Hospital readmissions prediction coursework

1 Dataset description

The dataset includes ten years (1999-2008) of clinical care records of patients diagnosed with diabetes, compiled from 130 US hospitals and integrated delivery networks.[3] It includes 49 features, with a sample-to-feature ratio of 2035:1. There are eight discrete numeric features, the rest are categorical, excluding the encounter ID which is unique for each sample. There are several characteristics of the dataset that can be handled by our selected models (Section 3): the skewed distributions apparent in the numeric features (Figure 1); the sparsity in the drug-related columns (Figure 6); and the medium degree of correlation or association between some variables (Figures 3, 4).

However, several challenges remain. There are missing values in the dataset (Table 1). Many of the categorical features have significant class imbalances that could introduce biases to our model (all in Figure 5 except gender, change, and the target). The characteristics of several categorical features are challenging for data pre-processing: several have high cardinality (Figure 8), and the ID columns are categorized by numerical values which are not reflective of an ordinal meaning (Figure 5). Finally, the data is not independent and identically distributed (i.i.d.), with repeat samples from the same patients (Figure 7). We are unable to ascertain whether any patient encounters represent their first medical visit, as encounters may have occurred pre-1999 or outside of the database locations. The dataset also includes patients admitted from or to hospice, or who passed away in the hospital.

2 Data assembling and initial pre-processing

We undertook pre-processing as if we would use the entire dataset in our model, although ultimately we used a sub-sample given computational constraints. We accounted for non i.i.d. data by retaining one sample per patient in the database, and dropping patients with a discharge disposition or admission source ID related to expiry or hospice. We dropped all columns with $> 30\%$ missing data and the unknown values in gender, and imputed the mode for missing race values. We encoded the target variable as binary, converted age categories to ordinal values at the midpoint of each age bracket, and implemented ordinal encoding of each of the drug columns. We also used thresholding and dropped drugs prescribed to fewer than 50 patients.

With regards to the 915 unique ICD-9 codes in the three diagnosis columns, we dropped the missing rows in diag_1, as we would expect a patient to have some initial diagnosis. However, we retained a "?" category for missing values in diag_2 and diag_3, as we posit that these may be left blank in the case where a patient has only one or two diagnoses. We grouped all non-diabetes related codes in buckets according to their overarching health concern.[1] We hypothesised that patients with the same diagnoses may have similar characteristics, regardless of whether the diagnosis was primary/secondary/tertiary. We therefore dummy encoded the remaining 54 unique diagnosis categories using a custom OneHotEncodeAndAggregate transformer in our preprocessor (Section 3 and code part 2.3), regardless of which column the diagnosis was from. Alternative approaches could consider other encoding methods, retaining unbucketed diagnosis codes (e.g. see [3]); and/or exploring whether retaining the primary/secondary/tertiary diagnosis structure is informative.

We handled the rest of the encoding process in a preprocessor which was applied in our machine learning pipeline to avoid data leakage. We dummy encoded race, gender, admission type ID, change, and diabetes medication in addition to the custom encoding of diagnoses. We applied smoothed target encoding on the discharge disposition ID and admission source ID. We scaled all non-dummy encoded values using MinMaxScaler.

Given limitations on computational resources, we used a bespoke under-sampling technique to select approximately 5% of our dataset. From a fairness and bias perspective, we aimed to balance demographic features, namely race, gender, and age, along with retaining the balance in the target variable. We included all Asian, Hispanic and other race patient samples given they were significantly underrepresented, and then undertook weighted stratified sampling by age, gender, and readmission of African American and Caucasian patients (see code part 2.4). This resulted in a subsample that

somewhat rectified the severity of the overall dataset's class imbalance in race, that had a balance of gender and our target class, and that retained a distribution of ages similar to our original database (Figure 9. However, this subsampling approach presented significant issues where we inspected interpretability, discussed in Section 4 below. We then applied an 80-20% split for our training and test set for use in the first pipeline (Section 3), stratifying by age given the largest demographic class imbalances remained in this category.

3 Design and build a machine learning pipeline

In addition to our linear SVM baseline, we selected an SVM with an RBF kernel ("RBF SVM"), a RandomForest ("RF"), and XGBoost ("XGB") for our analysis. We wished to explore models in principle capable of handling the full dataset, with varying degrees of complexity and approaches to the bias-variance trade-off. The different SVM kernels control the approach to decision-making boundaries which in turn influences the bias-variance trade-off, whereas the tree ensemble models address this through using bagging or boosting, respectively.

We chose a F-β score (specifically, the F2 score $\beta = 2$) as the metric to optimize in our cross-validation ("CV") pipeline. This represents the weighted harmonic mean of precision and recall, assigning twice as much importance to recall as it does to precision. This patient-safety-first approach prioritizes the identification of at-risk patients over falsely flagging patients as vulnerable to readmission. We also tracked overall accuracy given that this is important in resource-constrained settings.

Given we had already considered demographics in our sub-sample and overall train-test split (Section 2), we chose to remain agnostic to stratification of predictive features in our CV approach. We used StratifiedKFold CV with five splits to simply ensure that the target class remained balanced across folds. We integrated our preprocessor and CV in GridSearchCV on the training set to consider a selection of hyperparameters for each model (see code part 3.1.1. and Table 2). Given computational limitations we chose between one-three hyperparameters per model that directly influenced the bias-variance trade off and/or the models' complexities. For each hyperparameter, we chose 5-6 possible values. Note that these values do not present a comprehensive hyperparameter tuning due to computational limitations, as discussed in Section 7. The best models from the GridSearchCV were then evaluated on our held-out test set.

The hyperparameters associated with the best CV performance can be found in Table 3. Relatively low regularization (C) resulted in better performance in both SVMs (see Figures 10, 11). This indicates that our dataset may be noisy, so allowing for more flexibility in the decision function improved generalization on the validation data. There was a clear peak in performance across the hyperparameters tuned for the RF, whereas the increase/decrease in performance levelled off for XGB. (Figures 12, 13.

The performance of the best-performing models from the CV pipeline on the held-out test set can be seen in Table 4. The RF achieved the highest performance across both metrics on the test set and during cross validation (Table 5, Figures 14, 15). The Linear SVM and RF models appeared to have relatively low variance in our cross validation pipeline compared to the other models.

4 Model interpretation

The most straightforward method to interpret feature importance differs across the selected models. The feature importances attribute in the RF and XGB calculates the Gini importance of each feature (Figure 16), and the gain (Figure 17), respectively. We can also isolate the weight of the coefficients of the Linear SVM (Figure 18). However, feature importance is not directly interpretable in the RBF SVM due to its transformation of data into the high-dimensional space where standard coefficients are not applicable - a weakness in its utility in the medical domain.

In order to directly compare the results, we assessed permutation importances of each best-performing model from the cross-validation pipeline on the test set (Figures 19, 20, 21, 22, 23). This revealed that weightings differed across models, although several consistent features were reasonable

such as number of diagnoses, inpatient and outpatient visits. However, the significant prominence of race as the most influential feature across all models is concerning, most substantially so for our SVMs given this is the only feature that had a strong impact during the permutation tests. This suggests that race is being leveraged as a predictor of diabetic readmissions, which may mirror systemic biases in our dataset and/or healthcare more broadly rather than causality. The proportions of readmitted patients differed per race in the dataset and our subsample 24 which likely contributed to this. Further analysis may consider combining our undersampling technique with upsampling of readmitted patients per race so that we have equal target class proportions for each race, and re-running these models.

5 Alternative machine learning pipeline

We implemented a nested CV pipeline, combining StratifiedKFold with GridSearchCV for hyperparameter tuning. We used 5-fold StratifiedKFold CV in both the inner loop (for hyperparameter tuning with GridSearchCV) and the outer loop for model evaluation. The stratification was selected as in our previous pipeline, to maintain the even proportion of target class labels across each fold. We used the same selection of hyperparameters for the GridSearchCV as in our previous pipeline, the best combinations can be found in 7.

This was applied across our entire subsample given the train/validation/test split can be managed internally within the pipeline. The results in Table 6 that the best performing model was still the RF on both metrics, with the lowest standard deviation as well (albeit tied with the Linear SVM). The metric differences between XGB and RF, and between the two SVMs, respectively, were insubstantial ($<= 0.01$) although the differences were more evident in the F2 score standard deviations. Furthermore, the relative stability in variance previously attributed to the Linear SVM and RF's performance in Section 3 did not persist under this expanded set of trials (Figures 25, 26). This demonstrates the importance of the nested cross-validation approach in providing more stable estimates as opposed to our initial comparison of the best performing models on a single test set in Section 3.

6 Identify limitations and propose potential solutions

As mentioned in Section 1, a significant limitation of the dataset is the class imbalance prevalent in many of the categorical features which can impact any model's ability to generalize to patients with those characteristics. The columns with a significant extent of missing data may also be a limitation (although imputation is possible), as these may be important predictors. As discussed in Section 1, the dataset was also not i.i.d.; another related element to consider was that the dataset spanned 10 years, during which the diagnostic criteria and treatment protocols may have evolved. The dataset could therefore be improved by collecting more data on minority categories and missing features, recording the dates of treatment, and recording patient admission history beyond the existing one-year features.

