pip install ucimlrepo

In [1]:

```
from ucimlrepo import fetch_ucirepo

# fetch dataset
diabetes_130_us_hospitals_for_years_1999_2008 = fetch_ucirepo(id=296)

# data (as pandas dataframes)
X = diabetes_130_us_hospitals_for_years_1999_2008.data.features
y = diabetes_130_us_hospitals_for_years_1999_2008.data.targets

# metadata
print(diabetes_130_us_hospitals_for_years_1999_2008.metadata)

# variable information
print(diabetes_130_us_hospitals_for_years_1999_2008.variables)
```

```
C:\Users\Nishant\AppData\Roaming\Python\Python313\site-packages\ucimlrepo\fetch.py:97: D
typeWarning: Columns (10) have mixed types. Specify dtype option on import or set low_me
mory=False.
  df = pd.read_csv(data_url)
{'uci_id': 296, 'name': 'Diabetes 130-US Hospitals for Years 1999-2008', 'repository_ur
l': 'https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-20
08', 'data_url': 'https://archive.ics.uci.edu/static/public/296/data.csv', 'abstract':
'The dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and i
ntegrated delivery networks. Each row concerns hospital records of patients diagnosed wi
th diabetes, who underwent laboratory, medications, and stayed up to 14 days. The goal i
s to determine the early readmission of the patient within 30 days of discharge.\nThe pr
oblem is important for the following reasons. Despite high-quality evidence showing impr
oved clinical outcomes for diabetic patients who receive various preventive and therapeu
tic interventions, many patients do not receive them. This can be partially attributed t
o arbitrary diabetes management in hospital environments, which fail to attend to glycem
ic control. Failure to provide proper diabetes care not only increases the managing cost
s for the hospitals (as the patients are readmitted) but also impacts the morbidity and
mortality of the patients, who may face complications associated with diabetes.\n', 'are
a': 'Health and Medicine', 'tasks': ['Classification', 'Clustering'], 'characteristics':
['Multivariate'], 'num_instances': 101766, 'num_features': 47, 'feature_types': ['Catego
rical', 'Integer'], 'demographics': ['Race', 'Gender', 'Age'], 'target_col': ['readmitte
d'], 'index_col': ['encounter_id', 'patient_nbr'], 'has_missing_values': 'yes', 'missing
_values_symbol': 'NaN', 'year_of_dataset_creation': 2014, 'last_updated': 'Tue Sep 24 20
24', 'dataset_doi': '10.24432/C5230J', 'creators': ['John Clore', 'Krzysztof Cios', 'Jon
DeShazo', 'Beata Strack'], 'intro_paper': {'ID': 225, 'type': 'NATIVE', 'title': 'Impact
of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database
Patient Record', 'authors': 'Beata Strack, Jonathan DeShazo, Chris Gennings, Juan Olmo,
Sebastian Ventura, Krzysztof Cios, John Clore', 'venue': 'BioMed Research International,
vol. 2014', 'year': 2014, 'journal': None, 'DOI': None, 'URL': 'https://www.hindawi.com/
journals/bmri/2014/781670/', 'sha': None, 'corpus': None, 'arxiv': None, 'mag': None, 'a
cl': None, 'pmid': None, 'pmcid': None}, 'additional_info': {'summary': 'The dataset rep
resents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delive
ry networks. It includes over 50 features representing patient and hospital outcomes. In
formation was extracted from the database for encounters that satisfied the following cr
iteria.\n(1)\tIt is an inpatient encounter (a hospital admission).\n(2)\tIt is a diabeti
c encounter, that is, one during which any kind of diabetes was entered into the system
as a diagnosis.\n(3)\tThe length of stay was at least 1 day and at most 14 days.\n(4)\tL
aboratory tests were performed during the encounter.\n(5)\tMedications were administered
during the encounter.\n\nThe data contains such attributes as patient number, race, gend
er, age, admission type, time in hospital, medical specialty of admitting physician, num
ber of lab tests performed, HbA1c test result, diagnosis, number of medications, diabeti
```

c medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.', 'purpose': None, 'funded_by': None, 'instances_represent': 'The instances represent hospitalized patient records diagnosed with diabetes.', 'recommended_data_splits': 'No recommendation. The standard train-test split could be used. Can use three-way holdout split (i.e., train-validation-test) when doing model selection.', 'sensitive_data': 'Yes. The dataset contains information about the age, gender, and race of the patients.', 'preprocessing_description': None, 'variable_info': 'Detailed description of all the atrributes is provided in Table 1 Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.\n\nhttp://www.hindawi.com/journals/bmri/2014/781670/', 'citation': 'Please cite:\nBeata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.'}}

	name	role	type	demographic
0	encounter_id	ID		None
1	patient_nbr	ID		None
2	race	Feature	Categorical	Race
3	gender	Feature	Categorical	Gender
4	age	Feature	Categorical	Age
5	weight	Feature	Categorical	None
6	admission_type_id	Feature	Categorical	None
7	discharge_disposition_id	Feature	Categorical	None
8	admission_source_id	Feature	Categorical	None
9	time_in_hospital	Feature	Integer	None
10	payer_code	Feature	Categorical	None
11	medical_specialty	Feature	Categorical	None
12	num_lab_procedures	Feature	Integer	None
13	num_procedures	Feature	Integer	None
14	num_medications	Feature	Integer	None
15	number_outpatient	Feature	Integer	None
16	number_emergency	Feature	Integer	None
17	number_inpatient	Feature	Integer	None
18	diag_1	Feature	Categorical	None
19	diag_2	Feature	Categorical	None
20	diag_3	Feature	Categorical	None
21	number_diagnoses	Feature	Integer	None
22	max_glu_serum	Feature	Categorical	None
23	A1Cresult	Feature	Categorical	None
24	metformin	Feature	Categorical	None
25	repaglinide	Feature	Categorical	None
26	nateglinide	Feature	Categorical	None
27	chlorpropamide	Feature	Categorical	None
28	glimepiride	Feature	Categorical	None
29	acetohexamide	Feature	Categorical	None
30	glipizide	Feature	Categorical	None
31	glyburide	Feature	Categorical	None
32	tolbutamide	Feature	Categorical	None
33	pioglitazone	Feature	Categorical	None
34	rosiglitazone	Feature	Categorical	None
35	acarbose	Feature	Categorical	None
36	miglitol	Feature	Categorical	None
37	troglitazone	Feature	Categorical	None
38	tolazamide	Feature	Categorical	None
39	examide	Feature	Categorical	None
40	citoglipton	Feature	Categorical	None
41	insulin	Feature	Categorical	None

42	glyburide-metformin	Feature	Categorical	None	
43	glipizide-metformin	Feature	Categorical	None	
44	glimepiride-pioglitazone	Feature	Categorical	None	
45	metformin-rosiglitazone	Feature	Categorical	None	
46	metformin-pioglitazone	Feature	Categorical	None	
47	change	Feature	Categorical	None	
48	diabetesMed	Feature	Categorical	None	
49	readmitted	Target	Categorical	None	

	description	units	missing_values
0	Unique identifier of an encounter	None	no
1	Unique identifier of a patient	None	no
2	Values: Caucasian, Asian, African American, Hi...	None	yes
3	Values: male, female, and unknown/invalid	None	no
4	Grouped in 10-year intervals: [0, 10), [10, 20...	None	no
5	Weight in pounds.	None	yes
6	Integer identifier corresponding to 9 distinct...	None	no
7	Integer identifier corresponding to 29 distinc...	None	no
8	Integer identifier corresponding to 21 distinc...	None	no
9	Integer number of days between admission and d...	None	no
10	Integer identifier corresponding to 23 distinc...	None	yes
11	Integer identifier of a specialty of the admit...	None	yes
12	Number of lab tests performed during the encou...	None	no
13	Number of procedures (other than lab tests) pe...	None	no
14	Number of distinct generic names administered ...	None	no
15	Number of outpatient visits of the patient in ...	None	no
16	Number of emergency visits of the patient in t...	None	no
17	Number of inpatient visits of the patient in t...	None	no
18	The primary diagnosis (coded as first three di...	None	yes
19	Secondary diagnosis (coded as first three digi...	None	yes
20	Additional secondary diagnosis (coded as first...	None	yes
21	Number of diagnoses entered to the system	None	no
22	Indicates the range of the result or if the te...	None	no
23	Indicates the range of the result or if the te...	None	no
24	The feature indicates whether the drug was pre...	None	no
25	The feature indicates whether the drug was pre...	None	no
26	The feature indicates whether the drug was pre...	None	no
27	The feature indicates whether the drug was pre...	None	no
28	The feature indicates whether the drug was pre...	None	no
29	The feature indicates whether the drug was pre...	None	no
30	The feature indicates whether the drug was pre...	None	no
31	The feature indicates whether the drug was pre...	None	no
32	The feature indicates whether the drug was pre...	None	no
33	The feature indicates whether the drug was pre...	None	no
34	The feature indicates whether the drug was pre...	None	no
35	The feature indicates whether the drug was pre...	None	no
36	The feature indicates whether the drug was pre...	None	no
37	The feature indicates whether the drug was pre...	None	no
38	The feature indicates whether the drug was pre...	None	no
39	The feature indicates whether the drug was pre...	None	no
40	The feature indicates whether the drug was pre...	None	no
41	The feature indicates whether the drug was pre...	None	no
42	The feature indicates whether the drug was pre...	None	no
43	The feature indicates whether the drug was pre...	None	no
44	The feature indicates whether the drug was pre...	None	no
45	The feature indicates whether the drug was pre...	None	no
46	The feature indicates whether the drug was pre...	None	no
47	Indicates if there was a change in diabetic me...	None	no

```
48  Indicates if there was any diabetic medication...  None             no
49  Days to inpatient readmission. Values: <30 if ...  None             no
```

