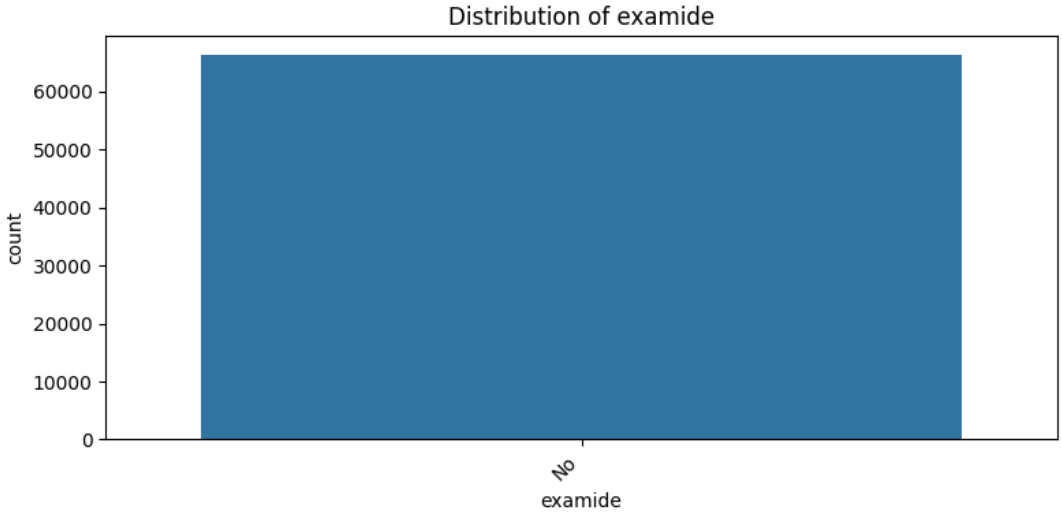

Distribution of miglitol
60000
50000
40000
30000
20000
10000
0
count
No
Steady
Down
Up
miglitol

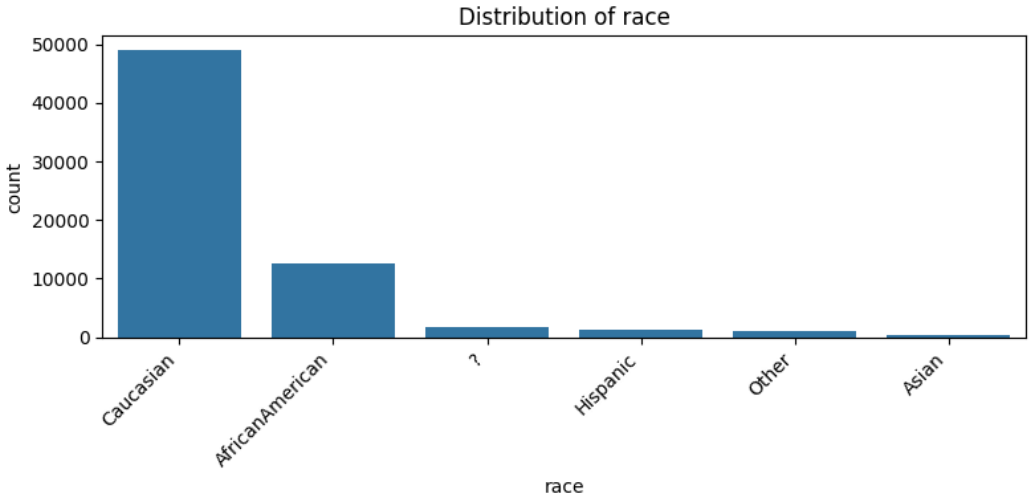

Distribution of gender
35000
30000
25000
20000
15000
10000
5000
0
count
Female
Male
Unknown/Invalid
gender

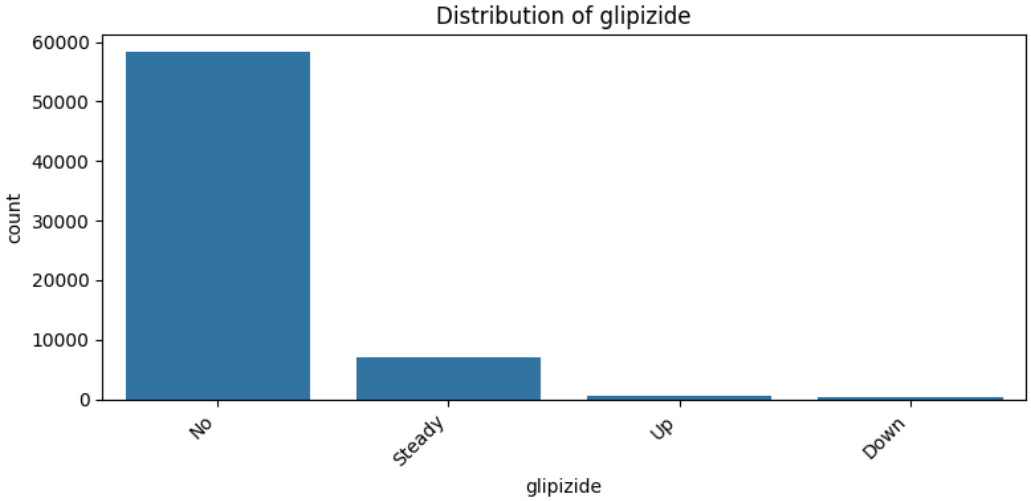

Distribution of change
35000
30000
25000
20000
15000
10000
5000
0
count
No
Ch
change

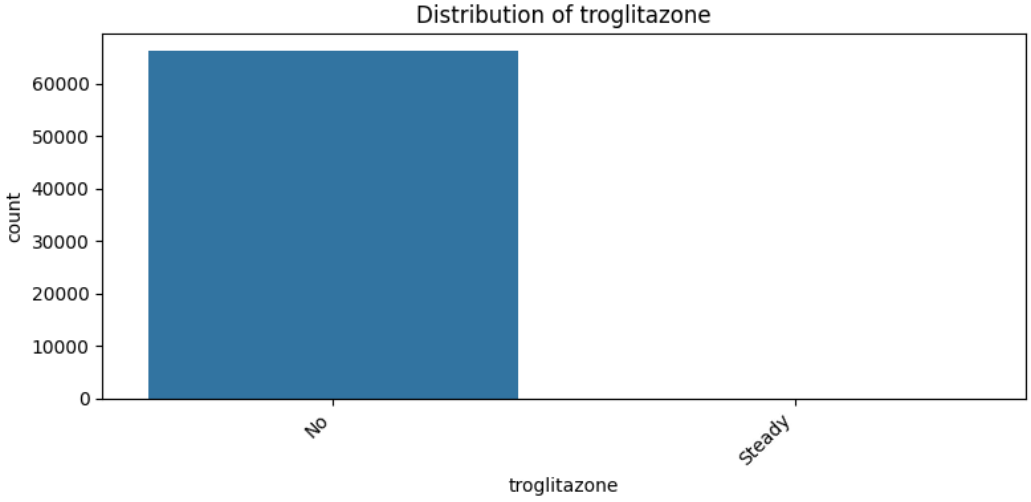

Distribution of discharge_disposition_id
count
discharge_disposition_id
40000
35000
30000
25000
20000
15000
10000
5000
0
1
3
6
18
11
22
2
5
25
4
7
14
13
23
28
8
15
24
9
17
19
16
27
12
10
20

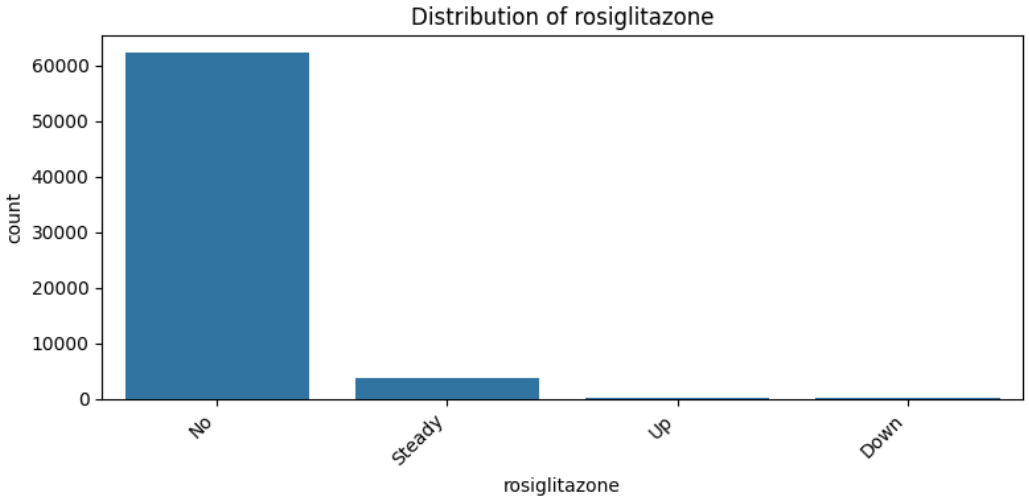

Distribution of tolbutamide
count
tolbutamide
60000
50000
40000
30000
20000
10000
0
No
Steady

