Hospital Readmission Prediction for Diabetic Patients

Table and Figures Appndix

List of Tables

List of Figures

Column Name	Dtype	Missing	Unique
encounter_id	int64	0	101766
patient_nbr	int64	0	71518
admission_type_id	int64	0	8
discharge_disposition_id	int64	0	26
admission_source_id	int64	0	17
time_in_hospital	int64	0	14
num_lab_procedures	int64	0	118
num_procedures	int64	0	7
num_medications	int64	0	75
number_outpatient	int64	0	39
number_emergency	int64	0	33
number_inpatient	int64	0	21
number_diagnoses	int64	0	16

Table 1. Numerical features in the full dataset with missing and unique value counts.

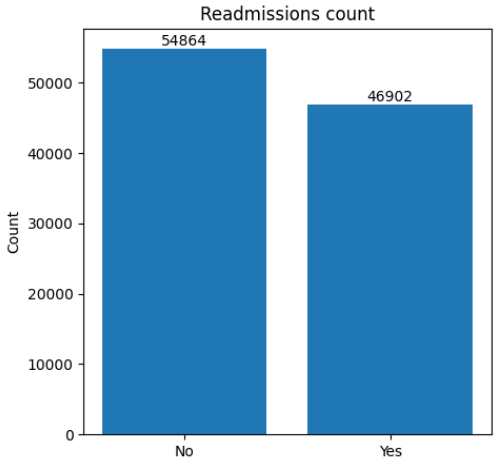

Figure 1. Distribution of readmissions label in the full dataset.

References

[1] CMS.gov. ICD-9-CM Diagnosis and Procedure Codes: Abbreviated and Full Code Titles. 1, 12

Column Name	Dtype	Missing	Unique
race	object	2273	5
gender	object	0	3
age	object	0	10
weight	object	98569	9
payer_code	object	40256	17
medical_specialty	object	49949	72
diag_1	object	21	716
diag_2	object	358	748
diag_3	object	1423	789
max_glu_serum	object	96420	3
A1Cresult	object	84748	3
metformin	object	0	4
repaglinide	object	0	4
nateglinide	object	0	4
chlorpropamide	object	0	4
glimepiride	object	0	4
acetohexamide	object	0	2
glipizide	object	0	4
glyburide	object	0	4
tolbutamide	object	0	2
pioglitazone	object	0	4
rosiglitazone	object	0	4
acarbose	object	0	4
miglitol	object	0	4
troglitazone	object	0	2
tolazamide	object	0	3
examide	object	0	1
citoglipton	object	0	1
insulin	object	0	4
glyburide-metformin	object	0	4
glipizide-metformin	object	0	2
glimepiride-pioglitazone	object	0	2
metformin-rosiglitazone	object	0	2
metformin-pioglitazone	object	0	2
change	object	0	2
diabetesMed	object	0	2
readmitted	object	0	3

Table 2. String features in the full dataset with missing and unique value counts.

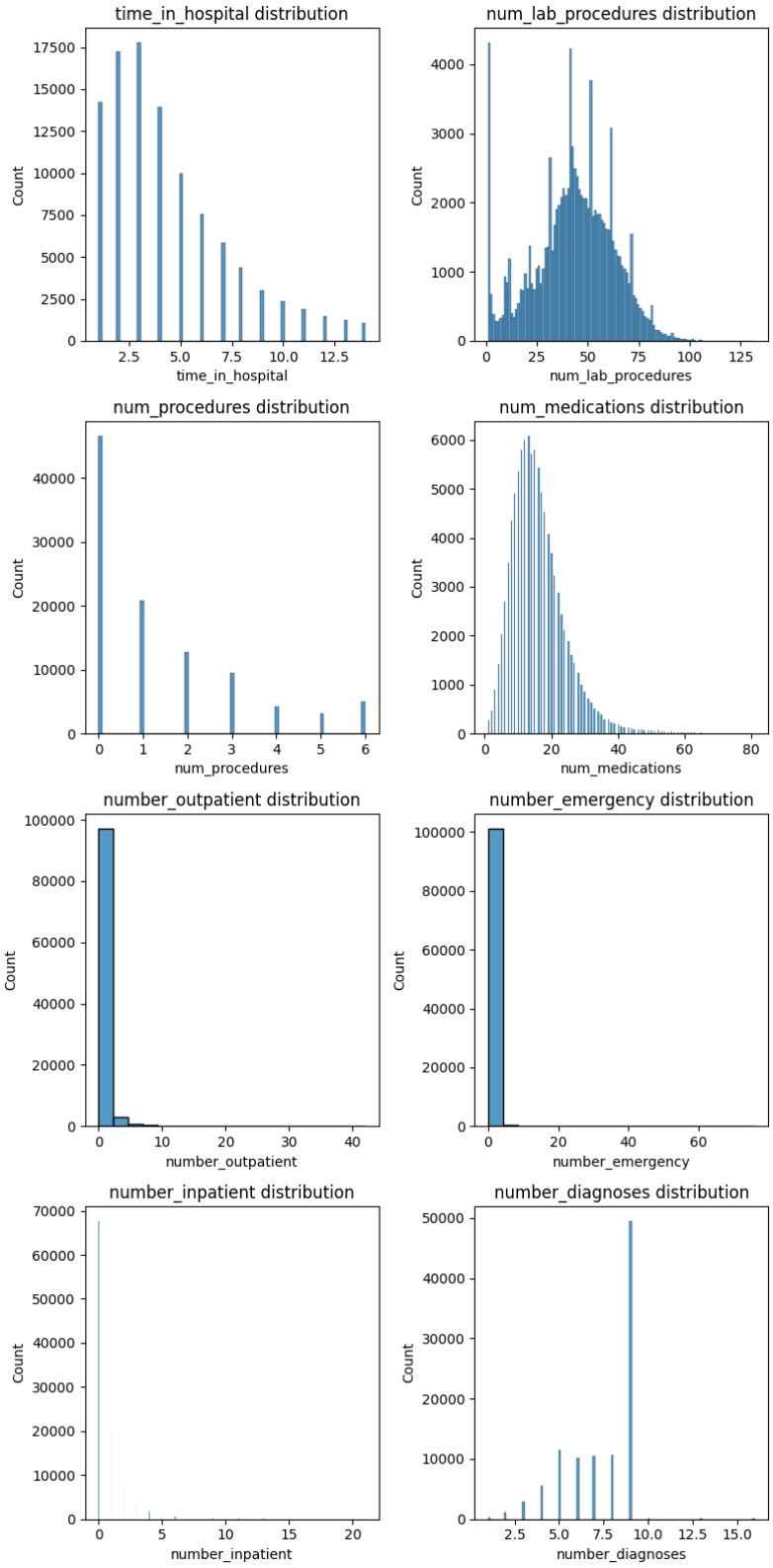

Figure 2. Numerical features distribution in the full dataset.