Our models and pipeline may have been limited by our approach to the challenges in the dataset discussed in Section 2, particularly the subsampling approach. Beyond this, the primary limitations of our models and pipeline were that given computational resources and the timeline for this coursework, we were unable to use the full dataset, do more extensive hyperparameter tuning (e.g. on all possible hyperparameters and a wider range), and repeat and average the results of our pipelines. This could be explored to improve the rigour of our approach. Another limitation of our models in a real-world context is that they are unable to directly account for new features we may want to add - e.g. incorporating additional information about a patient that may be important from a medical practitioner's perspective, but that is not already being recorded in our training text. It is not clear that our models perform well enough for real-world deployment, both in terms of the metrics we tracked as well as in terms of biases. Further exploration could include a detailed analysis of fairness across demographic groups (e.g. [2]).

7 Figures

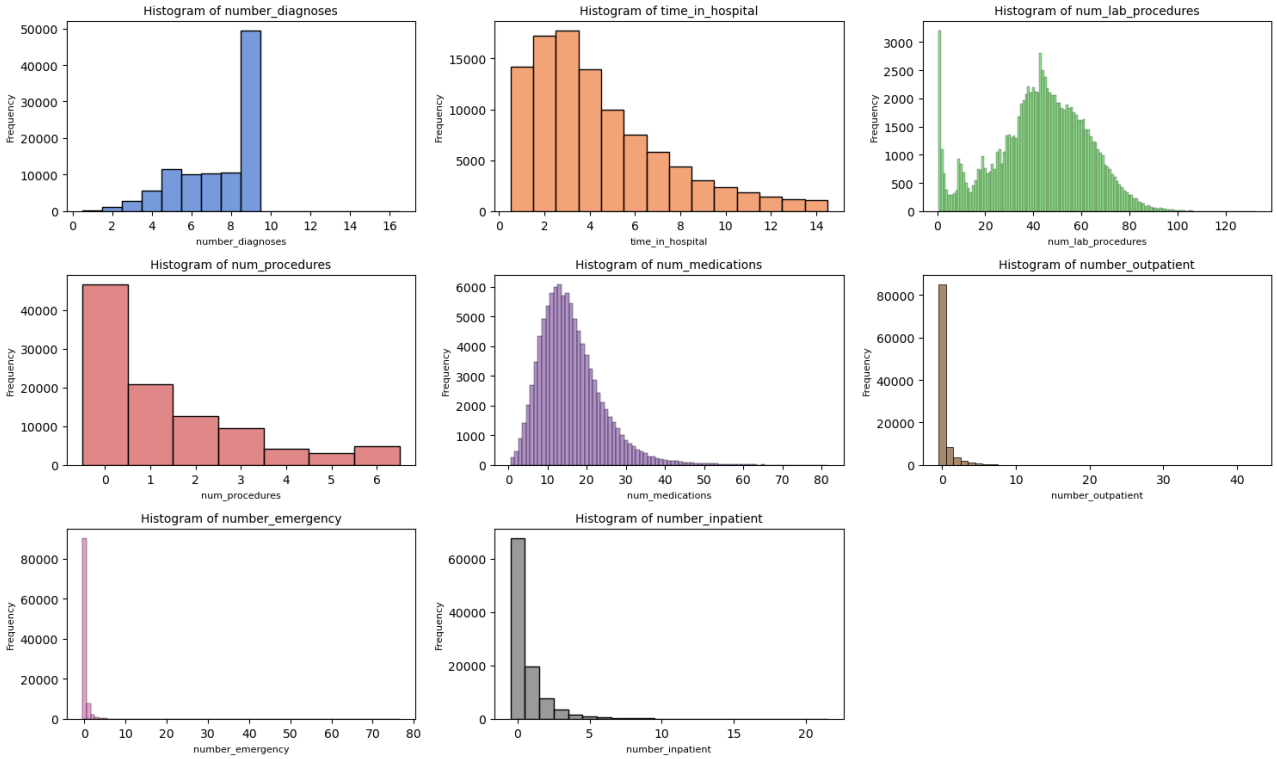

Figure 1: **Histograms of discrete columns:** These show that our discrete columns have varying scales and distributions. All appear to be skewed, with some appearing to have a somewhat exponential distribution or a skewed Gaussian distribution.

	Missing values	Percentage of total
encounter˙id	0	0.0
patient˙nbr	0	0.0
race	2273	2.2
gender	3	0.0
age	0	0.0
weight	98569	96.9
admission˙type˙id	0	0.0
discharge˙disposition˙id	3691	3.6
admission˙source˙id	6781	6.7
time˙in˙hospital	0	0.0
payer˙code	40256	39.6
medical˙specialty	49949	49.1
num˙lab˙procedures	0	0.0
num˙procedures	0	0.0
num˙medications	0	0.0
number˙outpatient	0	0.0
number˙emergency	0	0.0
number˙inpatient	0	0.0
diag˙1	21	0.0
diag˙2	358	0.4
diag˙3	1423	1.4
number˙diagnoses	0	0.0
max˙glu˙serum	96420	94.7
A1Cresult	84748	83.3
metformin	0	0.0
repaglinide	0	0.0
nateglinide	0	0.0
chlorpropamide	0	0.0
glimepiride	0	0.0
acetohexamide	0	0.0
glipizide	0	0.0
glyburide	0	0.0
tolbutamide	0	0.0
pioglitazone	0	0.0
rosiglitazone	0	0.0
acarbose	0	0.0
miglitol	0	0.0
troglitazone	0	0.0
tolazamide	0	0.0
examide	0	0.0
citoglipton	0	0.0
insulin	0	0.0
glyburide-metformin	0	0.0
glipizide-metformin	0	0.0
glimepiride-pioglitazone	0	0.0
metformin-rosiglitazone	0	0.0
metformin-pioglitazone	0	0.0
change	0	0.0
diabetesMed	0	0.0
readmitted	0	0.0

Table 1: **Counts and percentages of missing values in each column:** Including NaN's, "?" or "Unknown/Invalid" values.

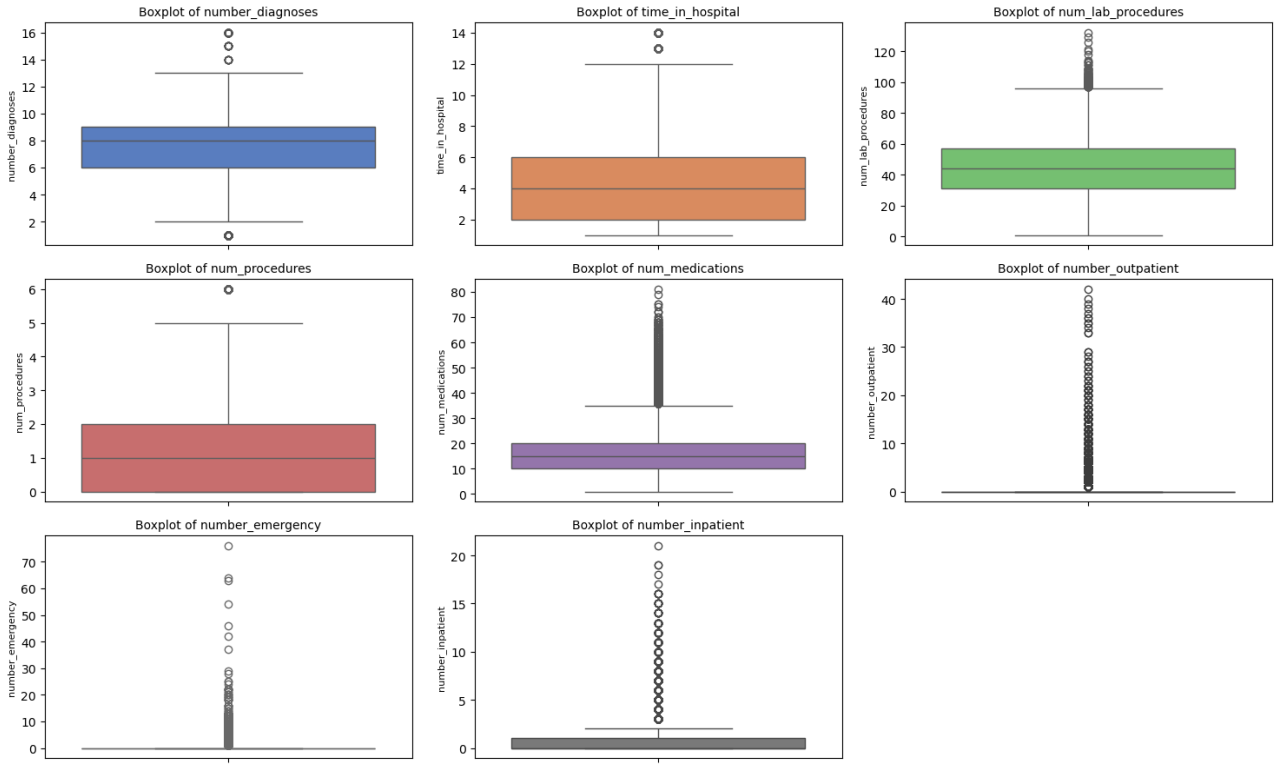

Figure 2: **Box plots of discrete columns:** The box plots also evidence the skew apparent in Figure 1