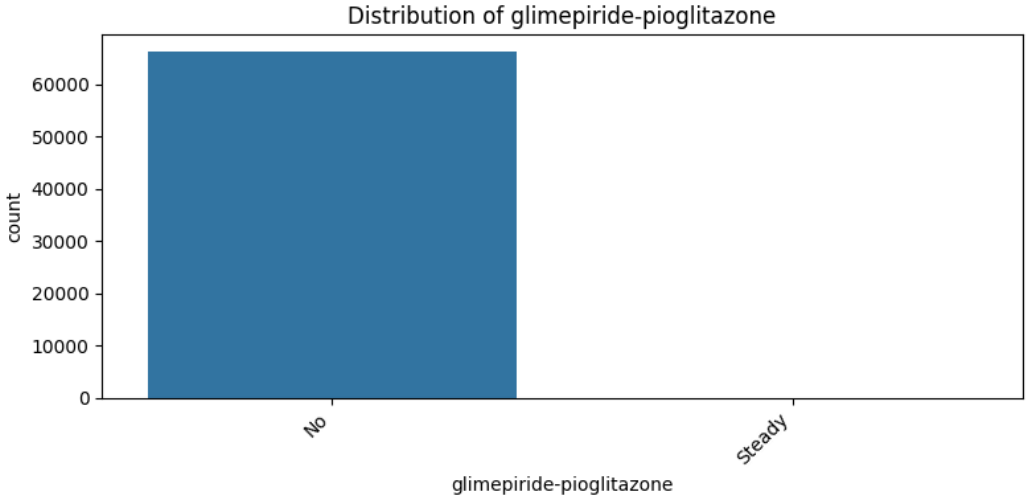

Distribution of examide
count
examide
60000
50000
40000
30000
20000
10000
0
No

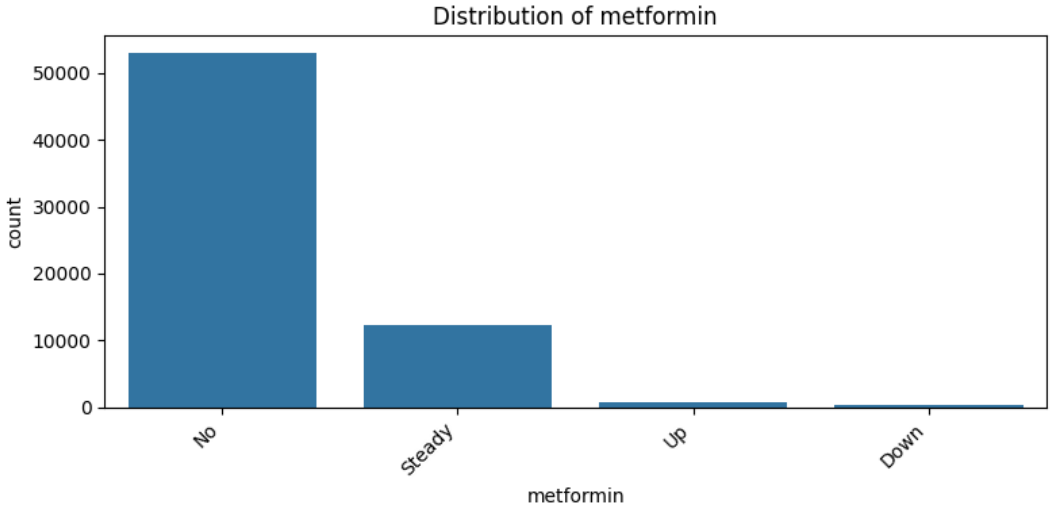

Distribution of race
50000
40000
30000
20000
10000
0
count
Caucasian
AfricanAmerican
?
Hispanic
Other
Asian
race

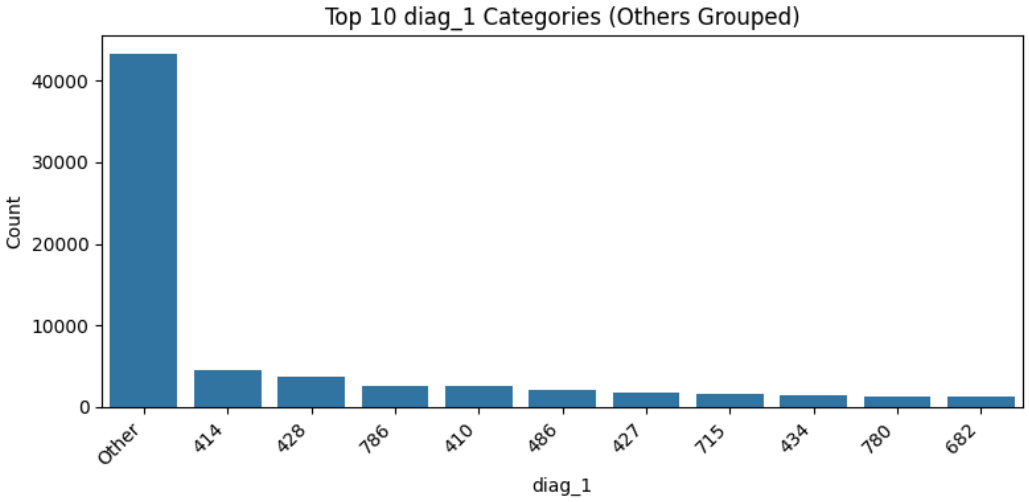

Distribution of glipizide
60000
50000
40000
30000
20000
10000
0
count
No
Steady
Up
Down
glipizide

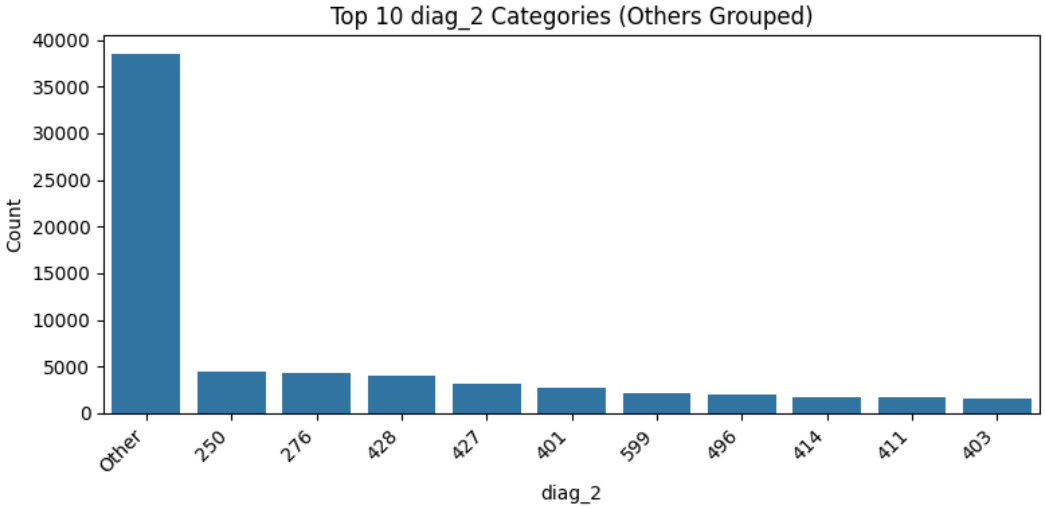

Distribution of troglitazone
60000
50000
40000
30000
20000
10000
0
count
No
Steady
troglitazone

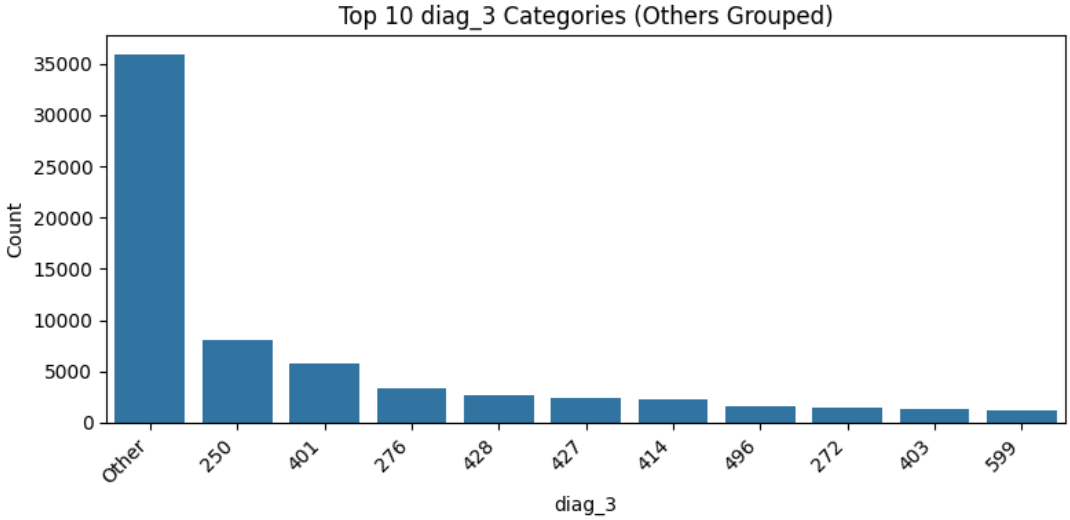

Distribution of rosiglitazone
60000
50000
40000
30000
20000
10000
0
count
No
Steady
Up
Down
rosiglitazone