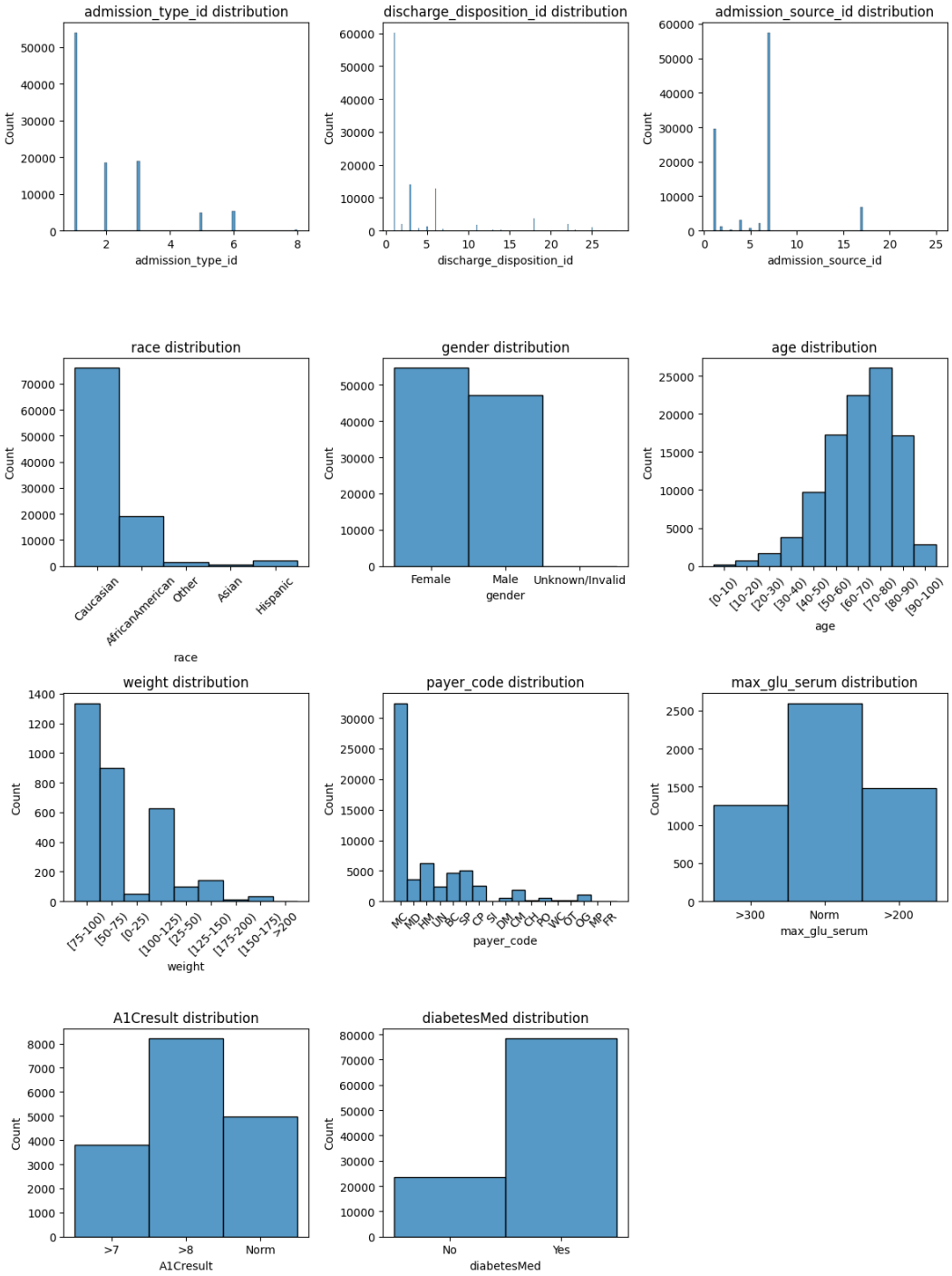

Figure 3. Regular categorical features distribution in the full dataset, excluding high-cardinality and medication ones.

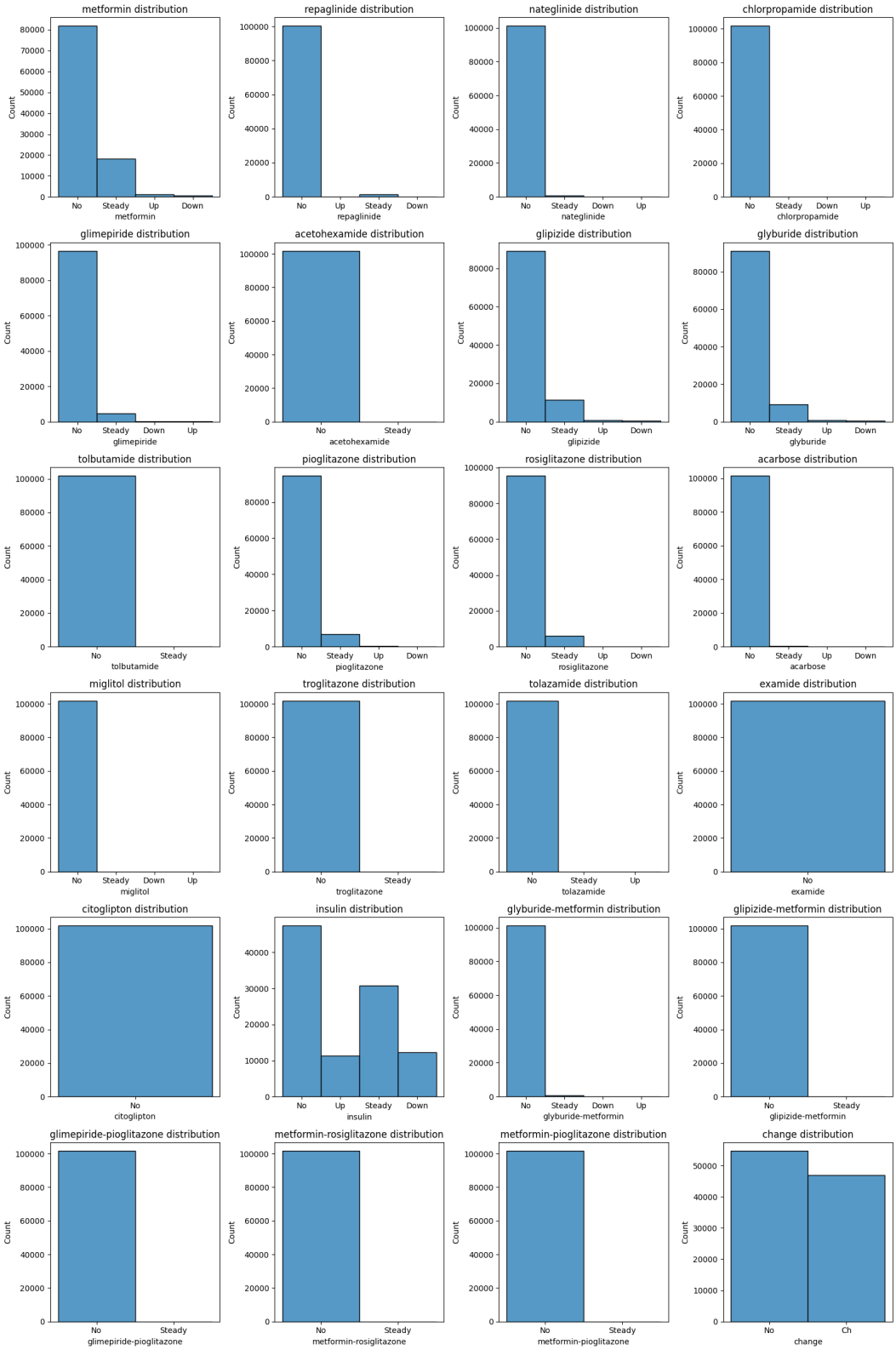

Figure 4. Medication features distribution in the full dataset.

diag_1 code	Value count
428	6862
414	6581
786	4016
410	3614
486	3508

Table 3. Value counts of top 5 diag_1 codes by frequency.

diag_2 code	Value count
276	6752
428	6662
250	6071
427	5036
401	3736

Table 4. Value counts of top 5 diag_2 codes by frequency.

diag_3 code	Value count
250	11555
401	8289
276	5175
428	4577
427	3955

Table 5. Value counts of top 5 diag_3 codes by frequency.

medical_specialty	Value count
InternalMedicine	14635
Emergency/Trauma	7565
Family/GeneralPractice	7440
Cardiology	5352
Surgery-General	3099

Table 6. Value counts of top 5 medical_specialty codes by frequency.

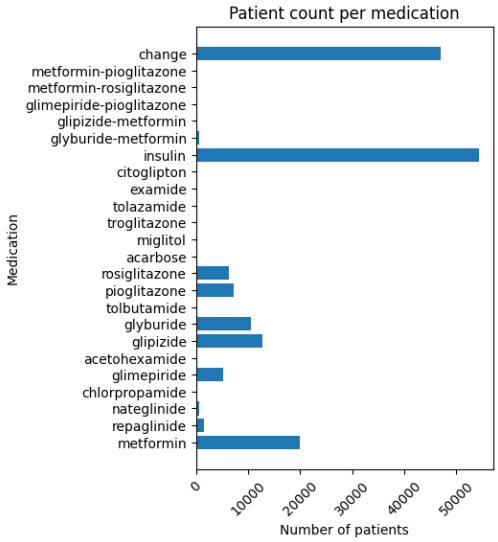

Figure 5. Counting number of patients each medication was taken by.

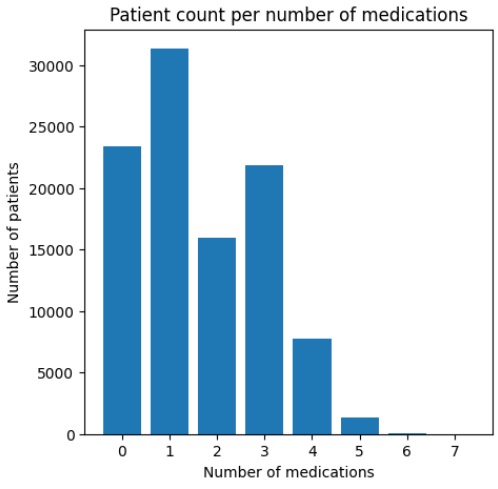

Figure 6. Distribution of patients per number of medications taken.