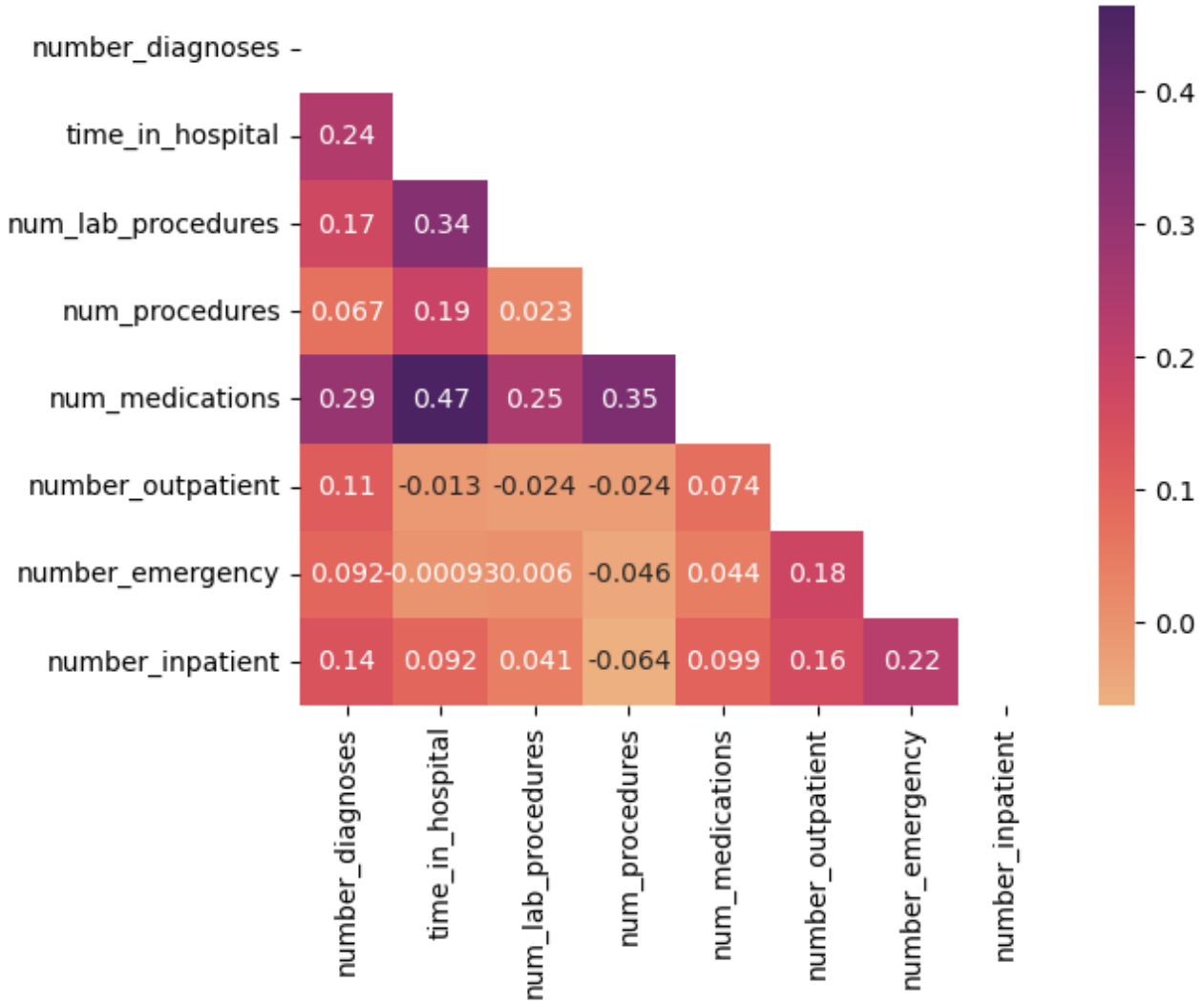

Figure 3: **Heatmap of Spearman's correlation between discrete features:** The heatmap indicates that most variables are weakly correlated except for the time spent in the hospital with the number of lab procedures and the number of medications. This makes sense and we do not use this information to do any feature engineering - our selected models can handle some extent of correlation, and these features indicate related but different information.

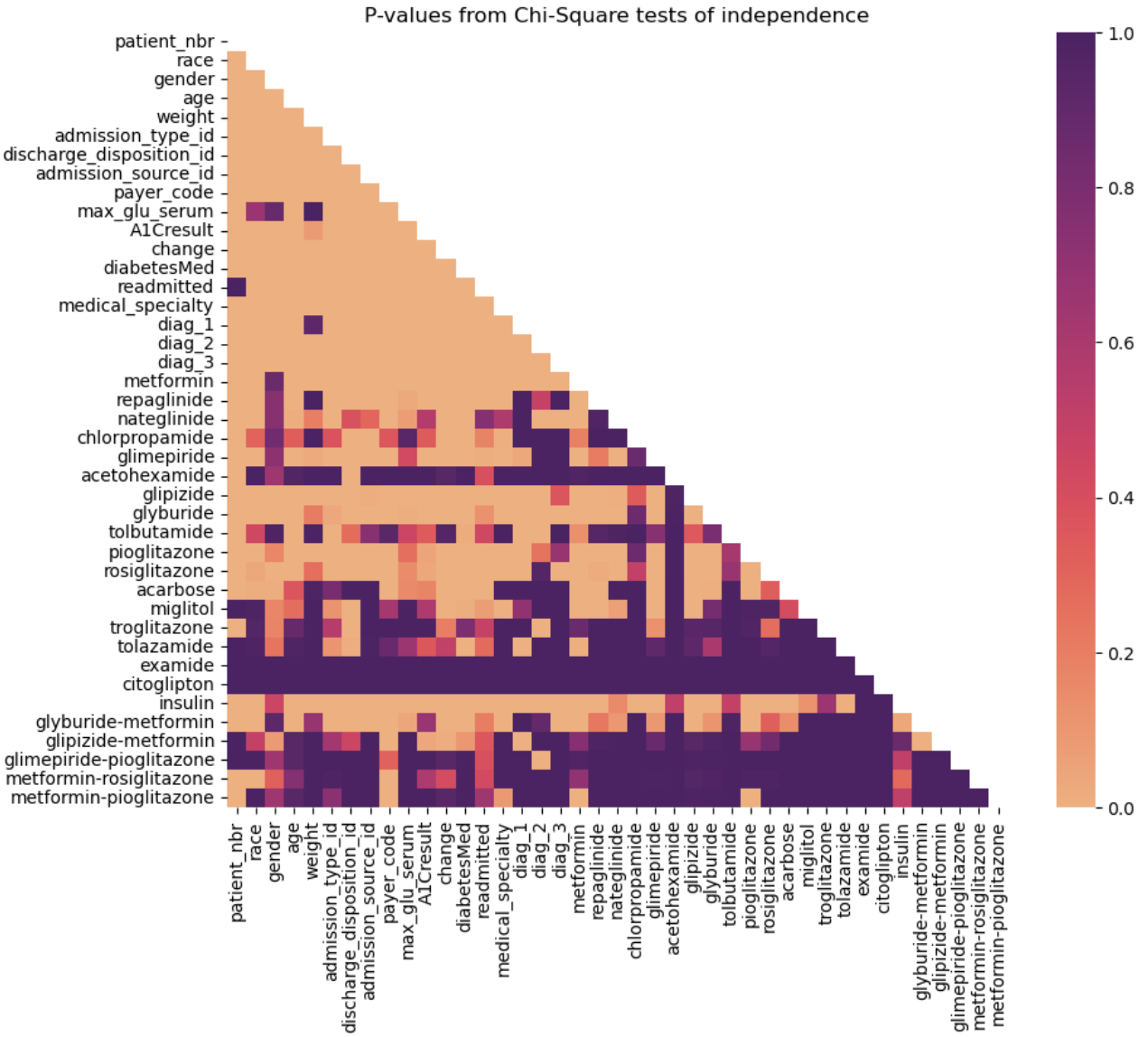

Figure 4: **Heatmap of Chi-squared test of independence between categorical features:** At first inspection the heatmap seems worrying in that there is a significant amount of evidence of a strong relationship between many of the features. However closer inspection shows that these features are those that are sparse or otherwise have a high number of missing values. Therefore the takeaway can be that there are likely some relationships, although the data sparsity may be obscuring the strength of this relationship.