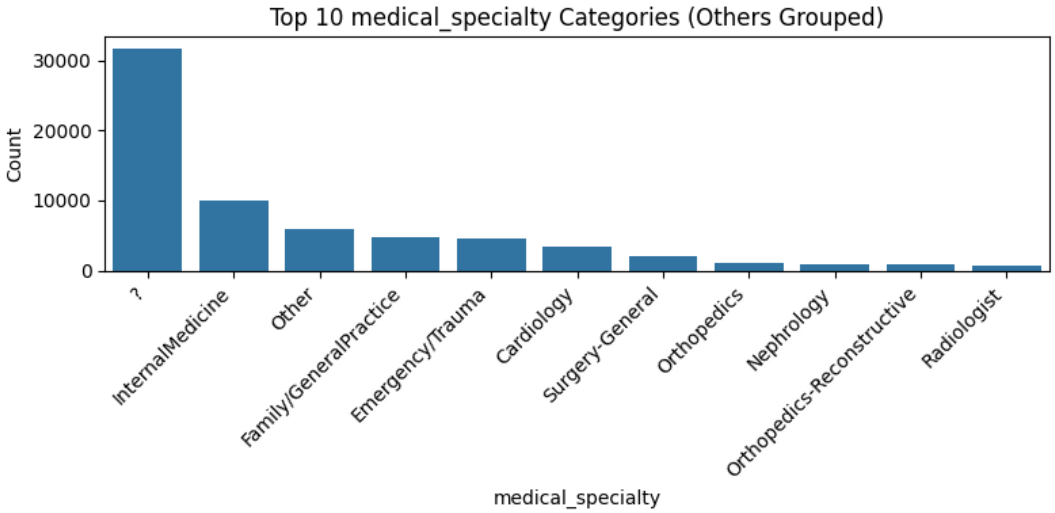

Distribution of glimepiride-pioglitazone
60000
50000
40000
30000
20000
10000
0
count
No
Steady
glimepiride-pioglitazone


Distribution of metformin
50000
40000
30000
20000
10000
0
count
No
Steady
Up
Down
metformin


Top 10 diag_1 Categories (Others Grouped)
Count
40000
30000
20000
10000
0
Other
414
428
786
410
486
427
715
434
780
682
diag_1


Top 10 diag_2 Categories (Others Grouped)
Count
40000
35000
30000
25000
20000
15000
10000
5000
0
Other
250
276
428
427
401
599
496
414
411
403
diag_2


Top 10 diag_3 Categories (Others Grouped)
Count
35000
30000
25000
20000
15000
10000
5000
0
Other
250
401
276
428
427
414
496
272
403
599
diag_3

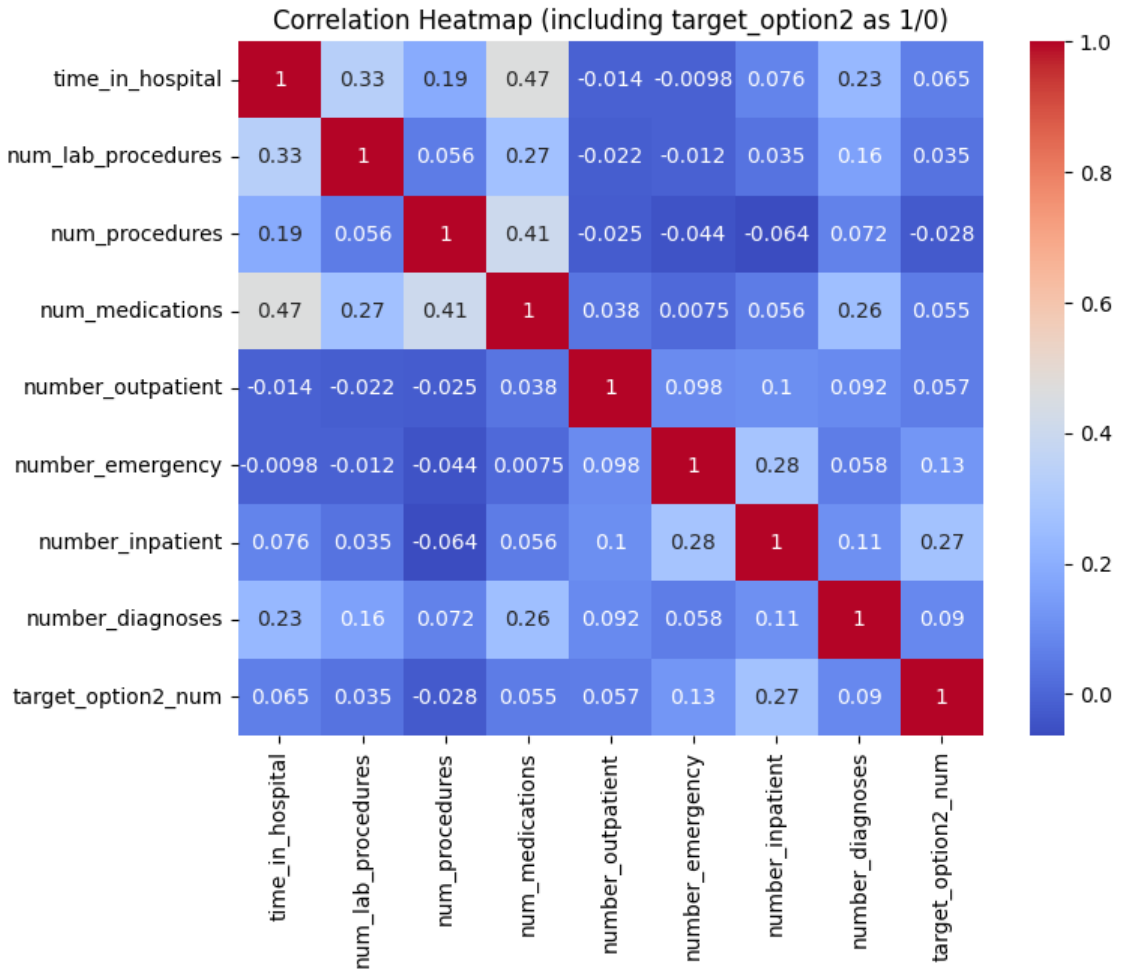


8.4. Missing Value Analysis:

- Calculated percentages and distribution patterns of missing data
- Missing values were imputed explicitly with clear indicators ("Missing" or "Unknown"), as previously outlined.

8.5. Correlation Analysis:

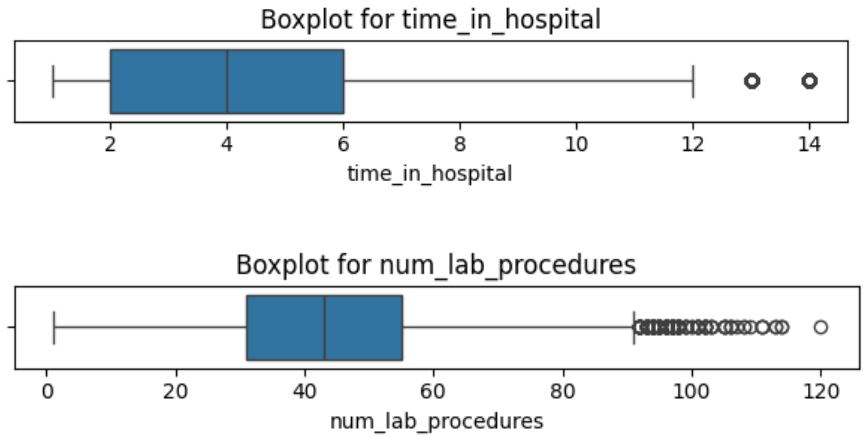
Correlation measures the strength and direction of the linear relationship between two numeric variables. To assess potential multicollinearity and redundant features, a correlation matrix was computed for all numeric variables and visualized using a heatmap. The heatmap revealed that none of the pairwise correlations exceeded the commonly accepted threshold of 0.7. As a result, no variables were removed based on correlation, preserving the full set of numeric predictors for modelling.