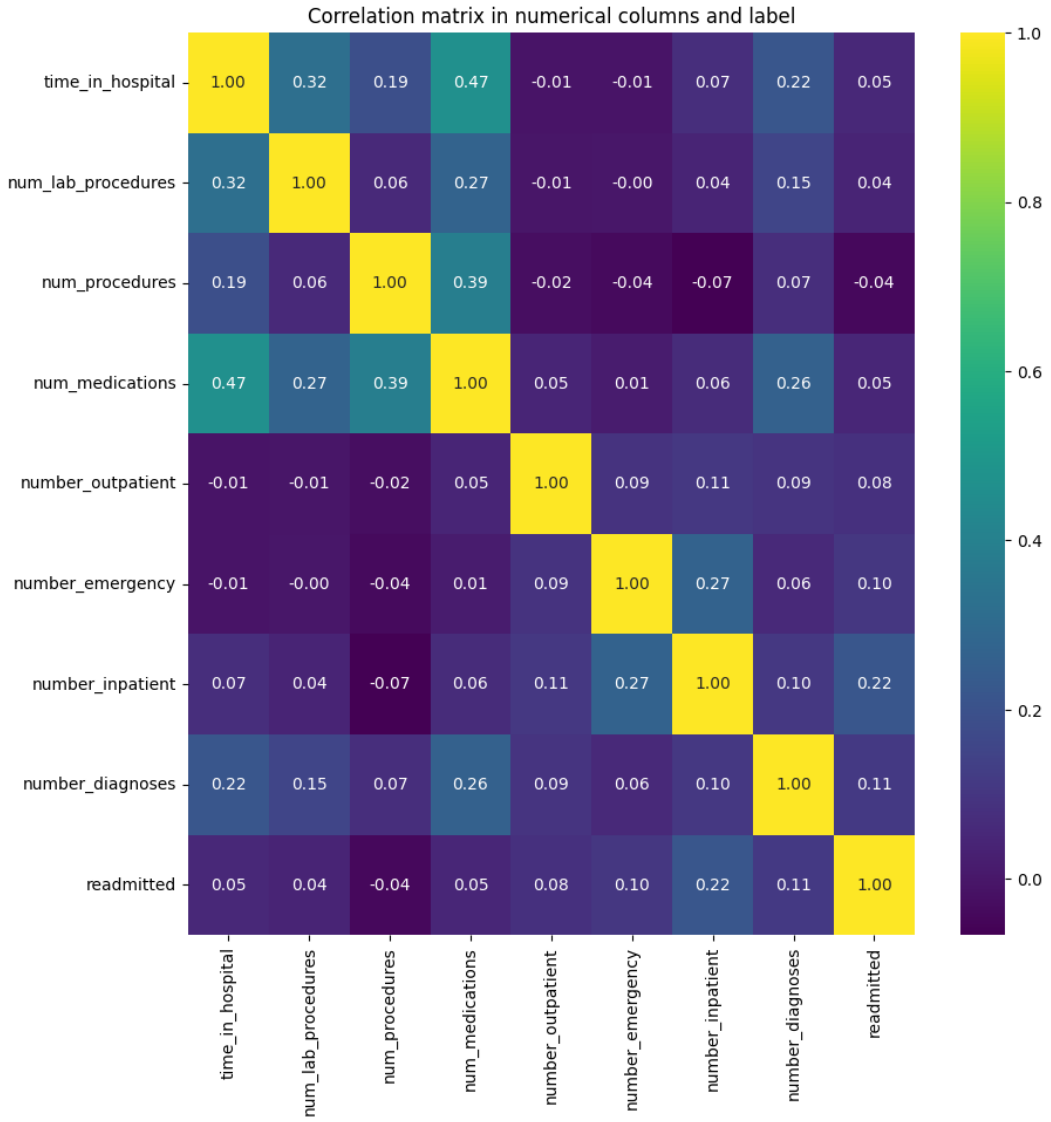

Figure 7. Correlation matrix of numerical features with the readmission label.

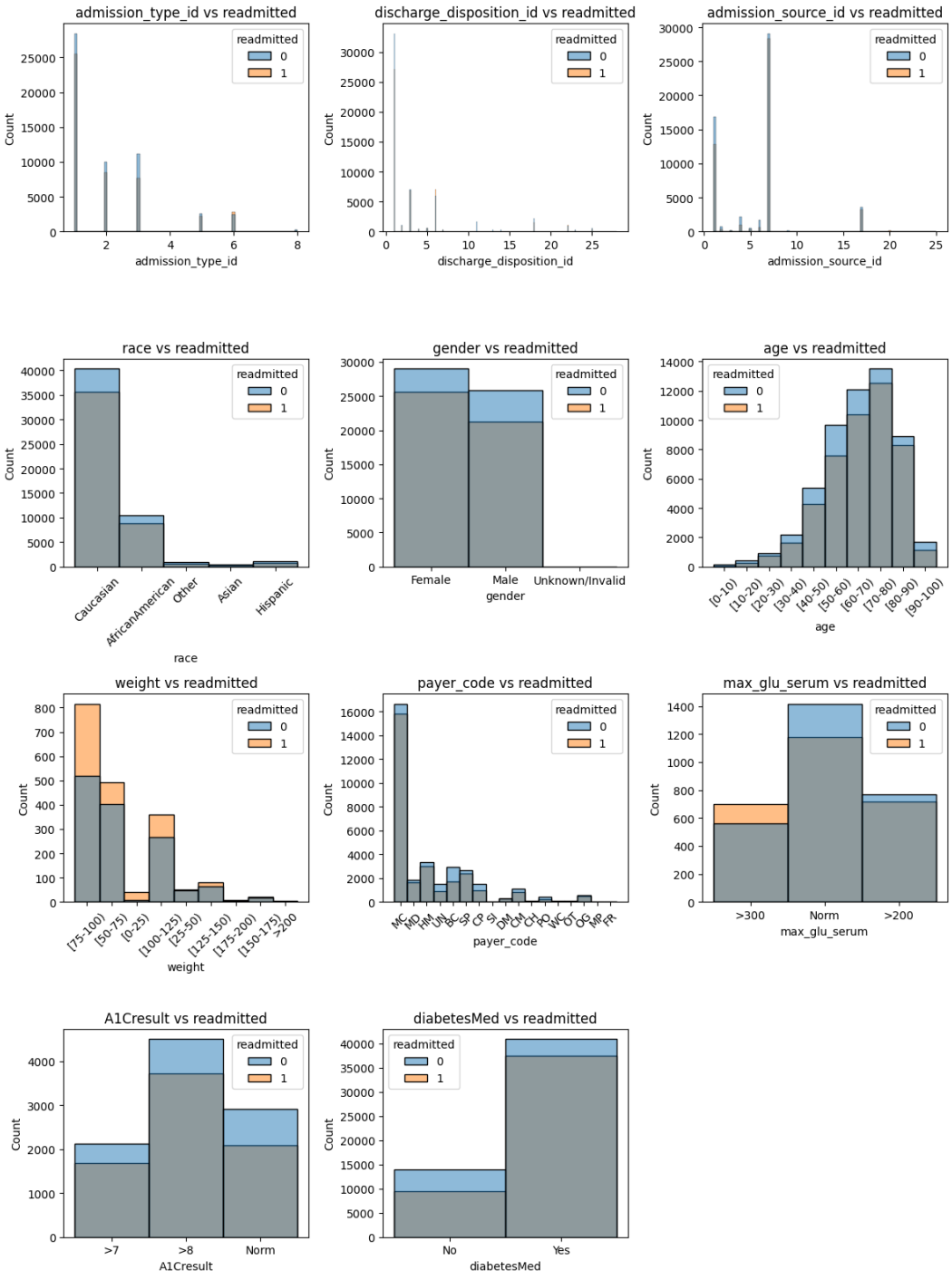

Figure 8. Overlapping distribution of regular categorical features with the readmission label as hue. Gray represents amount of data points both conditions has, whereas the surpassing color represents the amount of entries one has over the other.

Column Name	Dtype	Missing	Missing Ratio	Unique
race	object	1097	0.0215	5
weight	object	49285	0.9686	9
payer_code	object	20048	0.3940	16
medical_specialty	object	25072	0.4927	66
diag_1	object	14	0.0002	650
diag_2	object	195	0.0038	656
diag_3	object	720	0.0141	699
max_glu_serum	object	482	0.9474	3
A1Cresult	object	42414	0.8335	3

Table 7. Columns with missing values in the training dataset, their missing ratio and number of unique values.