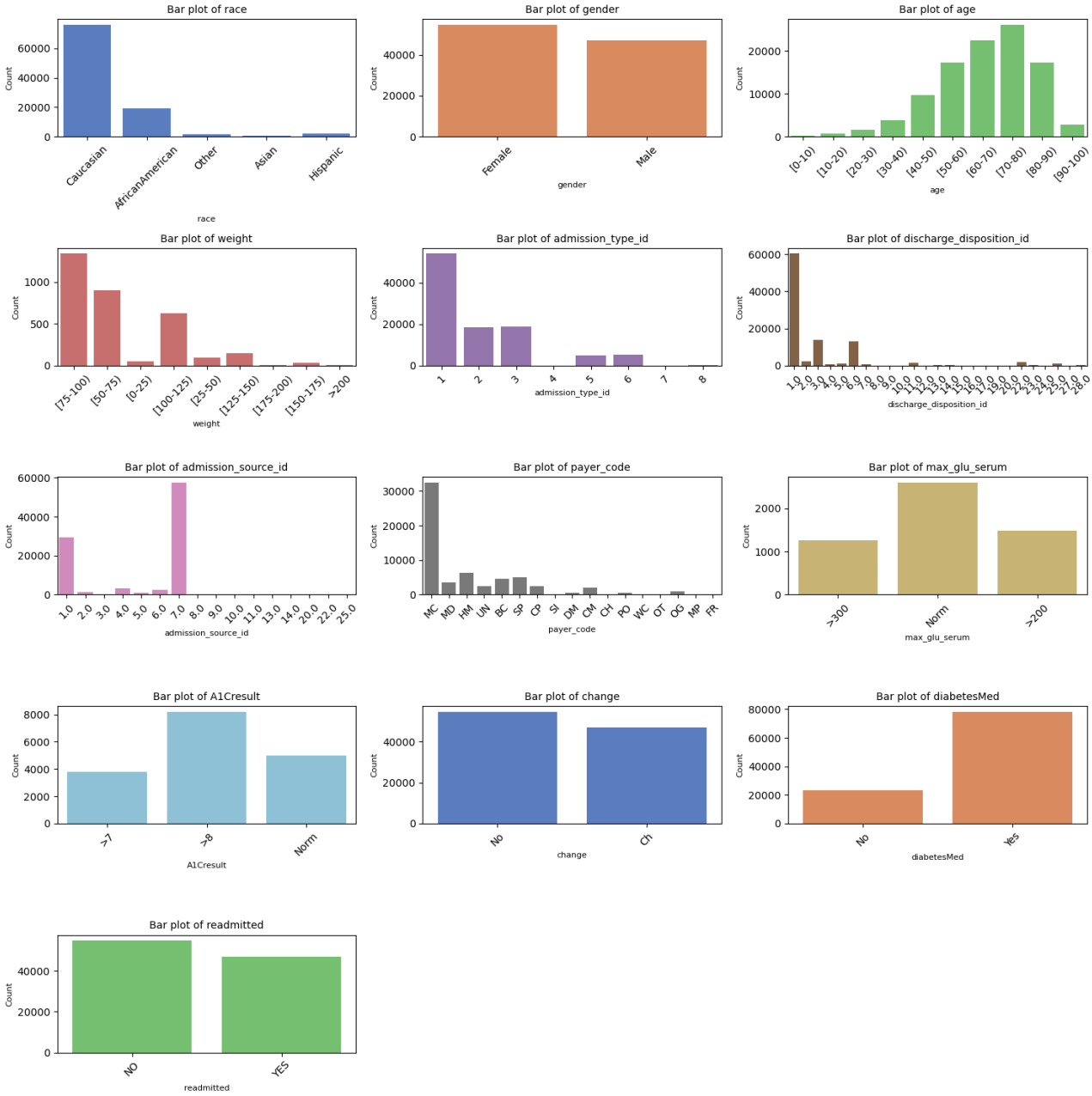

Figure 5: **Bar charts of categorical features and the target:** These charts show a significant amount of imbalance in the categories within these features. The only categories that seem relatively balanced are gender, change, and the target.

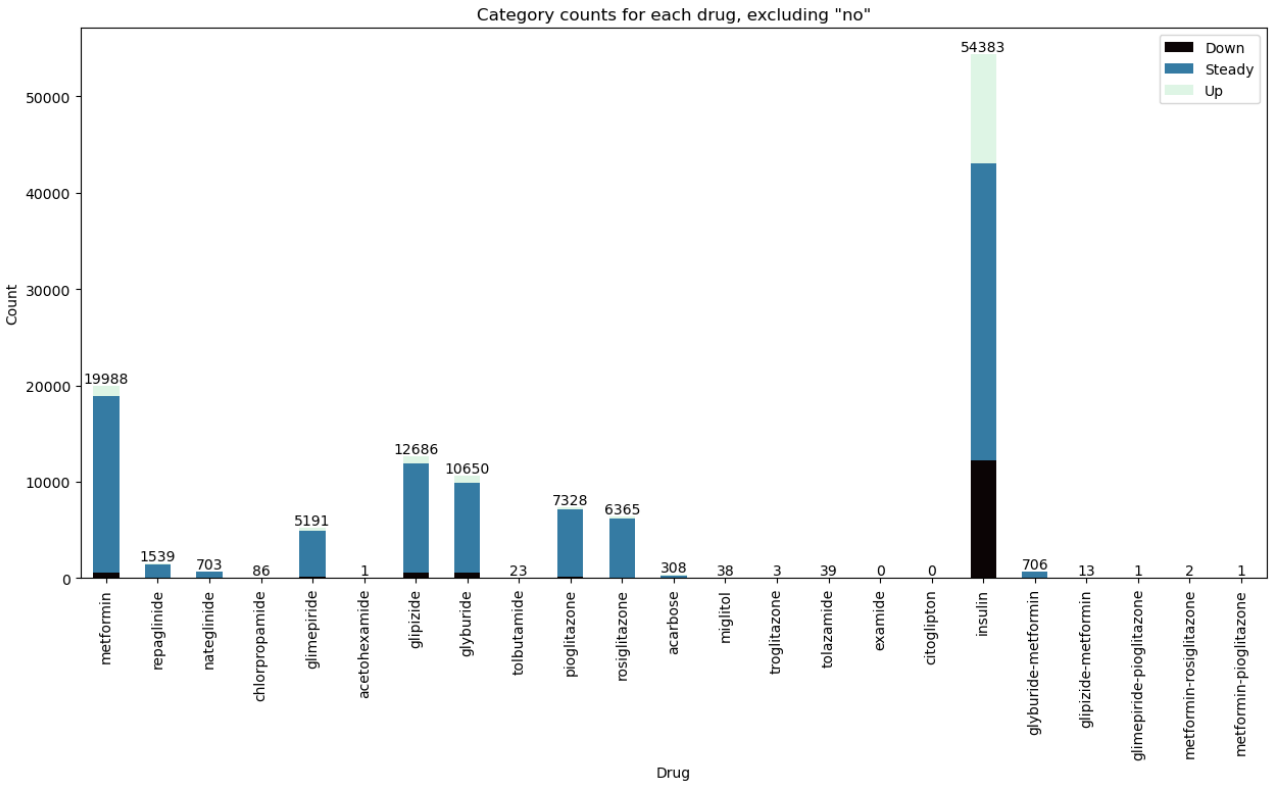

Figure 6: **Stacked bar chart of drug-related features, excluding "no" prescription:** This shows the significant degree of sparsity in patients actually prescribed each drug, as well as that most taking a drug have a "steady" prescription.

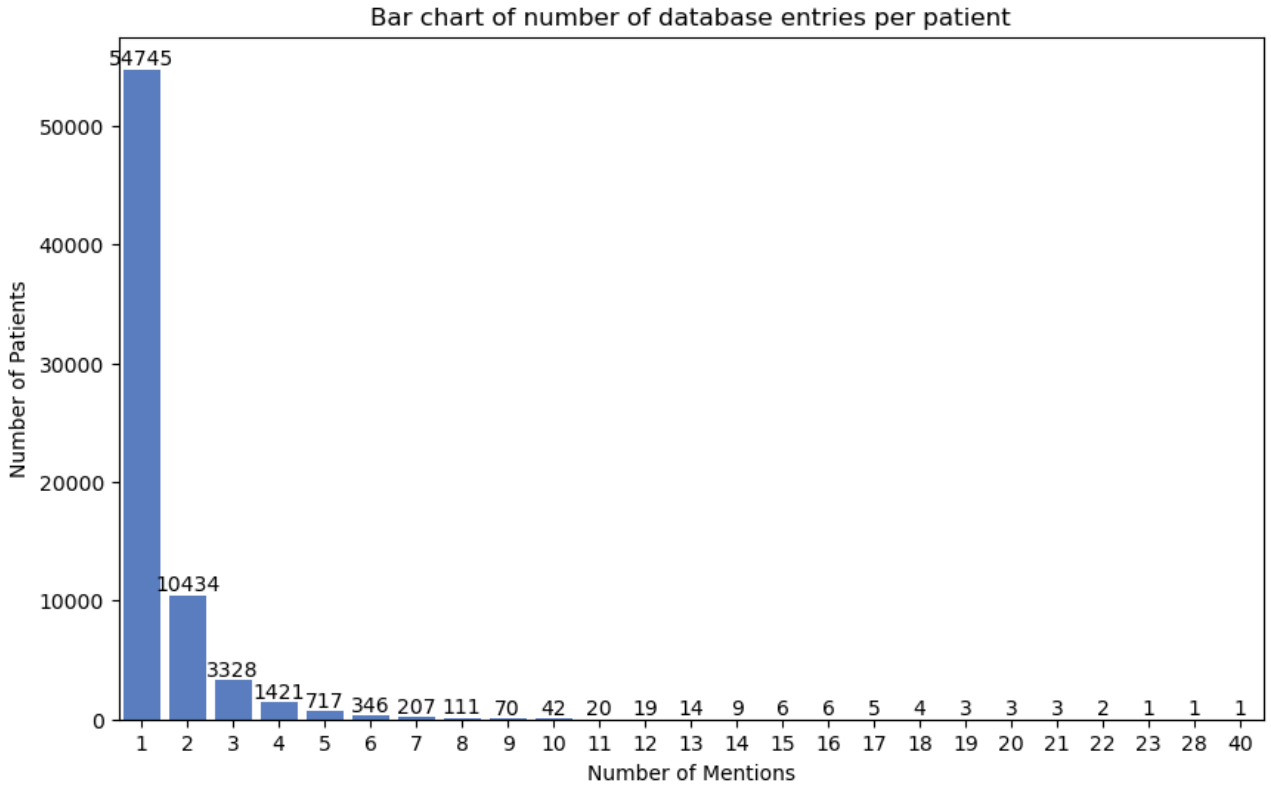

Figure 7: **Bar chart of number of database entries per unique patient:** This shows that the majority of patients in the database are only included once. However, there are 16,773 patients who have multiple encounters recorded in the database. This suggests that those entries are not independent.