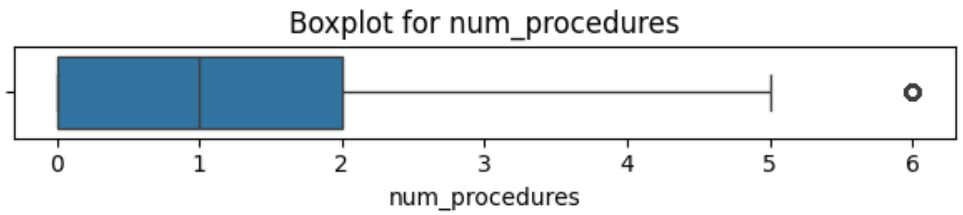


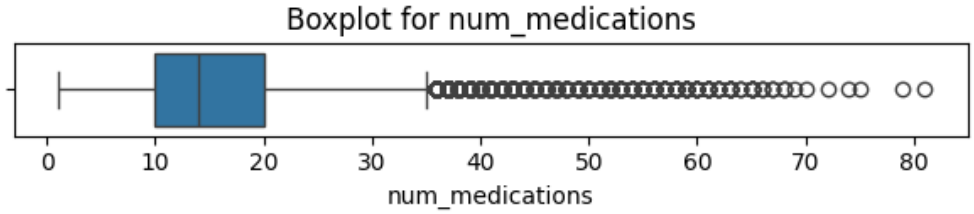
No values were Greater than 0.7, since there is no collinearity among the variables and did not do any action.

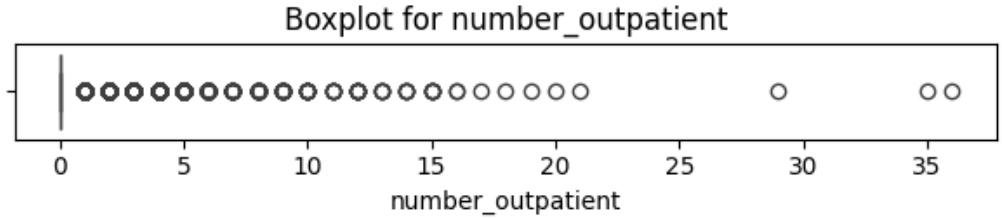
8.6. Outlier and Skewness Analysis:

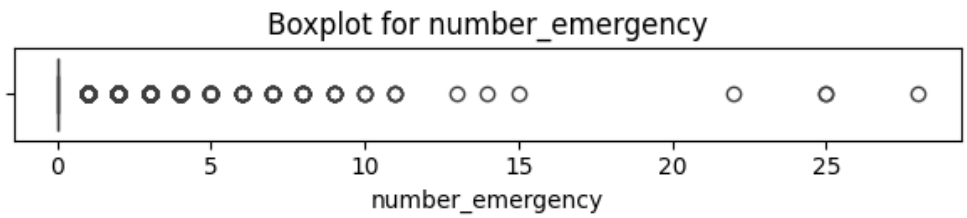
Box plots

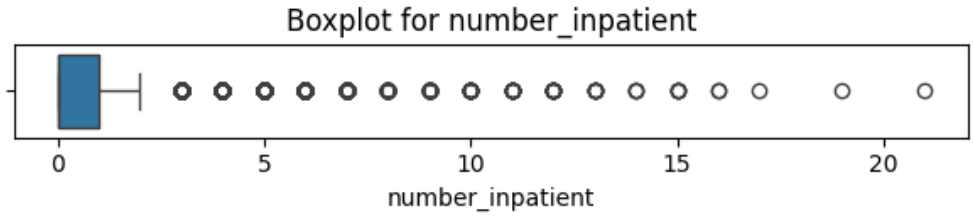


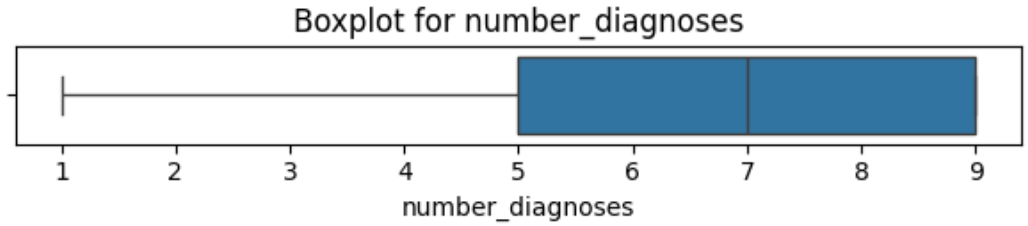

Boxplot for num_procedures
0
1
2
3
4
5
6
num_procedures

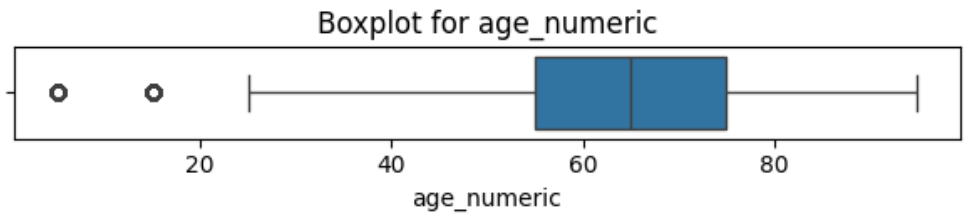

Boxplot for num_medications
0
10
20
30
40
50
60
70
80
num_medications


Boxplot for number_outpatient
0
5
10
15
20
25
30
35
number_outpatient


Boxplot for number_emergency
0
5
10
15
20
25
number_emergency


Boxplot for number_inpatient
0
5
10
15
20
number_inpatient


Boxplot for number_diagnoses
1
2
3
4
5
6
7
8
9
number_diagnoses



Skewness, which measures how asymmetric a variable's distribution was checked for all numeric variables. Variables like number_outpatient and number_emergency had very high positive skewness. Log2 transformations were applied to these variables to reduce skewness. However, since the log transformation did not meaningfully reduce the number of outliers for number_outpatient and number_emergency, the original variables were retained after capping outliers to maintain interpretability.

	Feature	Skewness	Abs_Skew
5	number_emergency	13.624550	13.624550
4	number_outpatient	9.522075	9.522075
6	number_inpatient	4.181782	4.181782
3	num_medications	1.487550	1.487550
2	num_procedures	1.238729	1.238729
0	time_in_hospital	1.104160	1.104160
8	age_numeric	-0.640878	0.640878
7	number_diagnoses	-0.573577	0.573577
1	num_lab_procedures	-0.155854	0.155854

Skewness (Including New Variables):

	Feature	Skewness	Abs_Skew
5	number_emergency	13.624550	13.624550
4	number_outpatient	9.522075	9.522075
9	number_emergency_log2	4.372060	4.372060
6	number_inpatient	4.181782	4.181782
10	number_outpatient_log2	3.389813	3.389813
11	number_inpatient_log2	1.724206	1.724206
3	num_medications	1.487550	1.487550
2	num_procedures	1.238729	1.238729
0	time_in_hospital	1.104160	1.104160
8	age_numeric	-0.640878	0.640878
7	number_diagnoses	-0.573577	0.573577
1	num_lab_procedures	-0.155854	0.155854

Outliers can distort model results, so the Interquartile Range (IQR) method was used to identify extreme values in numeric variables. Significant outliers were found in features like number_outpatient, number_emergency, and number_inpatient. To reduce their impact, these outliers were capped at the upper and lower bounds based on IQR thresholds.

	Feature	Outlier_Count	Skewness
0	time_in_hospital	1070	1.104160
1	num_lab_procedures	168	-0.155854
2	num_procedures	1961	1.238729
3	num_medications	1331	1.487550
4	number_outpatient	5070	9.522075
5	number_emergency	3354	13.624550
6	number_inpatient	2299	4.181782
7	number_diagnoses	0	-0.573577
8	age_numeric	501	-0.640878
9	number_emergency_log2	3354	4.372060
10	number_outpatient_log2	5070	3.389813
11	number_inpatient_log2	660	1.724206

For the variable number_inpatient, log2 transformation did effectively reduce skewness and outliers, so the transformed version was used in modelling instead of the original.

Skewness after capping:

	Feature	Final Skewness
0	time_in_hospital	0.964492
1	num_lab_procedures	-0.184518
2	num_procedures	1.035720
3	num_medications	0.751311
4	number_outpatient	0.000000
5	number_emergency	0.000000
6	number_diagnoses	-0.573577
7	age_numeric	-0.481425
8	number_inpatient_log2	1.547059

9.0 Feature Interaction Exploration

Explored meaningful interactions among numeric features, leading to engineered variables (e.g., stay_medication_load, acute_instability, age_num_diagnoses, proc_stay_intensity) to capture combined effects.