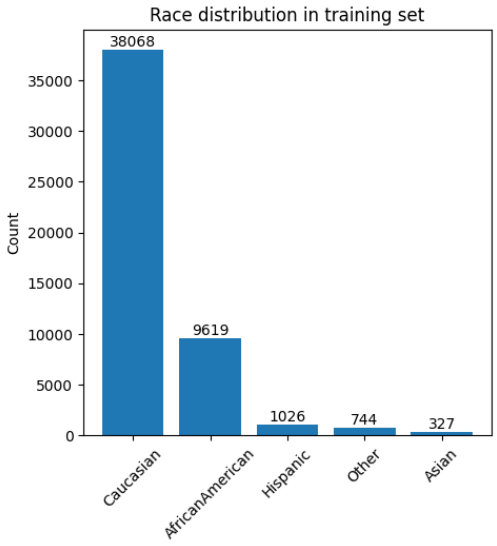

Figure 9. Race distribution in training dataset.

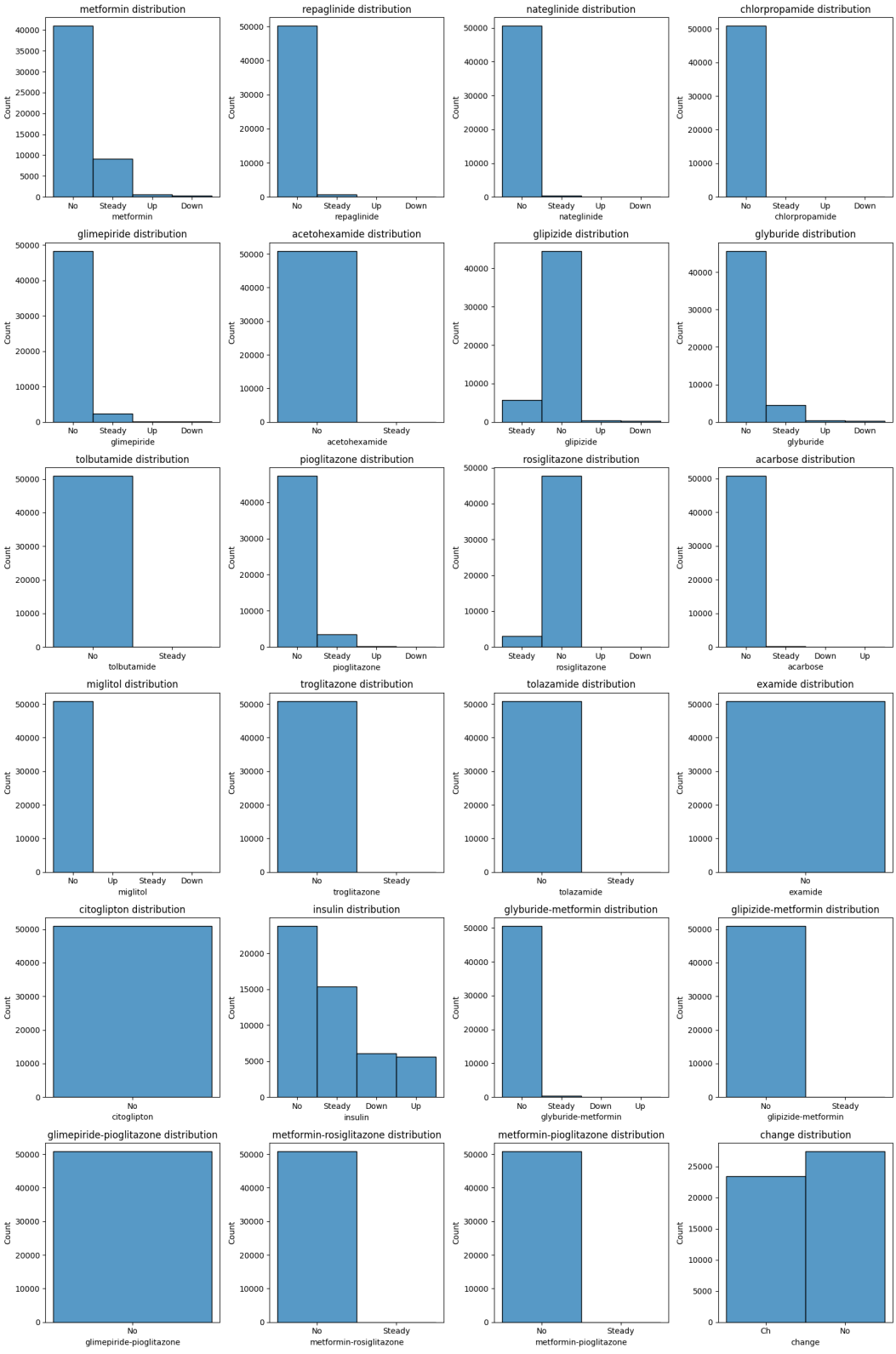

Figure 10. Medication features distribution in the training dataset.

ICD-9 Range	Category Description
001-139	Infectious and parasitic diseases
140-239	Neoplasms
240-279	Endocrine, nutritional and metabolic diseases, and immunity disorders
280-289	Diseases of the blood and blood-forming organs
290-319	Mental disorders
320-389	Diseases of the nervous system and sense organs
390-459	Diseases of the circulatory system
460-519	Diseases of the respiratory system
520-579	Diseases of the digestive system
580-629	Diseases of the genitourinary system
630-679	Complications of pregnancy, childbirth, and the puerperium
680-709	Diseases of the skin and subcutaneous tissue
710-739	Diseases of the musculoskeletal system and connective tissue
740-759	Congenital anomalies
760-779	Certain conditions originating in the perinatal period
780-799	Symptoms, signs, and ill-defined conditions
800-999	Injury and poisoning
E	External causes of injury
V	Supplemental classification
missing	missing

Table 8. ICD-9 code ranges and their corresponding category descriptions. Source: [1], link 1 and link 2.

Column Group	Columns
Columns to drop (7)	weight, payer_code, medical_specialty, max_glu_serum, A1Cresult, examide, citoglipton
Columns to binary encode (22)	21 medication columns, diabetesMed
Columns to one-hot encode (9)	gender, race, insulin, diag_1, diag_2, diag_3, admission_type_id, discharge_disposition_id, admission_source_id
Columns to ordinal encode (1)	age
Columns to scale (8)	time_in_hospital, num_lab_procedures, num_procedures, num_medications, number_outpatient, number_emergency, number_inpatient, number_diagnoses

Table 9. After imputing and mapping, columns to transform and their respective transformation in the training dataset.

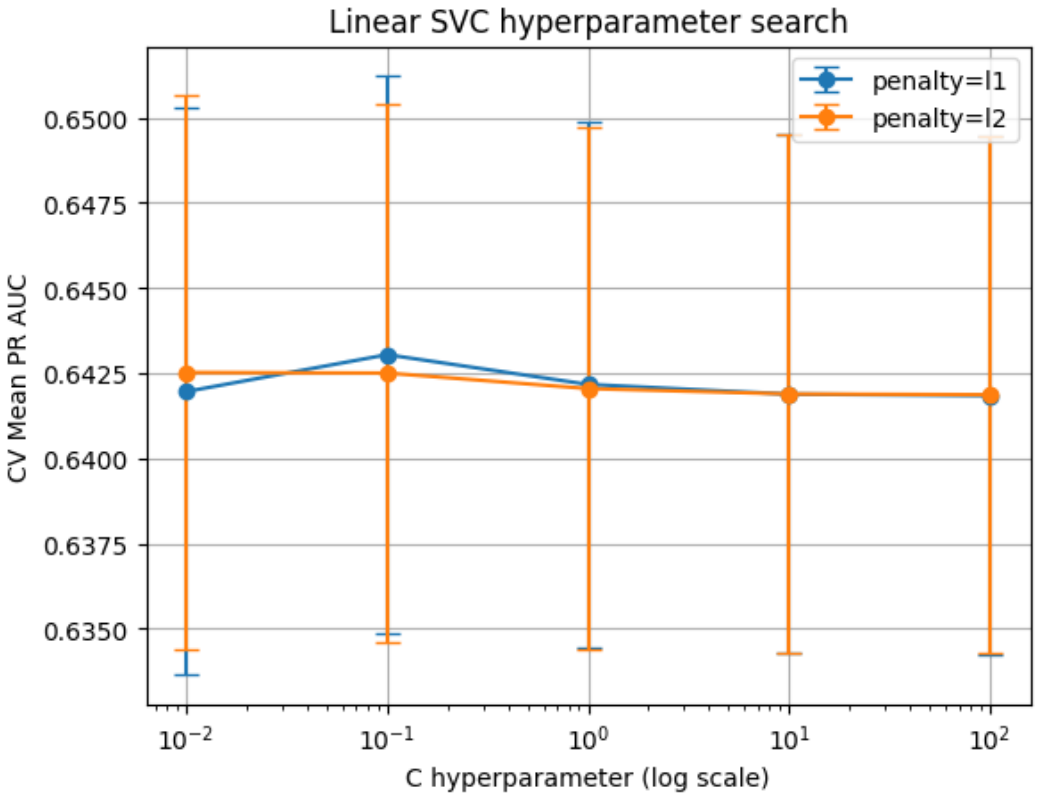

Figure 11. Hyperparameter optimisation visualisation for the LinearSVC model. C shown with log scale in x-axis while penalty shown with colour.