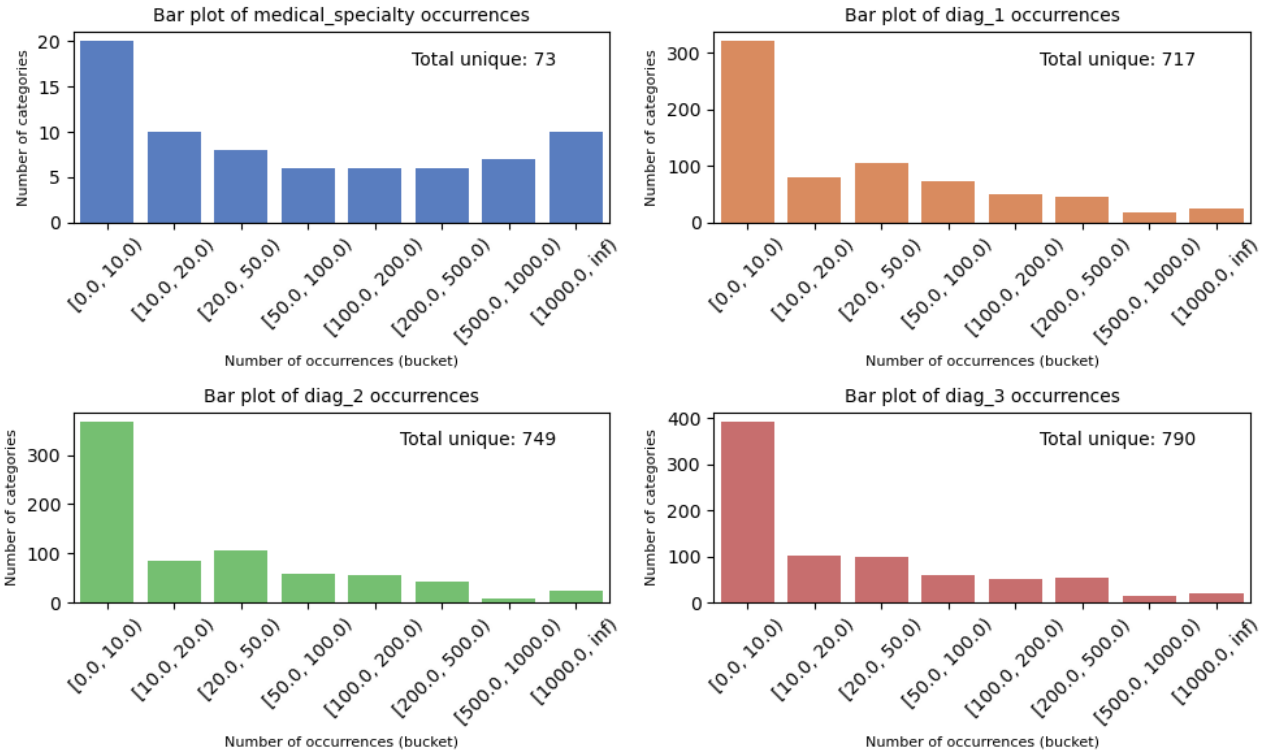

Figure 8: **Bar plots of counts for high-cardinality categorical features:** These bar charts show the counts of samples from each unique category for each selected categorical column with high cardinality (e.g. approx. 300 diagnoses were mentioned between 0-10 times in the database). This gives us an initial overview of the distribution of these features.

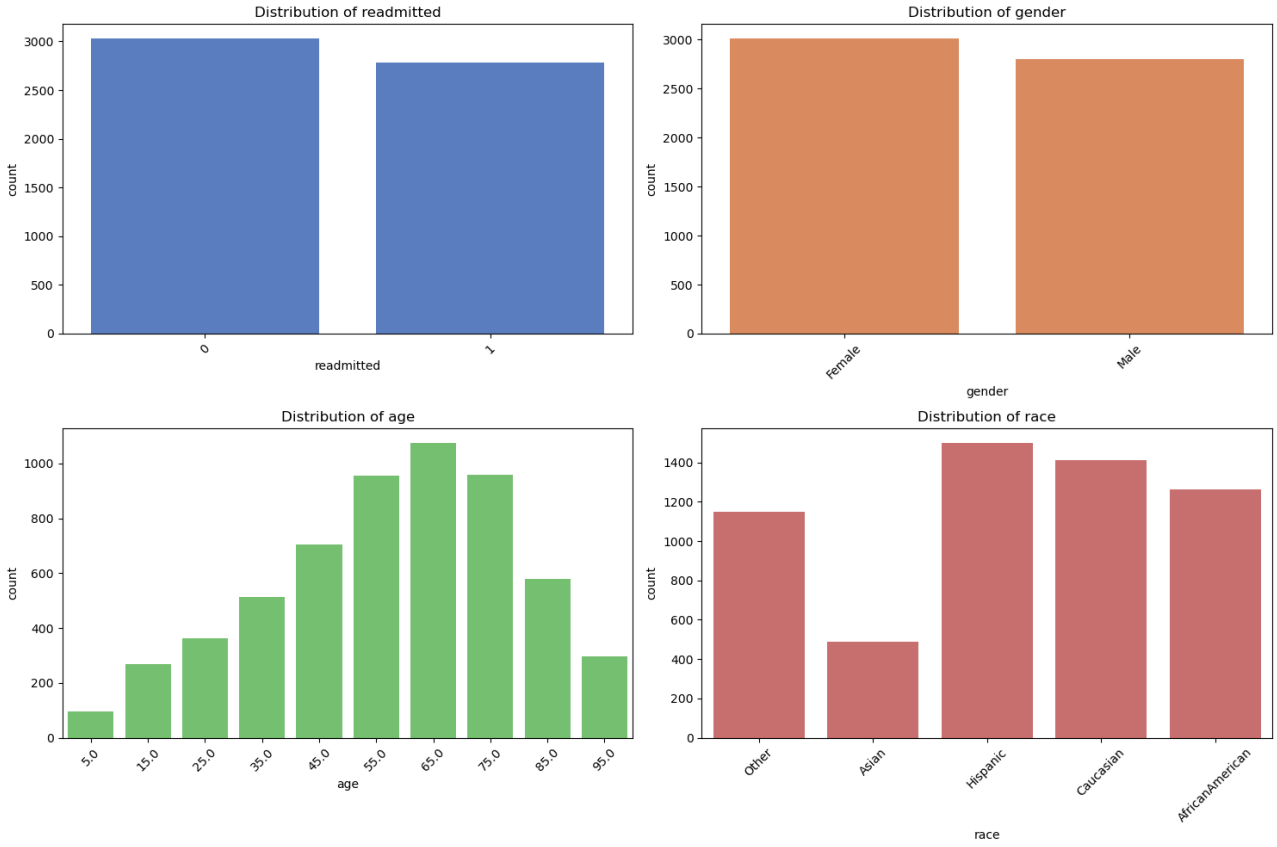

Figure 9: **Box plots of distribution of demographic features in subsample:** As can be seen, our final subsample still has an imbalance in race and ages, but this is improved with regards to the original database. We retained the balanced distributions of our target and gender. However, it is important to note that the distribution of these demographic features within each race is imbalanced - for example, Asian, Hispanic, and Other patients in the database were less likely to be readmitted so we sampled slightly more readmitted patients from African American and Caucasian races; these groups also had even more significant imbalances in age which contrasted with even sampling across age groups of African American and Caucasian patients. This may have contributed to the dynamics discussed in Section 4 and explored in Figure 24.