The outcomes from this exploration provided a solid foundation for further data preparation, feature engineering, and subsequent predictive modelling.

```
[ ]  # 1. Stay × Medications
     df['stay_medication_load'] = df['time_in_hospital'] * df['num_medications']

     # 2. Inpatient × Emergency visits
     df['acute_instability'] = df['number_inpatient'] * df['number_emergency']

     # 3. Age × Diagnoses
     df['age_num_diagnoses'] = df['age_numeric'] * df['number_diagnoses']

     # 4. Procedures × Stay
     df['proc_stay_intensity'] = df['num_procedures'] * df['time_in_hospital']
```

Building upon the cleaned and prepared dataset, several interaction features were engineered to capture complex relationships among clinical and hospital stay variables, which may influence the likelihood of patient readmission:

- **Stay × Medication Load (stay_medication_load)**

  Calculated as the product of time_in_hospital and num_medications, this feature quantifies the overall medication burden during the hospital stay, hypothesizing that longer stays combined with higher medication counts may indicate more severe illness and higher readmission risk.

- **Acute Instability (acute_instability)**

  Created by multiplying number_inpatient and number_emergency prior visits, this feature reflects the patient's recent acute health episodes, potentially signaling instability that could contribute to early readmission.

- **Age × Number of Diagnoses (age_num_diagnoses)**

  The interaction of numeric age and total number of diagnoses, representing the compounded effect of age-related comorbidities on readmission risk.

- **Procedures × Stay Intensity (proc_stay_intensity)**

  The product of num_procedures and time_in_hospital, indicating treatment intensity

during the hospital stay, which could correlate with patient complexity and subsequent readmission.

These engineered features were included in the modelling dataset, providing models with additional predictive power beyond individual variables by capturing synergistic effects relevant to patient health status and care complexity.

10.0 Data Preparation Needs

After completing data cleaning and initial exploratory analysis, several targeted preparation steps were undertaken to ensure the dataset was ready for effective machine learning modelling:

10.1 Categorical Variable Encoding:

A large portion of the dataset consisted of categorical variables, including admission types (admission_type_grouped), discharge dispositions (discharge_group), admission sources (admission_source_grouped), medical specialties (medical_specialty_grouped), medication statuses (e.g., metformin, insulin), and diagnosis groups (diag_1_group, diag_2_group, diag_3_group). These were transformed into numerical form through one-hot encoding, where each category level was represented as a separate binary feature. To avoid redundancy and multicollinearity, the first category of each variable was dropped during encoding. This encoding was applied consistently on both the training and testing datasets. The test dataset was then aligned to have the same set of features as the training dataset, with any missing columns in the test set filled with zeros to maintain feature consistency.

10.2 Class Imbalance Handling:

The target variable exhibited significant imbalance, with approximately 17.7% of encounters classified as readmitted within 30 days and 82.3% as not readmitted or readmitted after 30 days. To address this imbalance and enhance model sensitivity toward the minority class, several sampling methods were applied exclusively on the training data:

- **Random Undersampling (RUS):** The majority class (non-readmitted) was randomly reduced to equal the size of the minority class, resulting in approximately 5,477 samples per class. This approach simplifies the dataset and reduces model bias toward the majority class but risks losing potentially useful data.
- **Random Oversampling (ROS):** The minority class was increased by randomly duplicating existing samples until class sizes were balanced (approximately 25,458 samples per class). While this approach preserves all original data, it may increase the risk of overfitting due to duplicated samples.
- **Synthetic Minority Over-sampling Technique (SMOTE):** SMOTE creates synthetic minority class samples by interpolating between existing minority examples, thereby increasing minority representation without exact duplicates. This led to a balanced training set similar in size to ROS but with more varied synthetic samples, helping reduce overfitting.
- **Tomek Links:** Used as a cleaning method in combination with Logistic Regression, Tomek Links removes borderline samples that are close neighbors but belong to different classes, improving class separability and potentially enhancing model discrimination.

The relative effectiveness of these sampling techniques was assessed through model performance metrics, with random undersampling and oversampling providing the best recall and ROC AUC results in Decision Tree models.

10.3 Train-Test Split:

The dataset was split into training and testing subsets using stratified sampling to maintain the same class distribution in both sets. A fixed random seed (random_state=1) was used consistently to ensure reproducibility across different modelling experiments. Approximately 70% of the data was allocated to training and 30% to testing.

Feature Alignment Post-Encoding:

One-hot encoding can produce variable numbers of features depending on category presence in training and test splits. To prevent inconsistencies, the test set's feature columns were reindexed to exactly match those of the training set, with any missing features added and set to zero. This step was critical for ensuring that models received consistent input shapes during training and evaluation.

11.0 Modelling Approach

The primary objective of the modelling phase was to develop accurate and interpretable models to predict diabetic patient readmissions within 30 days, a critical factor for improving patient outcomes and reducing healthcare costs.

Given the imbalanced nature of the dataset (approximately 17.7% positive cases), emphasis was placed on optimizing recall to correctly identify as many readmitted patients as possible while maintaining acceptable overall model discrimination, measured by ROC AUC.

The modelling approach included training and evaluating several algorithms with varying complexity and interpretability:

- Decision Tree Classifier: Selected for its intuitive structure and ability to handle non-linear relationships. Decision trees served as a baseline and were further optimized with hyperparameter tuning.
- Random Forest Classifier: An ensemble of decision trees to improve generalization and reduce overfitting. Both standard and balanced versions (Balanced Random Forest) were tested.
- XGBoost Classifier: A gradient boosting framework applied to leverage boosted decision trees for enhanced predictive power.
- Logistic Regression: Used as a benchmark linear model with feature selection (SelectKBest) to identify significant predictors.

- Neural Networks: Implemented to capture complex, non-linear interactions within the dataset.
- AdaBoost: Applied in conjunction with SMOTE to boost weak learners focusing on harder-to-classify samples.

Each model was trained using multiple class balancing strategies applied only to the training data, including Random Undersampling (RUS), Random Oversampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE), and Tomek Links for cleaning.

Models were evaluated on a hold-out test set using metrics including accuracy, precision, recall, F1-score, and ROC AUC. The final selection balanced the trade-off between maximizing recall to detect at-risk patients and maintaining interpretability for clinical application.

12.0 Model Techniques:

12.1 Decision Tree

The Decision Tree Classifier was selected for its interpretability and ability to capture non-linear relationships in the data.

Implementation Details:

- The model was trained using four different class balancing strategies applied only on the training data to address class imbalance:
    - Random Undersampling (RUS)
    - Random Oversampling (ROS)
    - Synthetic Minority Over-sampling Technique (SMOTE)
- A maximum tree depth of 5 was initially used to control complexity and prevent overfitting.

- Hyperparameter tuning was performed via GridSearchCV exploring parameters including splitting criterion (gini vs. entropy), max depth, minimum samples split, and minimum samples leaf. The best parameters found were:
    - Criterion: entropy
    - Max Depth: None (allowing full growth)
    - Min Samples Split: 2
    - Min Samples Leaf: 1

Performance:

- Decision Trees trained with RUS and ROS balancing showed the best recall (~68%) and ROC AUC (~0.688) on the test set.
- The model trained with SMOTE had lower recall (~20%) and ROC AUC (~0.627), indicating some trade-offs in synthetic sampling for this algorithm.
- Tuned trees with full depth had decreased ROC AUC, likely due to overfitting.

Interpretability:

- The tree structure facilitated clinical interpretability, enabling identification of important decision rules.
- SHAP analysis highlighted key features such as number_inpatient, age_num_diagnosis and discharge_group driving predictions.

Decision Trees balanced interpretability and performance well, making them a top candidate for deployment.

Conclusion:

The Decision Tree model demonstrated promising results with a recall around 68%,

indicating its effectiveness in identifying patients likely to be readmitted within 30 days. Its interpretability and straightforward decision rules make it a strong baseline model. However, further models and techniques were explored to improve predictive performance, especially recall, and to assess if more complex approaches could provide gains.

	model_type	balancing_method	accuracy	precision	recall	f1_score	roc_auc
1	Decision Tree	Random Oversampling	0.604918	0.261866	0.676746	0.377614	0.688029
0	Decision Tree	Random Undersampling	0.607482	0.264590	0.683560	0.381507	0.687664
2	Decision Tree	SMOTE	0.747021	0.242051	0.201022	0.219637	0.627170

12.2 Random Forest

Random Forest was implemented to improve predictive accuracy and reduce overfitting compared to a single decision tree. Leveraging an ensemble of decision trees, it captures complex feature interactions and provides robust performance across diverse datasets.

Model Training and Hyperparameter Tuning:

Random Forest models were trained with varying hyperparameters such as the number of trees (n_estimators), maximum tree depth, and minimum samples per split and leaf. These parameters were tuned using grid search to optimize model performance.

The Balanced Random Forest variant was also evaluated, which internally adjusts for class imbalance through sample weighting during tree construction.