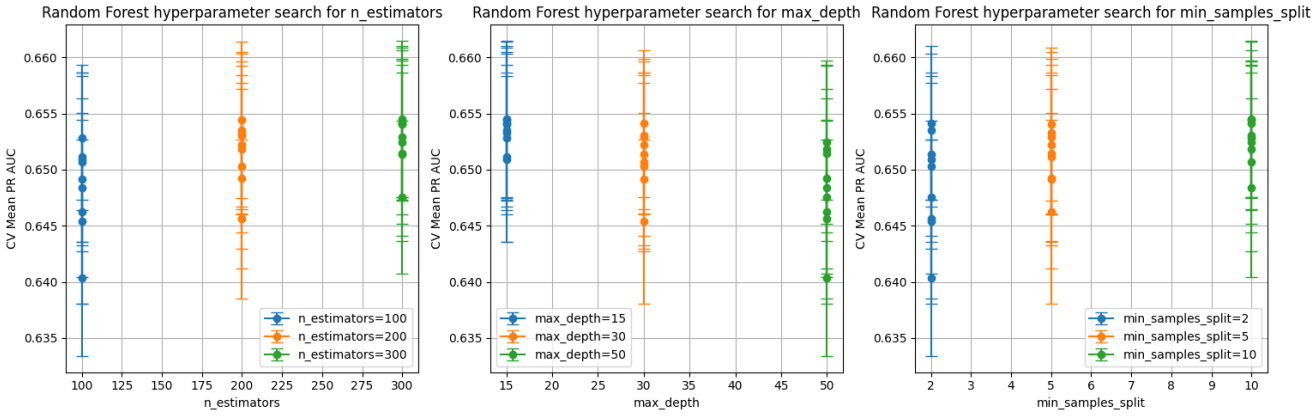

Figure 12. Full hyperparameter optimisation visualisation for the Random Forest model. Each hyperparameter shown in a separate subplot, diffusing the combination with other hyperparameters into the configs of this subplot.

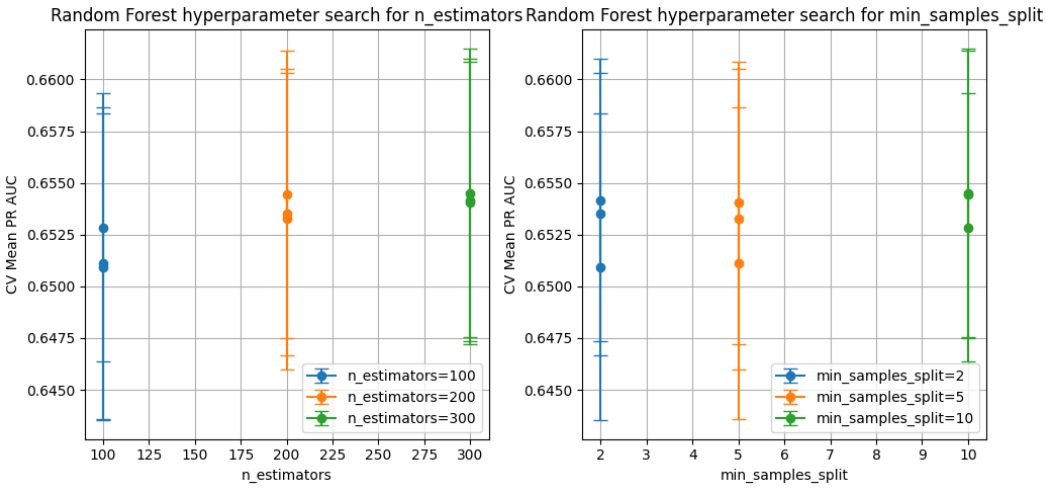

Figure 13. Subset hyperparameter optimisation visualisation for the Random Forest model, fixing max_depth to 15.

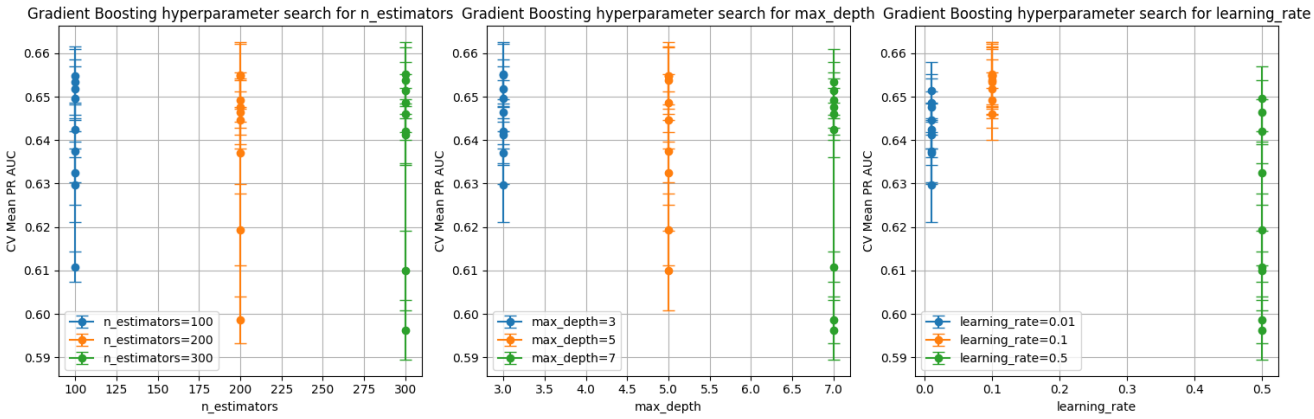

Figure 14. Full hyperparameter optimisation visualisation for the Gradient Boosting model. Each hyperparameter shown in a separate subplot, diffusing the combination with other hyperparameters into the configs of this subplot.

Validation folds	Linear SVC	Random Forest	Gradient Boosting
Accuracy	63.77 ± 0.49%	**71.25 ± 0.57%**	65.76 ± 0.43%
Precision	64.30 ± 0.99%	**74.41 ± 1.08%**	65.59 ± 0.82%
Recall	47.74 ± 0.83%	**57.16 ± 0.81%**	53.76 ± 0.47%
F1	54.79 ± 0.83%	**64.65 ± 0.78%**	59.09 ± 0.53%
ROC AUC	69.02 ± 0.42%	**80.20 ± 0.51%**	71.80 ± 0.39%
PR AUC	64.67 ± 0.58%	**78.30 ± 0.70%**	67.99 ± 0.61%

Table 10. Mean and standard deviation of the evaluation metrics for the best models (hyperparameter optimised) on the validation folds.

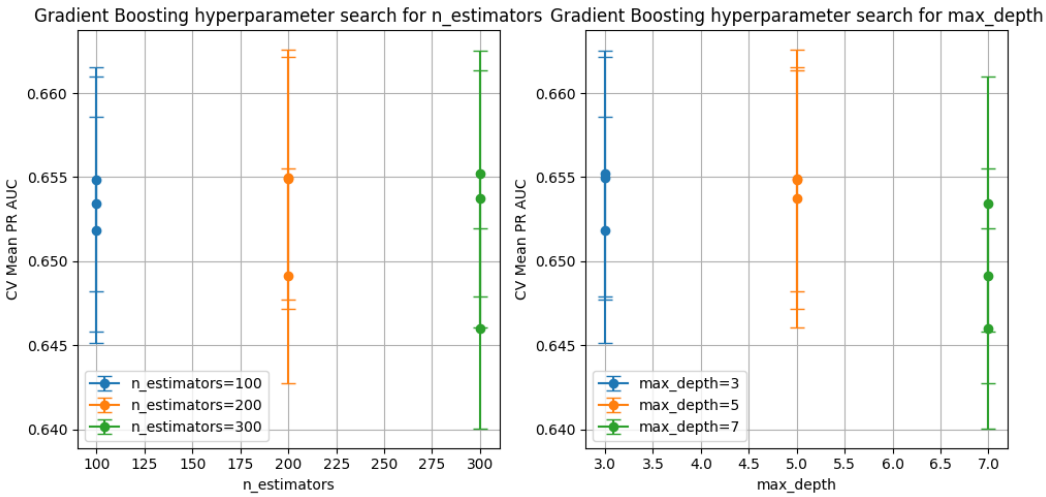

Figure 15. Subset hyperparameter optimisation visualisation for the Gradient Boosting model, fixing learning_rate to 0.1.