Table 2: Model hyperparameters selected for GridSearchCV

Model	**Hyperparameters**
Linear SVM	classifier__C: [0.001, 0.01, 0.1, 1, 10, 100]
RBF SVM	classifier__gamma: [1, 0.1, 0.01, 0.001] classifier__C: [0.01, 0.1, 1, 10, 100]
RF	classifier__n_estimators: [50, 100, 200, 500, 1000] classifier__max_depth: [None, 5, 20, 50]
XGB	classifier__n_estimators: [50, 100, 200, 500, 1000] classifier__learning_rate: [0.001, 0.01, 0.1, 0.3] classifier__max_depth: [None, 5, 20, 50]

Table 3: Best hyperparameters for each model

Model	**Best hyperparameters**
Linear SVM	classifier__C: 0.01
RBF SVM	classifier__C: 0.1, classifier__gamma: 0.1
RF	classifier__max_depth: 20, classifier__n_estimators: 500
XGB	classifier__learning_rate: 0.01, classifier__max_depth: None, classifier__n_estimators: 500

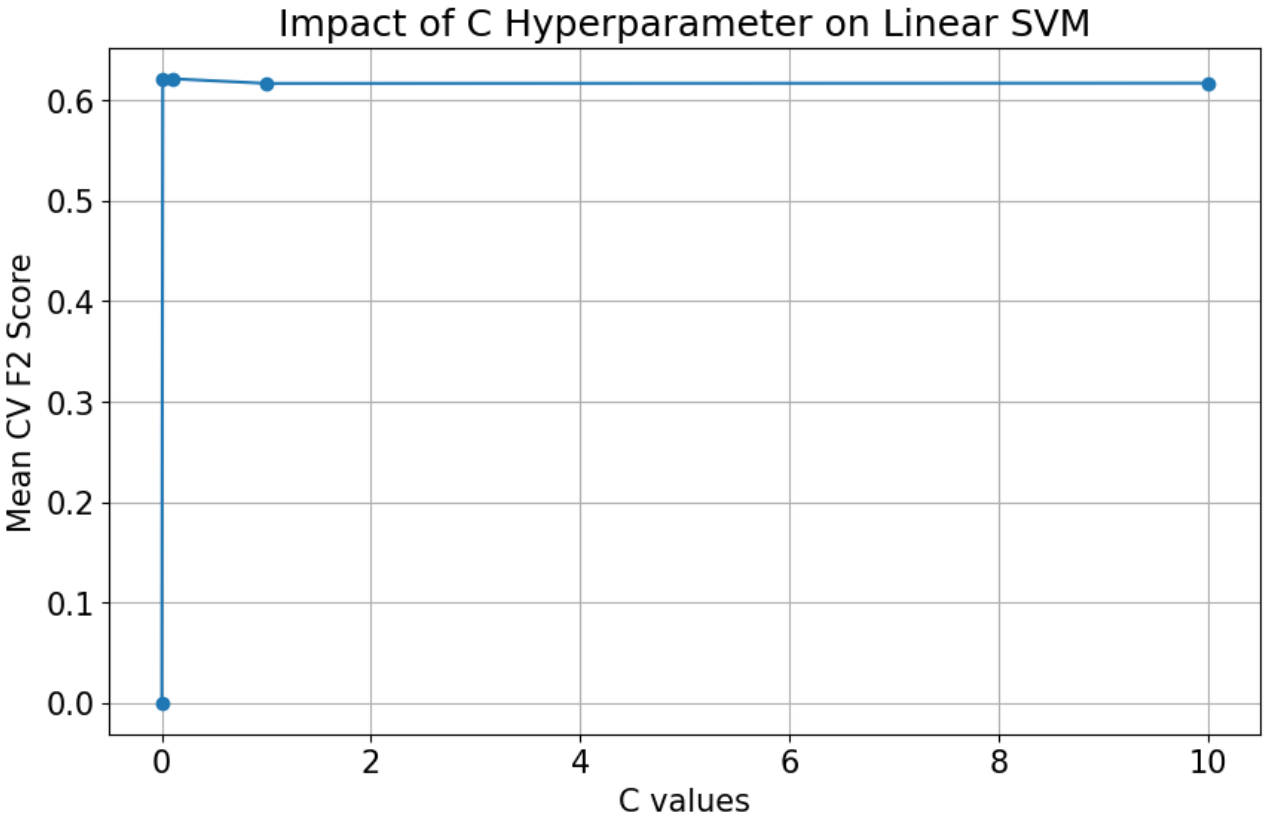

Figure 10: **Line plot of impact of different C values on mean CV F2 score of Linear SVM:** We can see that the best C value is obtained at 0.01. This suggests that the model benefits from a higher degree of regularization to prevent overfitting and improve its generalization. As with Figures 11, 13, ideally we could continue hyperparameter tuning past the 'plateau' until we see worse performance in our charts, ensuring we have explored a wide enough range of hyperparameters. While we did explore a wider range of additional values for each model (not pictured here), it would require more computational resources and time to explore beyond the plateaus (which continued for significantly higher x-value hyperparameters).

Table 4: Test scores for each model

Model	Test F2 Score	Test Accuracy
Linear SVM	0.615	0.652
RBF SVM	0.615	0.652
RF	0.656	0.674
XGB	0.633	0.665

Table 5: Cross-validation scores for each model

Model	Mean F2	Std Dev F2	Mean Accuracy	Std Dev Accuracy
Linear SVM	0.619	0.017	0.649	0.014
RBF SVM	0.393	0.014	0.592	0.010
RF	0.634	0.015	0.660	0.016
XGB	0.577	0.021	0.627	0.017

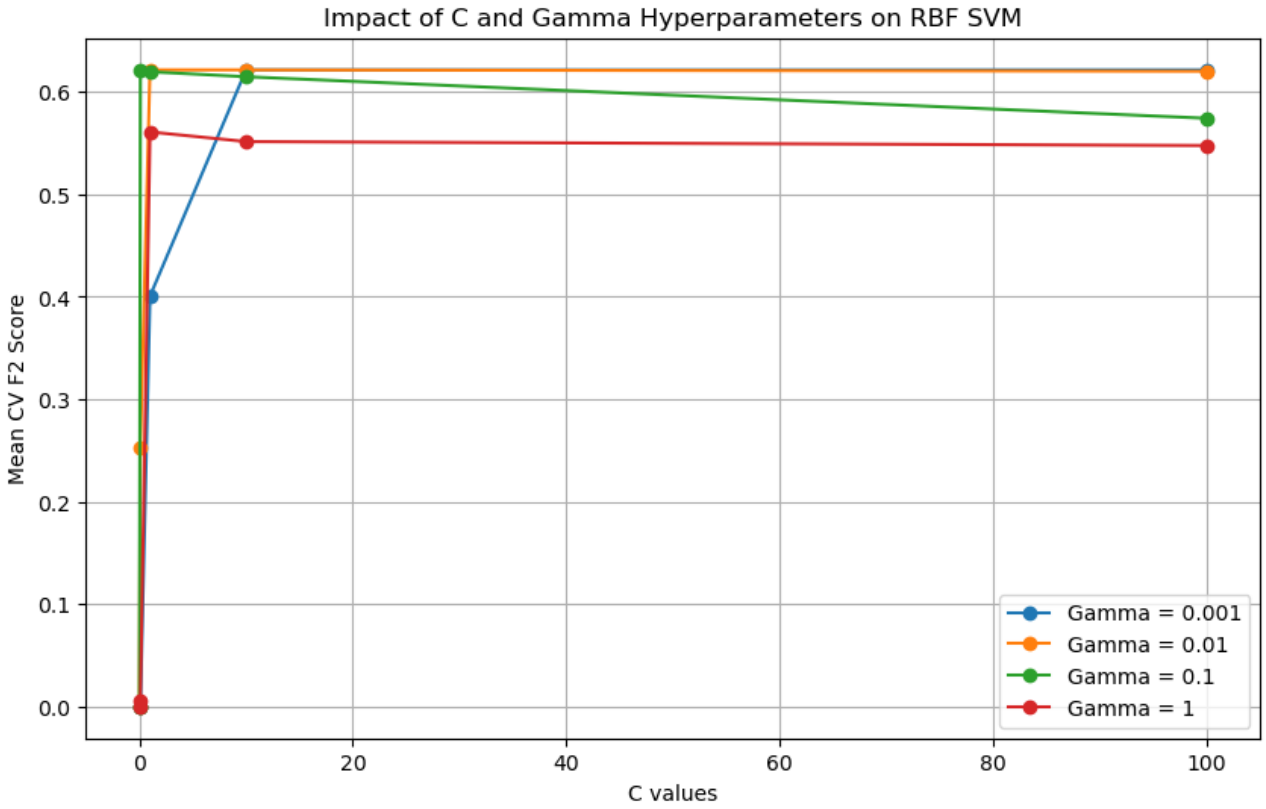

Figure 11: **Line plot of impact of different C and γ values on mean CV F2 score of RBF SVM:** It is evident that the best combination of hyperparameters was 0.1 for each.