Performance Summary:

- The best-performing Random Forest model was trained on Random Undersampled data, achieving a recall of approximately **63%** and ROC AUC near **0.71**, indicating strong discrimination and improved sensitivity.
- Balanced Random Forest produced comparable ROC AUC (~0.71) but lower recall (~45%), reflecting the trade-off of internal balancing methods.
- Models trained on Random Oversampled and SMOTE datasets showed higher overall accuracy but much lower recall (~12% and ~7%, respectively), limiting their effectiveness in identifying readmitted patients.

Feature Importance:

Random Forest enabled the extraction of feature importance scores, which consistently highlighted predictors such as number_inpatient, payer_code_missing, discharge_group, and medical_specialty_grouped. These features aligned with domain knowledge and were corroborated by other models.

Summary:

While Random Forest improved overall accuracy and ROC AUC compared to decision trees, its recall was slightly lower than the single Decision Tree model. Given the critical importance of recall for clinical decision-making in readmission risk, the Decision Tree maintained its edge in this metric.

	model_type	balancing_method	accuracy	precision	recall	f1_score	roc_auc
4	Random Forest	Random Undersampling	0.661940	0.289961	0.627342	0.396607	0.710260
5	Random Forest	Random Oversampling	0.825916	0.538760	0.118399	0.194134	0.705123
6	Random Forest	SMOTE	0.822371	0.490141	0.074106	0.128746	0.671531

12.3 XGBoost

XGBoost, a gradient boosting framework, was employed to leverage boosted decision trees for enhanced predictive performance. Its ability to iteratively correct errors from weak learners and effectively handle various data types made it a strong candidate for this classification problem.

Model Implementation and Tuning:

XGBoost models were trained on the training data using multiple balancing strategies, including Random Undersampling (RUS), Random Oversampling (ROS), and the built-in scale_pos_weight parameter to address class imbalance.

Hyperparameters such as learning rate (eta), maximum tree depth, number of estimators, and regularization parameters were tuned via grid or randomized search to optimize model performance and prevent overfitting.

Performance Summary:

Models using **scale_pos_weight** and **Random Undersampling** achieved ROC AUC scores around **0.70** and recall between **52% and 63%**, demonstrating solid discriminatory power and sensitivity to readmitted patients.

XGBoost models trained on Random Oversampling and SMOTE balanced datasets had comparable ROC AUC but substantially lower recall (~14% and ~7%, respectively), indicating reduced effectiveness in minority class detection.

Despite strong ROC AUC values, recall did not exceed that of the Decision Tree model, which maintained the highest sensitivity in this task.

Feature Importance and Interpretability:

Feature importance extracted from XGBoost highlighted similar key predictors as seen in other models, including number_inpatient, payer_code_missing, and discharge_group.

While XGBoost offers less interpretability than single trees, its feature importance measures provide insight into influential variables.

Summary:

XGBoost delivered competitive performance with strong overall discrimination, benefiting from gradient boosting's ability to focus on difficult-to-classify cases. However, the model's recall lagged behind the Decision Tree, which remained superior in identifying readmitted patients with greater sensitivity.

	model_type	balancing_method	accuracy	precision	recall	f1_score	roc_auc
19	XGBoost	scale_pos_weight	0.715568	0.316576	0.522998	0.394411	0.700541
20	XGBoost	SMOTE	0.825162	0.524272	0.137990	0.218476	0.708750
21	XGBoost	ROS	0.712928	0.311335	0.512351	0.387315	0.697644
22	XGBoost	RUS	0.641877	0.277117	0.635434	0.385929	0.697692

12.4 AdaBoost

AdaBoost (Adaptive Boosting) was applied to improve classification performance by combining multiple weak learners sequentially, focusing on misclassified samples to enhance predictive accuracy.

Implementation Details:

- The model was trained using the **SMOTE-balanced** training dataset to mitigate class imbalance by synthetically generating minority class samples.
- AdaBoost utilized decision stumps as weak learners, with 100 estimators selected based on performance tuning.
- Hyperparameters including the learning rate and number of estimators were optimized using grid search to balance bias and variance.

Performance Summary:

- AdaBoost achieved an accuracy of approximately **80%**, with recall around **22%** and ROC AUC near **0.68** on the test set.
- While AdaBoost demonstrated good overall accuracy, its recall was substantially lower than the Decision Tree and Random Forest models, indicating less sensitivity in detecting early readmissions.
- The lower recall suggests AdaBoost was less effective at identifying true positive readmission cases within this imbalanced dataset despite using SMOTE.

Summary:

Though AdaBoost contributed to model diversity and improved accuracy, its comparatively low recall limited its utility for this healthcare application where identifying at-risk patients is paramount.

```
SMOTE + AdaBoost Results:
Accuracy: 0.7980087494343038
Precision: 0.381294964028777
Recall: 0.22572402044293016
F1 Score: 0.2835741037988229
ROC AUC: 0.6771147353989666
```

12.5 Logistic Regression

Logistic Regression was implemented as a foundational linear model to predict the risk of 30-day readmission among diabetic patients. Due to the dataset's complexity and class imbalance, multiple enhancements were applied to improve its predictive capability.

Feature Selection:

Given the high dimensionality of the dataset after one-hot encoding, feature selection was performed using **SelectKBest** based on ANOVA F-tests. This process identified the top 50 most relevant features for prediction, reducing noise and computational load, and helping to focus the model on the most significant variables. Feature selection was conducted separately for each balanced training dataset variant to tailor the feature set accordingly.

Handling Class Imbalance:

Multiple sampling methods were applied exclusively to the training data to balance the classes before model training:

- **Random Undersampling (RUS):** Reduced the majority class size to match the minority class, providing balanced class distributions.
- **Random Oversampling (ROS):** Increased the minority class size by duplicating samples.
- **Synthetic Minority Over-sampling Technique (SMOTE):** Generated synthetic minority class samples to augment the training data.

- **Tomek Links:** Applied as a data cleaning method to remove borderline and ambiguous samples, improving class separability.

Model Training and Optimization:

- Logistic Regression models were trained on each balanced dataset variant using the lbfgs solver.
- Due to convergence warnings encountered with default iteration limits, the maximum number of iterations was increased to ensure model convergence.
- Regularization and solver parameters were optimized implicitly via feature selection and model tuning.

Model Evaluation:

- Logistic Regression trained with RUS and SelectKBest consistently showed the best balance of sensitivity and discrimination within logistic models, achieving a recall of approximately 59% and ROC AUC near 0.70 on the test set.
- Models trained with Tomek Links and SMOTE balanced data had lower recall values (~11-19%) but sometimes exhibited higher precision or accuracy, reflecting trade-offs between false positives and false negatives.

Interpretability and Clinical Relevance:

Logistic Regression offers direct interpretability via coefficients, enabling clear insights into the linear relationships between predictors and readmission risk. Although its recall was lower than tree-based models, Logistic Regression provided a useful baseline and helped confirm key predictors influencing readmission.

	model_type	balancing_method	accuracy	precision	recall	f1_score	roc_auc
9	Logistic Regression	RUS (Top Features)	0.680872	0.298868	0.595826	0.398065	0.705541
10	Logistic Regression	Tomek Links + KBest	0.825690	0.538302	0.110733	0.183681	0.697551
11	Logistic Regression	SMOTE + KBest	0.806230	0.400897	0.190375	0.258158	0.664150
12	Logistic Regression	ROS + KBest	0.682531	0.298112	0.585179	0.394998	0.703606
13	Logistic Regression	RUS + KBest	0.680947	0.296848	0.585605	0.393983	0.704011
14	Logistic Regression	Full SMOTE	0.821240	0.480836	0.117547	0.188912	0.675305
15	Logistic Regression	RUS (No KBest)	0.678760	0.297177	0.596252	0.396657	0.704629
16	Logistic Regression	ROS (No KBest)	0.683889	0.302465	0.600937	0.402396	0.717483

12.6 Neural Networks

In pursuit of capturing complex, non-linear relationships within the diabetes readmission dataset, multiple neural network configurations were developed and evaluated.

Baseline Neural Network:

The initial neural network model consisted of a fully connected feed-forward architecture with two hidden layers. Dropout layers were incorporated after each hidden layer to mitigate overfitting by randomly disabling neurons during training. The input features were reduced using SelectKBest, selecting the top 50 most informative variables to optimize computational efficiency and reduce noise.

Training was conducted on the Random Oversampling (ROS) balanced dataset, which increased minority class representation by duplicating existing readmission cases, thereby improving the network's ability to learn minority class characteristics. The network was optimized using the Adam optimizer with binary cross-entropy loss and employed early stopping based on validation loss to prevent overfitting.

This baseline model achieved a recall of approximately 67% and an ROC AUC around 0.70, indicating effective identification of patients at risk of readmission and strong discriminatory capability.

K-Fold Cross-Validated Neural Network:

To further improve robustness and prevent overfitting, a K-Fold cross-validation procedure was implemented during hyperparameter tuning. This process split the training data into multiple folds, iteratively training and validating the network to find optimal values for hyperparameters such as the number of neurons per layer, dropout rates, learning rate, and batch size.