Test folds	Linear SVC	Random Forest	Gradient Boosting
Accuracy	63.30 ± 0.43%	64.18 ± 0.46%	**64.29 ± 0.45%**
Precision	63.75 ± 0.64%	**64.58 ± 0.66%**	63.71 ± 0.61%
Recall	47.61 ± 0.71%	49.71 ± 0.70%	**52.71 ± 0.65%**
F1	54.51 ± 0.61%	56.17 ± 0.62%	**57.69 ± 0.56%**
ROC AUC	68.58 ± 0.55%	69.63 ± 0.53%	**69.78 ± 0.53%**
PR AUC	64.67 ± 0.67%	65.56 ± 0.68%	**65.94 ± 0.68%**

Table 11. Mean and standard deviation of the evaluation metrics for the best models (hyperparameter optimised) on the test folds.

Feature	Meaning
discharge_disposition_id_11	Expired
discharge_disposition_id_13	Hospice / home
discharge_disposition_id_14	Hospice / medical facility
discharge_disposition_id_15	Discharged/transferred within this institution to Medicare approved swing bed
discharge_disposition_id_28	Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
admission_source_id_4	Transfer from a hospital
admission_source_id_5	Transfer from a Skilled Nursing Facility (SNF)
admission_type_id_6	NULL
admission_type_id_7	Trauma Center
diag_2	Complications of pregnancy, childbirth, and the puerperium

Table 12. Feature names translation that appear in top 10 features, according to mapping tables.

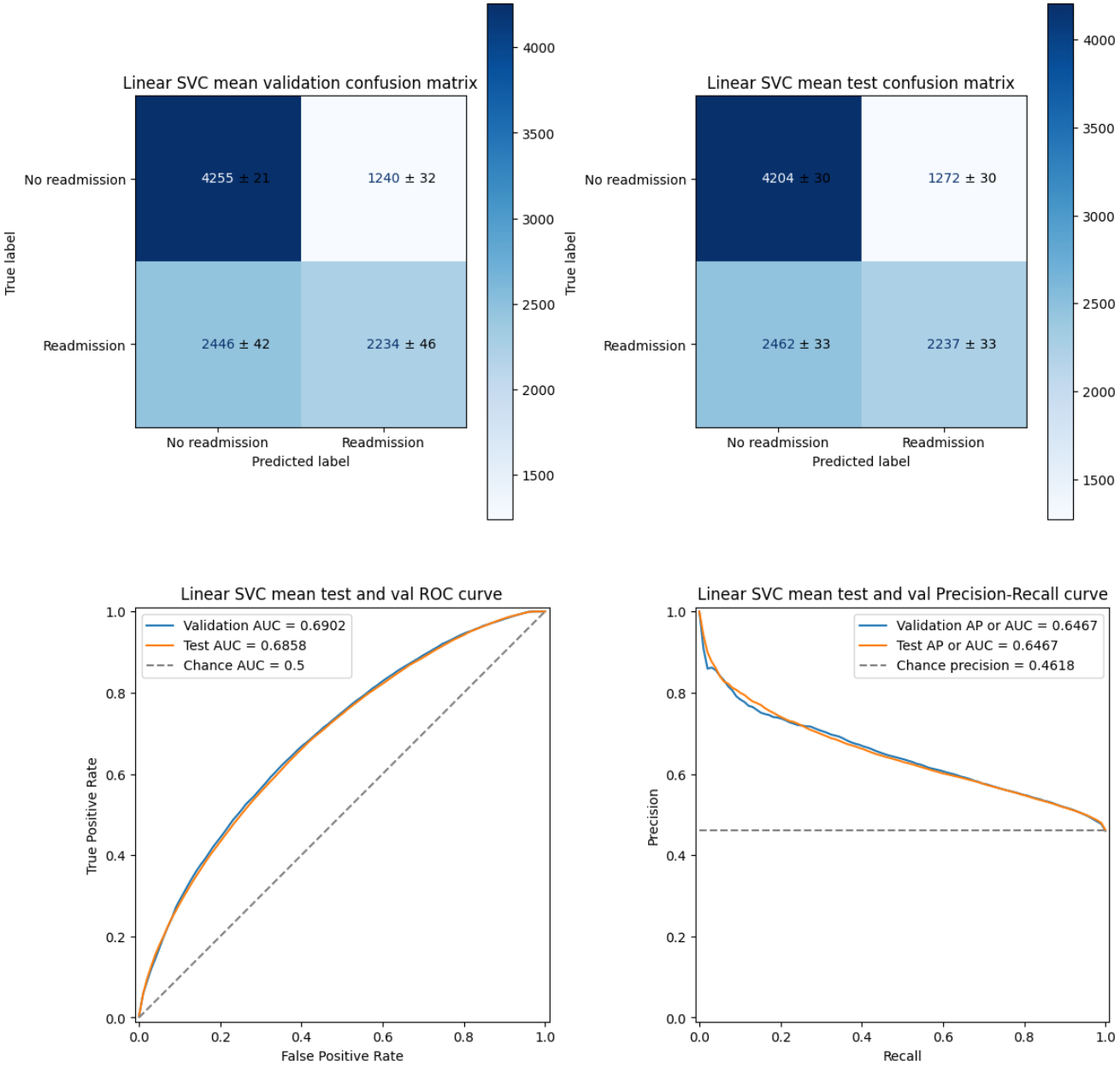

Figure 16. Validation confusion matrix (top-left), test confusion matrix (top-right), ROC curves (bottom-left) and PR curves (bottom-right) for the Linear SVC model, at validation and test folds. AUC: Area Under the Curve, AP: Average Precision.

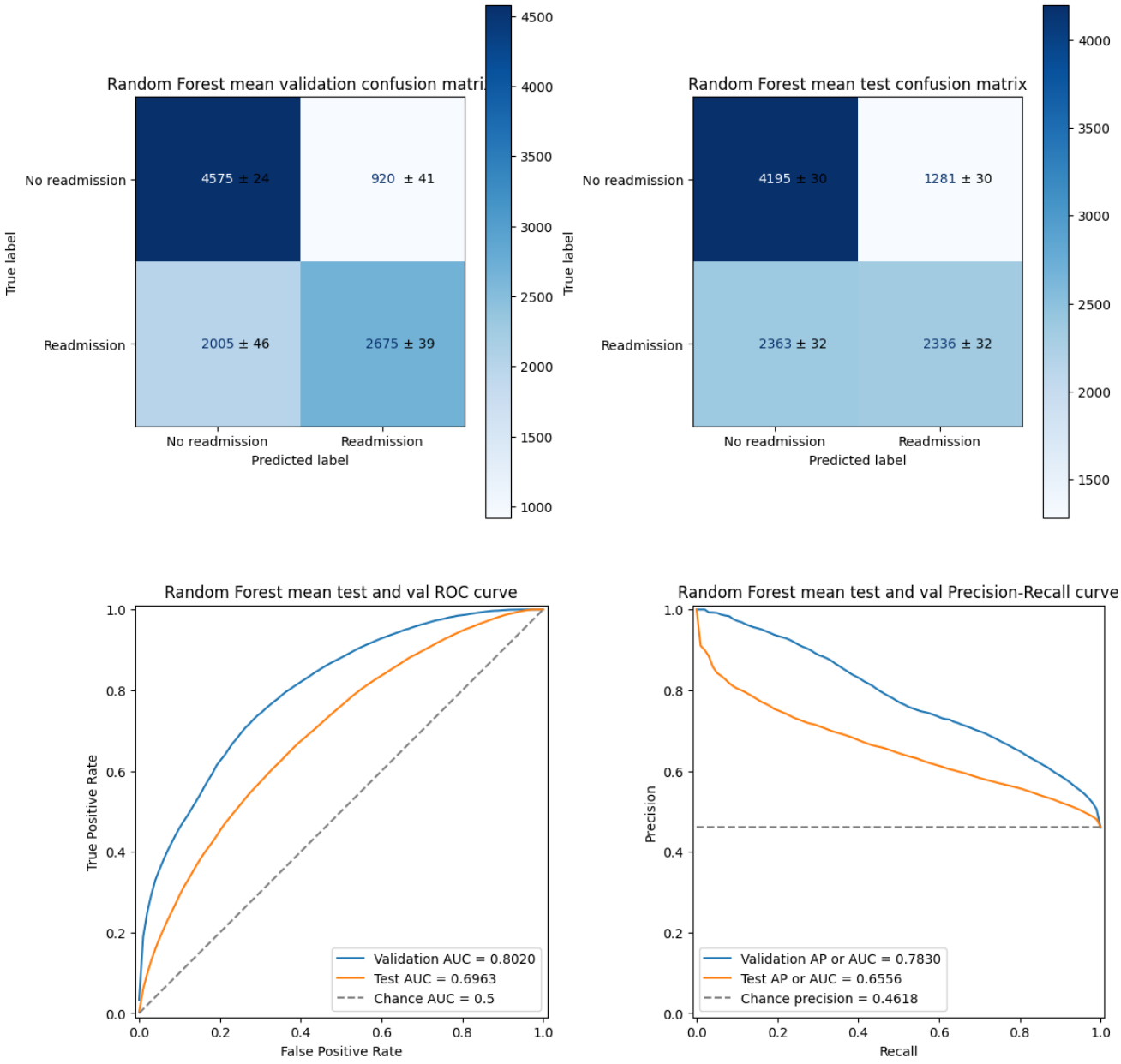

Figure 17. Validation confusion matrix (top-left), test confusion matrix (top-right), ROC curves (bottom-left) and PR curves (bottom-right) for the Random Forest model, at validation and test folds.

