DATA PREPARATION PLAN

OVERVIEW

- Data quality issues and actions
- Feature engineering decisions
- Feature selection decisions
- Dataset partitioning decisions
- Final dataset summary

DATA PREPARATION PLAN

DATA QUALITY ISSUES AND ACTIONS

Missing variables

Variables	Actions	Description
Number of Prior Visits	Impute with mode	Important features, discrete variables
Medications Prescribed	Impute with mode	Not highly correlated with other variables, discrete variables

USC Viterbi School of Engineering

DATA PREPARATION PLAN

FEATURE ENGINEERING DECISIONS

Variables	Actions	Description
Exercise Frequency	Ordinal encoding	None = 0, Occasional = 1, Regular = 2
BMI	BMI category	Underweight (<18.5), Normal (18.5–24.9), Overweight (25–29.9), Obese (≥30)
Age	Age group	<40, 40–64, 65+
Age	Remove outliers	Remove age > 100
Length of Stay	Log	Heavily skewed

USC Viterbi School of Engineering

DATA PREPARATION PLAN

FEATURE SELECTION DECISIONS

Variables	Actions	Description
Hospital ID	Dropped	no predictive value
Adjusted Weight	Dropped	High correlation with Weight or BMI; redundant
Weight	Dropped	Already reflected in BMI

DATA PREPARATION PLAN
DATASET PARTITIONING DECISIONS

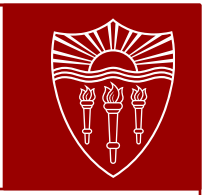
- Perform SMOTE oversampling due to unbalanced dataset

DATA PREPARATION PLAN
FINAL DATASET SUMMARY

Demographic variables

Name	Description
Age	Patient age
Age_Group_65+	65 <= age
Age_Group_<40	Age < 40
Gender_Male	Male = 1, Female = 0
Ethnicity_Caucasian	One hot encoding for Ethnicity
Ethnicity_Hispanic	One hot encoding for Ethnicity
Ethnicity_Other	One hot encoding for Ethnicity

DATA PREPARATION PLAN
FINAL DATASET SUMMARY



Clinical variables

Name	Description
Height(m)	Patient height in meters
Smoker	Boolean indicating if patient is a current smoker 1=patient is a current smoker 0=patient is not a current smoker
BMI	Patient Body Mass Index
BMI_Category_Obese	30 <= bmi
BMI_Category_Underweight	bmi < 18.5
BMI_Category_Overweight	25 <= bmi < 30
Has Diabetes	Boolean indicating if patient has diabetes 1=patient has diabetes 0=patient does not have diabetes
Has Hypertension	Boolean indicating if patient has hypertension 1=patient has hypertension 0=patient does not have hypertension

DATA PREPARATION PLAN
FINAL DATASET SUMMARY

Behavioral variables

Name	Description
Exercise_Encoded	None = 0, Occasional = 1, Regular = 2
Diet Type_High-fat	One hot encoding for Diet Type
Diet Type_Other	One hot encoding for Diet Type
Diet Type_Vegetarian	One hot encoding for Diet Type

DATA PREPARATION PLAN
FINAL DATASET SUMMARY

Treatment-related variables

Name	Description
Number of Prior Visits	Number of previous hospitalizations of the patient
Medications Prescribed	Number of different prescription medications patient is currently taking
LOS_Log	Log of Length of Stay
Type of Treatment_Minor Surgery	One hot encoding of Type of Treatment
Type of Treatment_None	One hot encoding of Type of Treatment
Type of Treatment_Other Treatment	One hot encoding of Type of Treatment