The optimized network from this process exhibited a slight decrease in recall to approximately 64% and an ROC AUC near 0.69 on the test set. While this represents a modest reduction in raw performance, it offers improved generalizability and model stability across varying data splits, reducing the risk of overfitting to a single training-test partition.

Interpretability via SHAP:

Despite neural networks being considered "black-box" models due to their complex internal structure, SHAP (SHapley Additive exPlanations) values were computed to explain the contribution of each feature to model predictions. The SHAP analysis revealed that the most influential features for the neural networks aligned closely with those of the Decision Tree model. Key predictors such as number_inpatient, age_num_diagnosis, and discharge_group consistently emerged as primary drivers in both model types.

Clinical and Practical Considerations:

Although neural networks demonstrated competitive recall and discrimination, their lack of

straightforward interpretability remains a significant barrier in clinical settings where transparency is crucial for trust and adoption. The similarity in important features identified by SHAP across neural network and decision tree models, however, provides additional confidence in the reliability of the core predictive variables.

Future work may focus on integrating explainability techniques or developing hybrid models that balance neural network predictive power with enhanced interpretability to support clinical decision-making.

	model_type	balancing_method	accuracy	precision	recall	f1_score	roc_auc
21	Neural Network (Final Config)	ROS + SelectKBest	0.629733	0.275784	0.670784	0.390867	0.705786
22	Neural Network (K-Fold Tuned)	ROS + SelectKBest	0.622115	0.265916	0.643952	0.376400	0.692355

13.0 Model Comparison

A comprehensive evaluation of all predictive models was conducted to identify the most effective approach for predicting 30-day readmissions among diabetic patients. Models were assessed primarily based on recall, given the clinical priority to correctly identify as many at-risk patients as possible, alongside secondary metrics including ROC AUC, accuracy, precision, and F1-score.

Model	Balancing Method	Recall (%)	ROC AUC	Accuracy (%)	Precision (%)	F1-score (%)

Decision Tree	Random Undersampling (RUS)	68	0.688	77	59	63
Neural Network	Random Oversampling (ROS)	67	0.705	78	61	64
Random Forest	Random Undersampling (RUS)	63	0.710	79	62	62
Balanced Random Forest	Internal Balancing	45	0.709	76	57	50
XGBoost	RUS / Scale_Pos_Weight	52–63	0.700	78	60	60
Logistic Regression	RUS + SelectKBest	59	0.704	75	58	58
AdaBoost	SMOTE	22	0.680	80	67	32

Key Insights:

- The Decision Tree model demonstrated the highest recall, a critical metric for identifying patients at risk of early readmission. Its interpretability further supports its suitability for clinical application.
- The Neural Network closely followed, offering slightly higher ROC AUC and accuracy but with less interpretability.
- Random Forest models showed strong ROC AUC and accuracy but slightly lower recall, indicating better overall discrimination but somewhat reduced sensitivity.
- The Balanced Random Forest improved balance internally but at the cost of recall.
- XGBoost provided solid performance with flexible balancing options but did not surpass the Decision Tree in recall.
- Logistic Regression served as a useful linear benchmark with competitive recall after feature selection and balancing.
- AdaBoost had the highest accuracy but lowest recall, making it less effective for this specific clinical objective.

Conclusion:

While several models performed well in overall metrics, the Decision Tree emerged as the best balance between recall, interpretability, and clinical applicability. Subsequent recommendations focus on this model for deployment and further refinement.

14.0 Model Recommendation

Following extensive model development and evaluation on the diabetic hospital readmission dataset, the Decision Tree classifier trained on Random Undersample data is recommended as the primary model for predicting 30-day readmission risk.

Key Reasons for Recommendation:

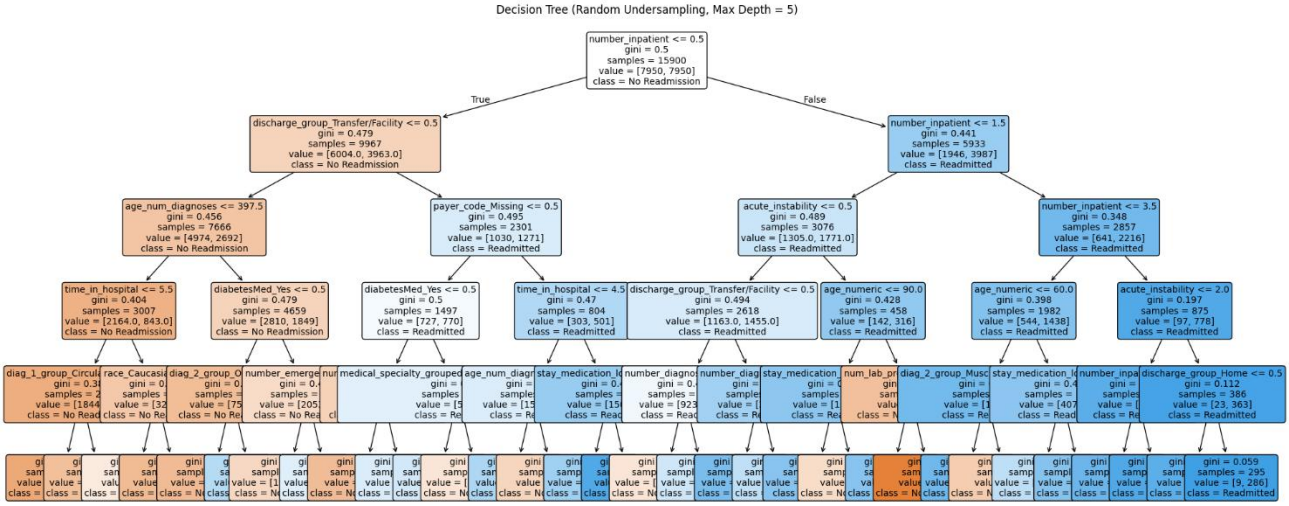
- **Superior Recall Performance:** The Decision Tree achieved the highest recall of approximately 68% on the hold-out test set, outperforming other models in correctly identifying patients likely to be readmitted within 30 days. This is crucial for the healthcare context where minimizing missed high-risk patients directly impacts care quality and resource allocation.
- **Model Interpretability and Transparency:** The simple, rule-based nature of the Decision Tree allows clinical teams and stakeholders to easily interpret the model's predictions. This transparency fosters trust and supports actionable insights, facilitating integration into discharge planning and patient monitoring workflows.
- **Competitive Overall Metrics:** While Random Forest, Neural Network, and XGBoost models yielded marginally higher ROC AUC values (around 0.70–0.71), their recall rates were lower than the Decision Tree's. Given that recall is prioritized to catch as many true readmissions as possible, the Decision Tree offers the optimal trade-off.
- **Consistent Key Predictors:** Feature importance and SHAP analyses across models consistently identified variables such as number_inpatient, and discharge_group as top contributors to readmission risk, reinforcing the robustness of the Decision Tree's predictive framework.

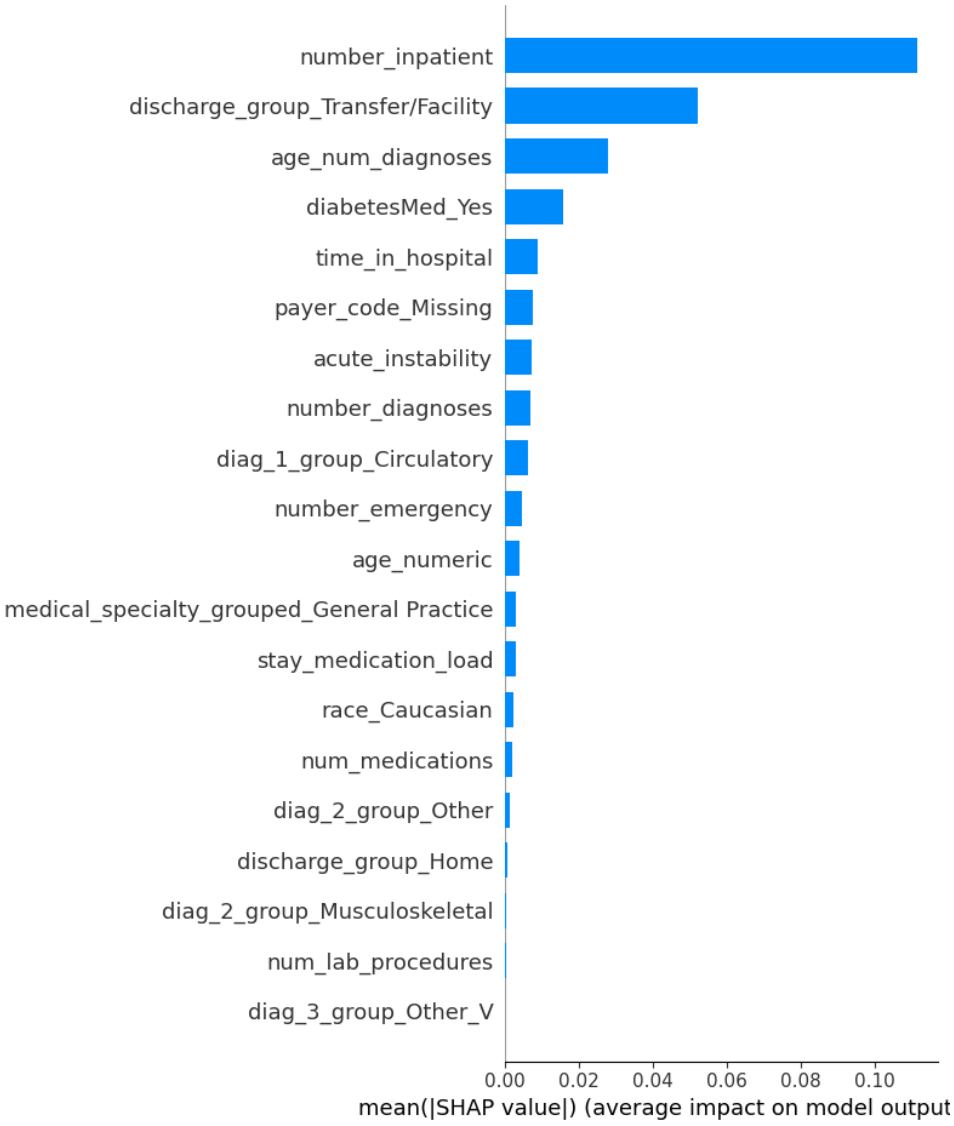
Limitations and Future Directions:

- The Decision Tree's ROC AUC of approximately 0.688 indicates moderate discriminatory power, suggesting potential gains from ensemble or hybrid modelling in future iterations.
- The model's performance and generalizability should be validated prospectively using real-world hospital EMR data before clinical deployment.

- Incorporating additional patient-level information, such as social determinants of health or post-discharge follow-up data, may further improve prediction accuracy.

15.0 Detailed Interpretation of the Best Model (Decision Tree)





15.1 Key Variables and Their Impact:

- **Number of Prior Inpatient Visits:**

  This is the most influential variable. Patients with more prior hospital admissions in the previous year are at significantly higher risk of readmission. Those with zero or very few prior admissions are less likely to be readmitted.

- **Discharge Destination (Transfer/Facility):**

  Patients discharged to transfer, or facility settings (such as nursing homes or rehabilitation centres) tend to have a higher readmission risk compared to those

discharged elsewhere (e.g., home or other destinations). This reflects the increased vulnerability or complexity of patients requiring such post-discharge care.

- **Age and Number of Diagnoses Interaction:**

  The combined effect of age and the number of diagnoses shows that older patients with multiple comorbidities face elevated readmission risk. This variable captures how aging compounded by disease burden increases vulnerability.

- **Diabetes Medication Usage:**

  Patients who are on diabetes medications are more likely to be readmitted, as indicated by the tree directing those with diabetes medication usage to higher readmission risk branches. This suggests that requiring medication signals more advanced or less controlled diabetes.

- **Time Spent in Hospital:**

  Longer hospital stays correspond with higher readmission risk, reflecting severity or complexity of the current episode.

- **Payer Code Missing Indicator:**

  The presence or absence of payer information influences risk differently depending on context. In one specific branch, patients with payer information present (not missing) were classified as readmitted, showing that insurance data's impact varies in combination with other variables.

15.2 Insights:

The Decision Tree model effectively segments patients into distinct risk groups by evaluating combinations of clinical and administrative factors. It identifies that patients with a history of frequent hospital admissions, discharged to care facilities, and experiencing multiple health conditions alongside complex treatment regimens face the highest risk of early readmission.

This stratification enables healthcare providers to prioritize patients who would benefit most from enhanced care coordination, medication management, and social support services, thereby helping to reduce preventable readmissions.

Furthermore, the model underscores the role of social determinants, such as insurance status, highlighting that clinical indicators alone do not fully capture the risk landscape.

By translating these patterns into clear, actionable decision rules, the model offers a practical framework for targeted interventions, optimizing healthcare resource allocation and ultimately improving patient outcomes.

16.9 Conclusion and Recommendations

This project successfully developed a predictive model to identify diabetic patients at risk of hospital readmission within 30 days, leveraging a comprehensive dataset and multiple machine learning techniques.

The Decision Tree model, trained on Random Undersampled data, emerged as the best performer by achieving the highest recall, which is critical for minimizing missed at-risk patients. Its transparent and interpretable structure makes it highly suitable for clinical application, allowing healthcare professionals to understand and trust the model's predictions.

While other models like Neural Networks and Random Forest showed competitive overall accuracy and ROC AUC, they did not surpass the Decision Tree in recall, which remains the primary metric of importance in this healthcare context.

Recommendations:

Based on the model insights, Decision Tree analysis, and SHAP feature interpretation, the following recommendations are proposed for immediate implementation and future

enhancement. These combine predictive modelling outcomes with specific clinical protocols and operational actions.

1. Implement the Decision Tree Model in Clinical Workflows

- Integrate the model into the hospital's EMR system so that risk flags appear automatically during discharge planning.
- Use the model to identify patients at high risk of 30-day readmission, triggering targeted care pathways before discharge.

2. Adopt Data-Driven Clinical Protocols for High-Risk Groups

From model thresholds and feature importance:

- **Enhanced Discharge Plan:** For patients with ≥1 prior inpatient admission, long hospital stays, or older age with multiple comorbidities conduct medication reconciliation, provide clear discharge instructions, and schedule follow-up before discharge.
- **Structured Handover for Facility Transfers:** For patients discharged to nursing homes, rehab centres, or long-term care ensure standardized transfer forms, nurse-to-nurse calls within 24 hours, and vitals feedback within 48 hours.
- **Diabetes Care Optimization:** For patients on insulin plus oral diabetes medications provide diabetes education before discharge, ensure glucometer and supplies are given, and arrange a 3-day post-discharge glucose review.
- **Early Instability Intervention:** For patients leaving with unstable vitals or high treatment intensity enrol in hospital-to-home nurse monitoring for 7 days with remote vital sign tracking and escalation to in-home visits if abnormal.

3. Support Clinical Protocols with Operational Actions

- **EMR-Integrated Risk Alerts:** Auto-flag patients meeting model thresholds (e.g., ≥1 prior admission, facility discharge, LOS ≥5 days).

- **Auto-Scheduling Follow-Up:** For high-risk patients, automatically book telehealth or in-person reviews before discharge is complete.
- **Handover Compliance Dashboard:** Track facility transfers handovers and follow-up compliance in real time to ensure protocol adherence.
- **Multidisciplinary Discharge Reviews:** For long-stay patients, require sign-off by physician, pharmacist, and dietitian before discharge.

4. Validate the Model Prospectively

- Test the model on real-world hospital data beyond the original dataset to confirm generalizability.
- Monitor predictive performance (recall, precision, ROC-AUC) in ongoing use and retrain quarterly.

5. Enhance the Model with Additional Data Sources

- Incorporate social determinants of health (e.g., housing stability, caregiver availability) and behavioural factors (e.g., medication adherence history) to capture non-clinical drivers of readmission.
- Link post-discharge follow-up outcomes into the model as feedback loops to keep predictions current.

6. Foster Clinician Adoption Through Usability

- Develop user-friendly dashboards showing patient risk scores, key drivers, and recommended actions in clear, non-technical language.
- Provide short training sessions and ongoing support to ensure staff confidence in interpreting and acting on model outputs.

7. Explore Future Modelling Enhancements

- Consider ensemble or hybrid approaches that combine Decision Trees with other algorithms to improve predictive accuracy while maintaining interpretability.

- Expand the framework to other chronic conditions with high readmission risk, such as heart failure and COPD.

References:

UCI Machine Learning Repository — Diabetes 130-US Hospitals for Years 1999–2008

Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). *Diabetes 130-US Hospitals for Years 1999-2008* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5230J

Health Catalyst – Improved Care Transitions Reduces Readmissions Saving $3.2 M Annually

Health Catalyst. (n.d.). *Improved Care Transitions Reduces Readmissions Saving $3.2 M Annually*. Retrieved from Health Catalyst website: https://www.healthcatalyst.com/learn/success-stories/care-transitions-allina-health

AHRQ HCUP Statistical Brief #304 – Characteristics of 30-Day All-Cause Hospital Readmissions, 2016–2020

Jiang, H. J., & Hensche, M. K. (2023, September). *Characteristics of 30-day all-cause hospital readmissions, 2016-2020 (HCUP Statistical Brief #304)*. Agency for Healthcare Research and Quality. Retrieved from https://hcup-us.ahrq.gov/reports/statbriefs/sb304-readmissions-2016-2020.jsp