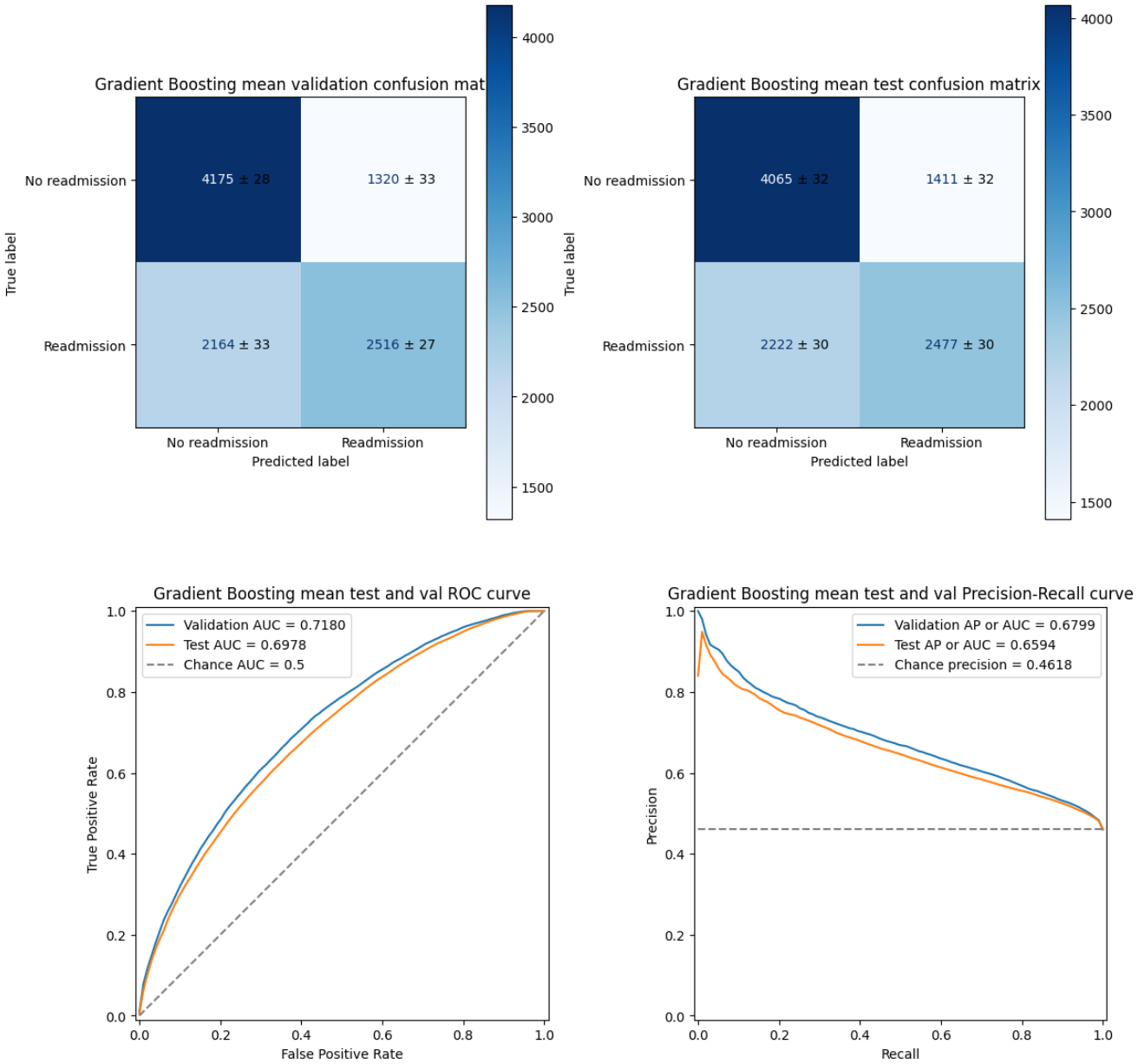

Figure 18. Validation confusion matrix (top-left), test confusion matrix (top-right), ROC curves (bottom-left) and PR curves (bottom-right) for the Gradient Boosting model, at validation and test folds.

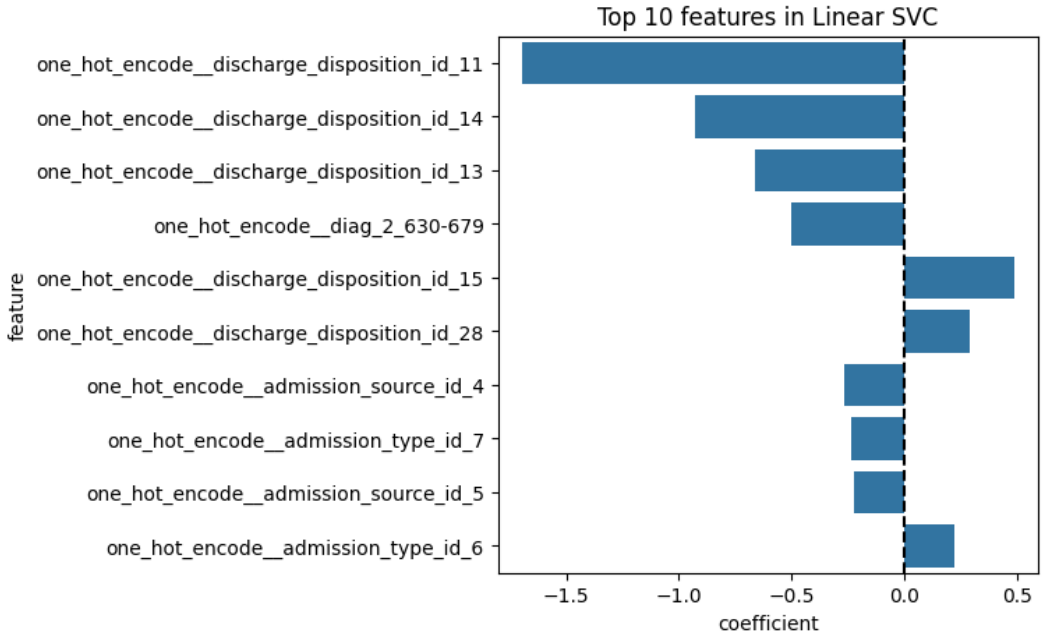

Figure 19. Top 10 features ranked by absolute coefficients in the Linear SVC model.

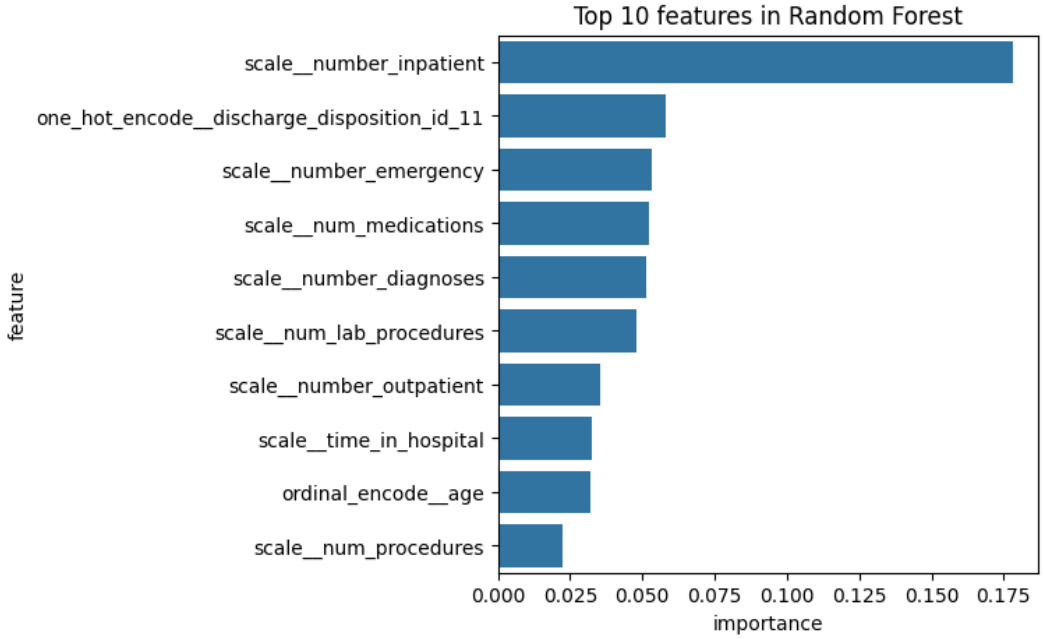

Figure 20. Top 10 features ranked by absolute coefficients in the Random Forest model.