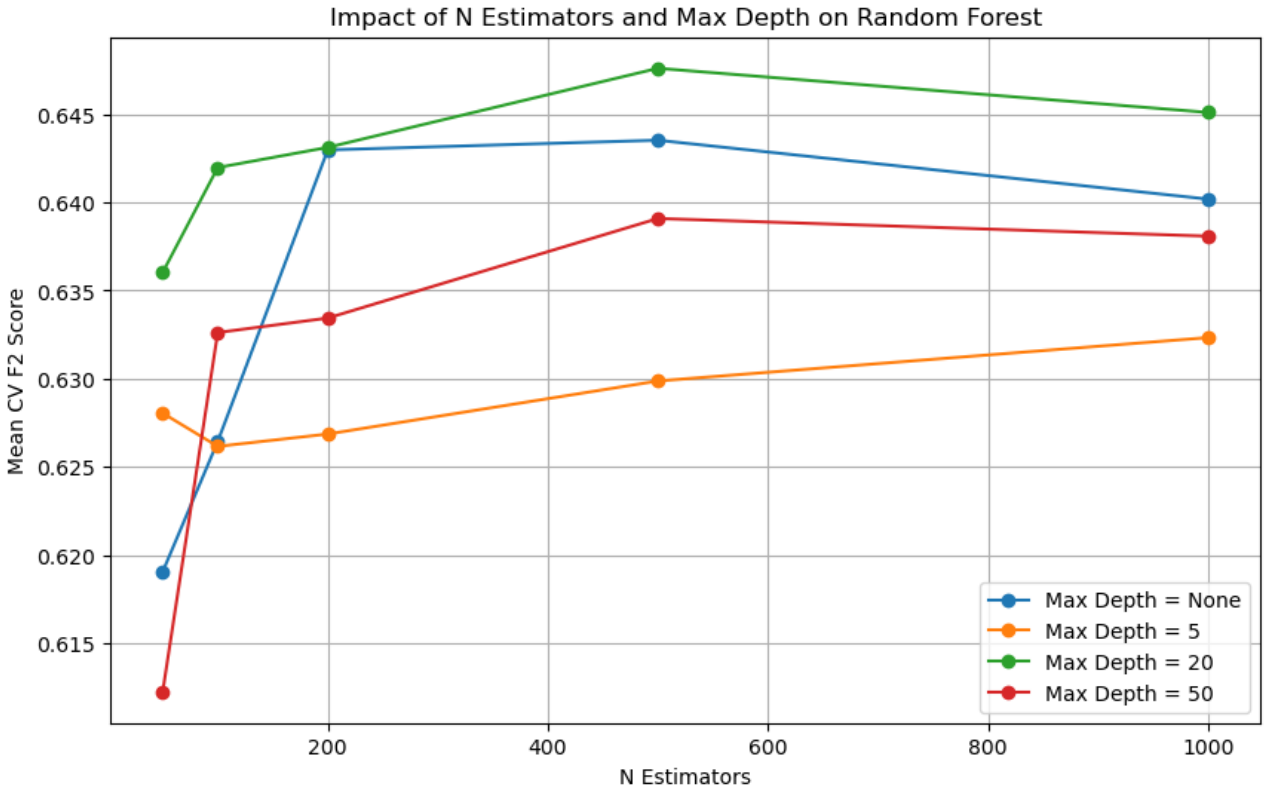

Figure 12: **Line plot of impact of number of estimators and max depth on mean CV F2 score of RF:** This indicates that the best performance was found by taking a max depth of 50 and 500 estimators. However, the performance was quite varied across different combinations of hyper-parameters and we did not see the expected peak/curve downwards of our tuning results. However, given that further exploration would require significant computational resources as we continue to increase the number of estimators, we selected the clear best performer from our range.

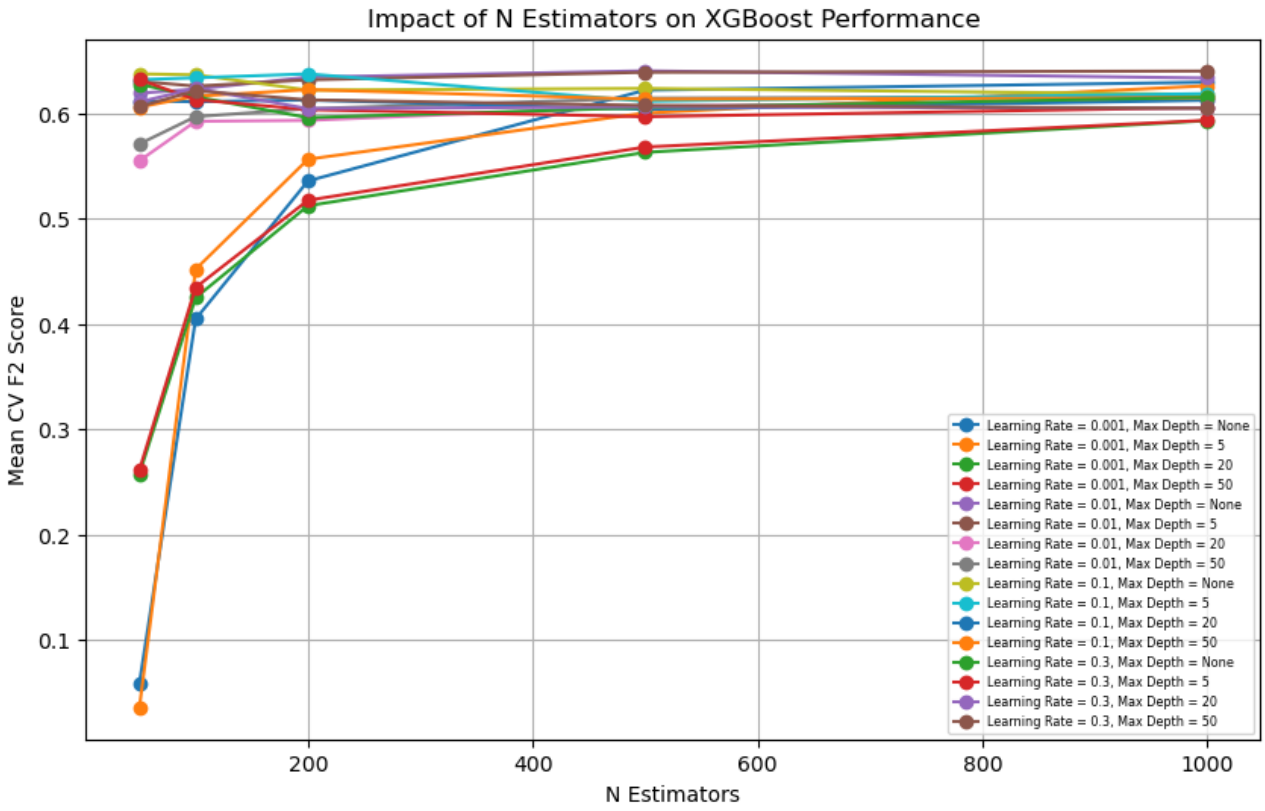

Figure 13: **Line plot of impact of number of estimators and max depth on mean CV F2 score of XGB:** Although this is difficult to visualise, the peak is found at 500 estimators.

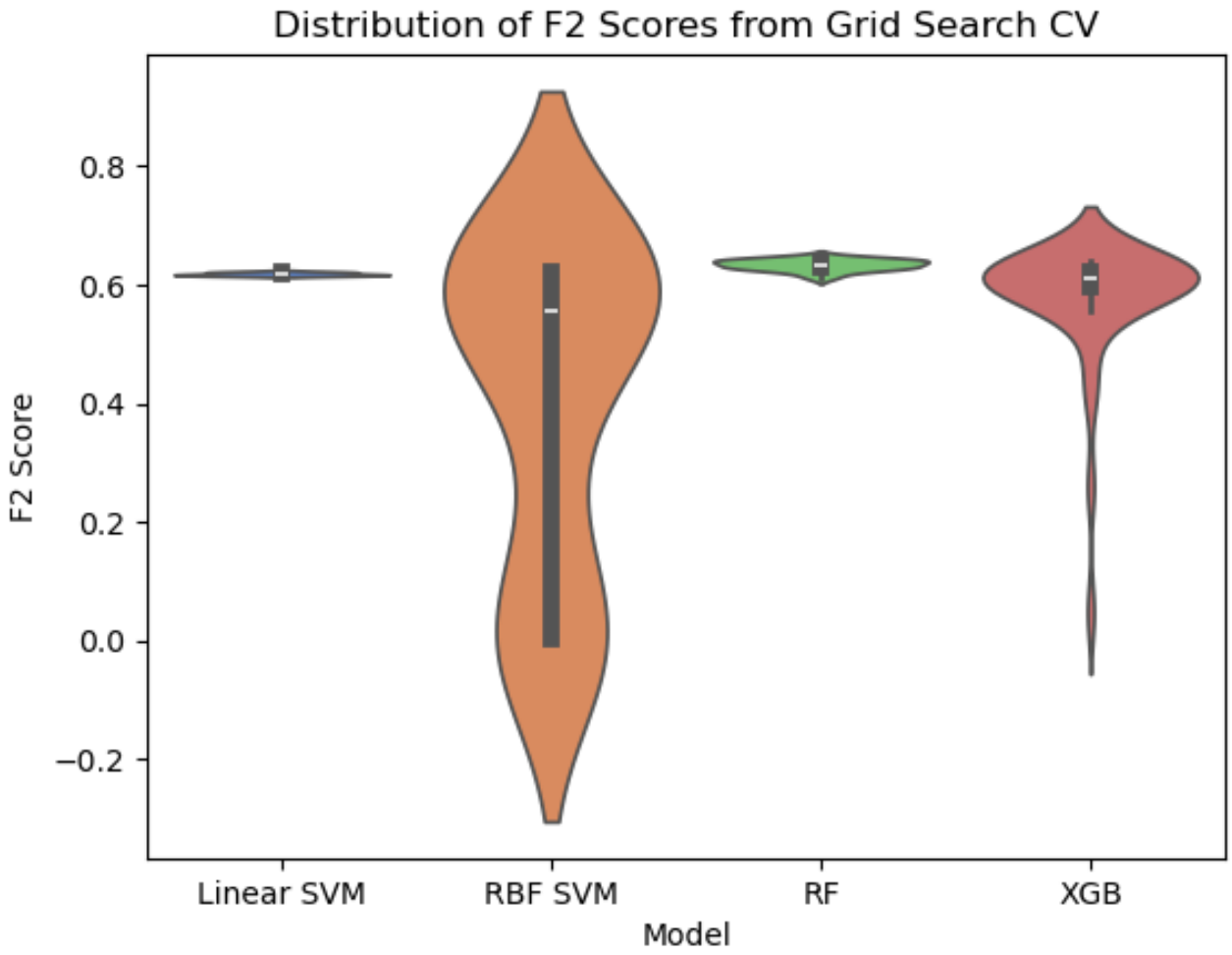

Figure 14: **Distribution of F2 scores on cross-validation sets**: This indicates the F2 score results from each validation set in the cross-validation folds.