import pandas as pd y.info()

In [ ]:

```
X.info()
```

In [ ]:

```
X.isnull()
```

In [ ]:

```
X.isnull().sum()
```

In [ ]:

```
import numpy as np
import pandas as pd
X.fillna("NAN")
```

In [ ]:

```
X.isnull().sum()
```

In [ ]:

```
X.fillna(0, inplace=False)
```

In [ ]:

```
X.isnull().sum()
```

In [ ]:

```
X.drop('weight',axis=1).head()
```

In [ ]:

```
X.columns
```

In [ ]:

```
X.columns
```

In [ ]:

```
X.columns=X.columns.str.lower()
X.columns=X.columns.str.replace(' ','_')
```

In [ ]:

```
X.columns
```

In [ ]:

```
y.isnull().sum().head()
```

In [ ]:

```
y.head()
```

In [ ]:

```
mapping = {'<30': 1, '>30': 0, 'NO': 0}
y['readmission_binary'] = y['readmitted'].map(mapping)
```

In [ ]:

```
y.head(20)
```

In [ ]:

```
X['time_in_hospital'] = X['time_in_hospital'].replace(0, 1)
X['acuity_score'] = (X['num_lab_procedures'] +
                     X['num_medications'] +
                       X['number_diagnoses']) /X['time_in_hospital']
X['acuity_score'] =X['acuity_score'].round(2)
```

In [ ]:

```
print(X[['payer_code', 'acuity_score', 'time_in_hospital']].sort_values(by='acuity_score
```

In [ ]:

```
X.head()
```

In [ ]:

```
AVG_READMISSION_COST = 13500
X['historical_loss'] = y['readmission_binary'] * AVG_READMISSION_COST
dept_loss = X.groupby('medical_specialty')['historical_loss'].sum().sort_values(ascendin

print("Top 10 Cost-Draining Departments:")
print(dept_loss)
```

In [ ]:

```
X.head()
```

In [ ]:

```
import numpy as np


def calculate_nurse_demand(acuity):
    if acuity > 7:
        return 1/3
    else:
        return 1/5


X['nurse_required_count'] = X['acuity_score'].apply(calculate_nurse_demand)


staffing_needs = X.groupby('medical_specialty').agg({
    'payer_code': 'count',
    'nurse_required_count': 'sum',
    'acuity_score': 'mean'
}).rename(columns={'payer_code': 'Total_Patients', 'nurse_required_count': 'Nurses_Neede


staffing_needs['Nurses_Needed'] = np.ceil(staffing_needs['Nurses_Needed']).astype(int)

print(staffing_needs.sort_values(by='Nurses_Needed', ascending=False).head(10))
```

In [ ]:

```
X.head()
```

In [ ]:

```
# Use Parquet for faster loading and smaller file size

# Or standard CSV
X.to_csv('hospital_data_cleaned.csv', index=False)
```