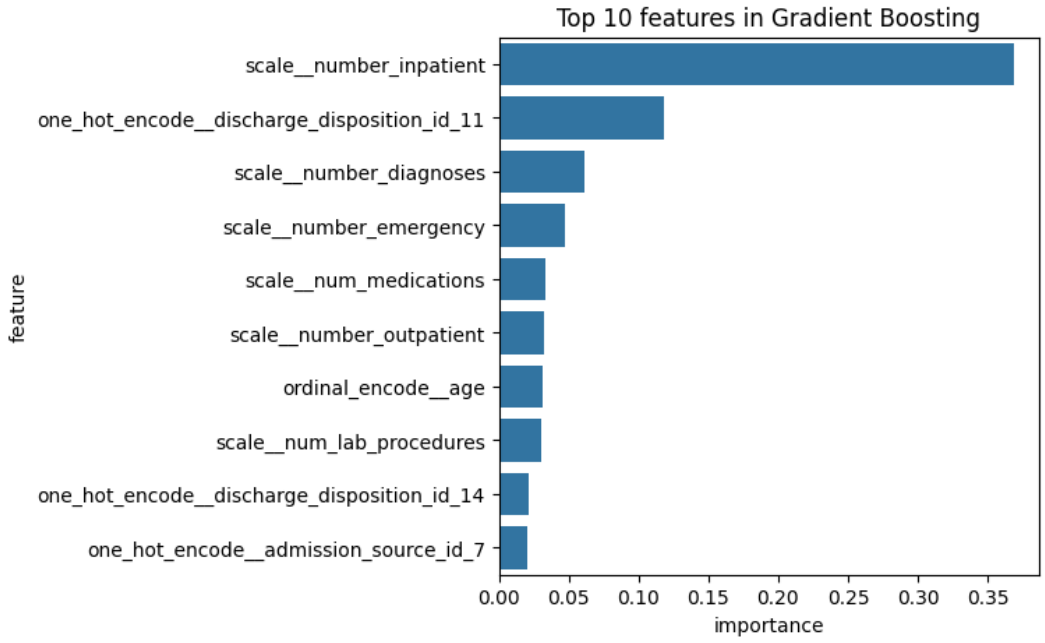

Figure 21. Top 10 features ranked by absolute coefficients in the Gradient Boosting model.

Validation folds	Gradient Boosting	With RFE
Accuracy	65.76 ± 0.43%	63.00 ± 0.38%
Precision	65.59 ± 0.82%	64.53 ± 0.59%
Recall	53.76 ± 0.47%	43.42 ± 0.56%
F1	59.09 ± 0.53%	51.91 ± 0.49%
ROC AUC	71.80 ± 0.39%	66.91 ± 0.51%
PR AUC	67.99 ± 0.61%	63.60 ± 0.63%

Table 13. Mean and standard deviation comparison of Gradient Boosting model with and without Recursive Feature Elimination (RFE) on the validation folds.

Test folds	Gradient Boosting	With RFE
Accuracy	64.29 ± 0.45%	62.60 ± 0.42%
Precision	63.71 ± 0.61%	63.96 ± 0.66%
Recall	52.71 ± 0.65%	43.56 ± 0.69%
F1	57.69 ± 0.56%	51.82 ± 0.63%
ROC AUC	69.78 ± 0.53%	66.63 ± 0.52%
PR AUC	65.94 ± 0.68%	63.60 ± 0.67%

Table 14. Mean and standard deviation comparison of Gradient Boosting model with and without Recursive Feature Elimination (RFE) on the test folds.

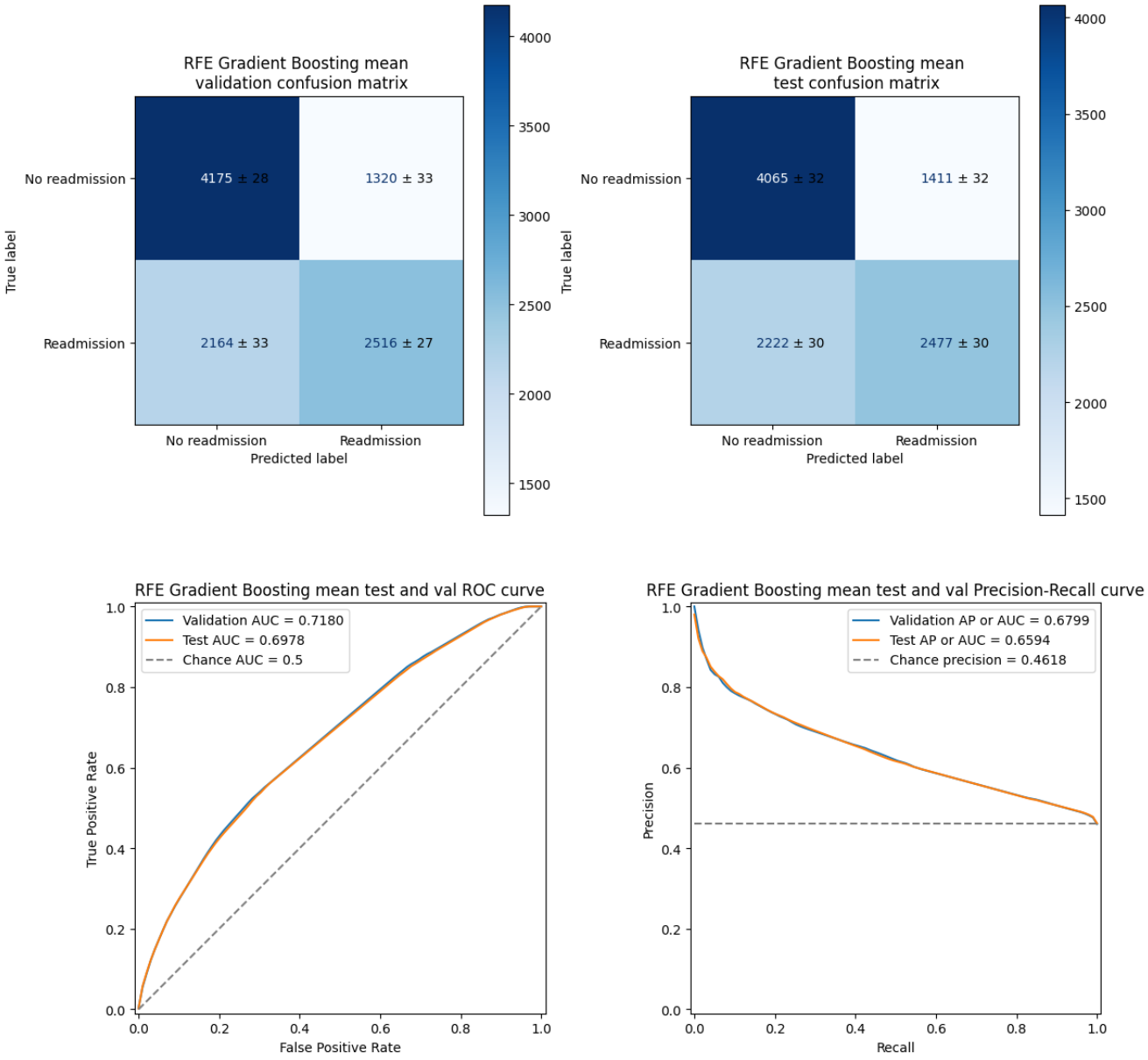

Figure 22. Validation confusion matrix (top-left), test confusion matrix (top-right), ROC curves (bottom-left) and PR curves (bottom-right) for the RFE Gradient Boosting model, at validation and test folds. AUC: Area Under the Curve, AP: Average Precision.