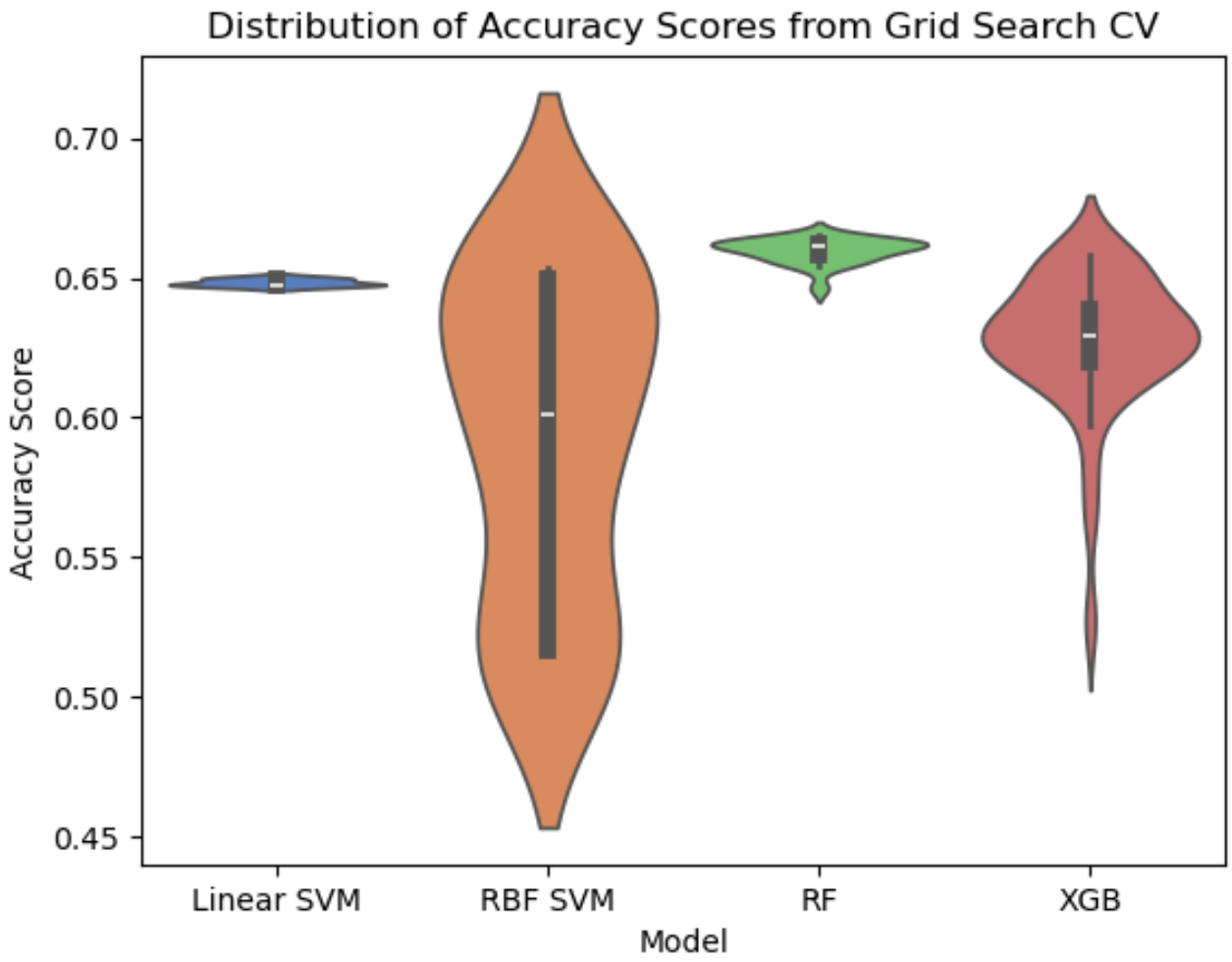

Figure 15: **Distribution of accuracy scores on cross-validation sets:** This indicates the accuracy score results from each validation set in the cross-validation folds. The behaviour on accuracy, although the models are not being optimized for this metric, is extremely similar visually to that evidenced for the F2 scores in Figure 14. This can be seen as evidence that the difference in the bias-variance trade-off is closely related to the model architectures as applied to this dataset.

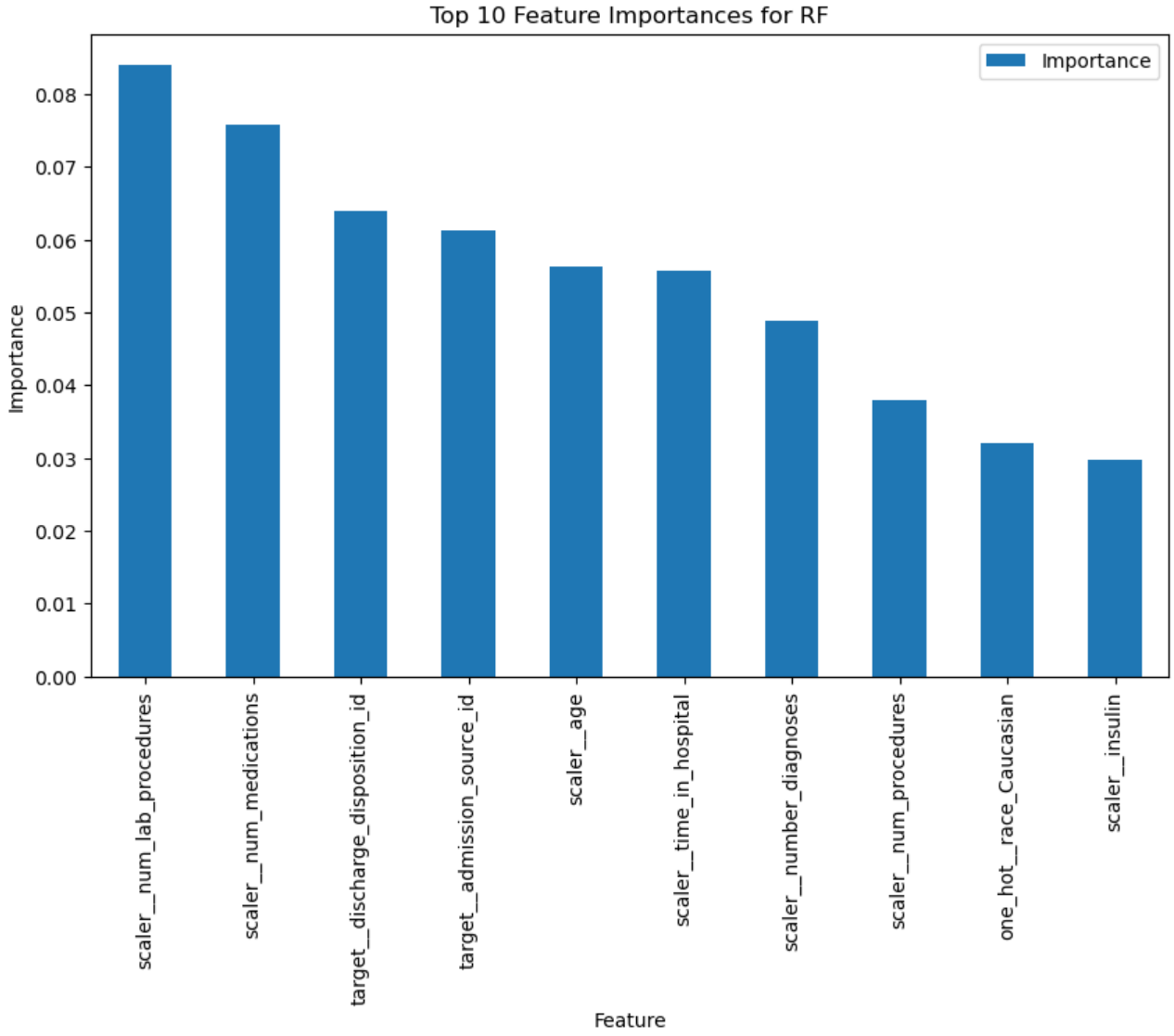

Figure 16: **Top 10 features by importance for RF:** These features are broadly reasonable, and quite different from those highlighted in Figures 17 and 18. In addition, the analysis here does not indicate that the RF prioritizes 'race' as highly as evident in the permutation test. This is interesting and may be explained by the different metrics used to calculate importance between the two methods.).

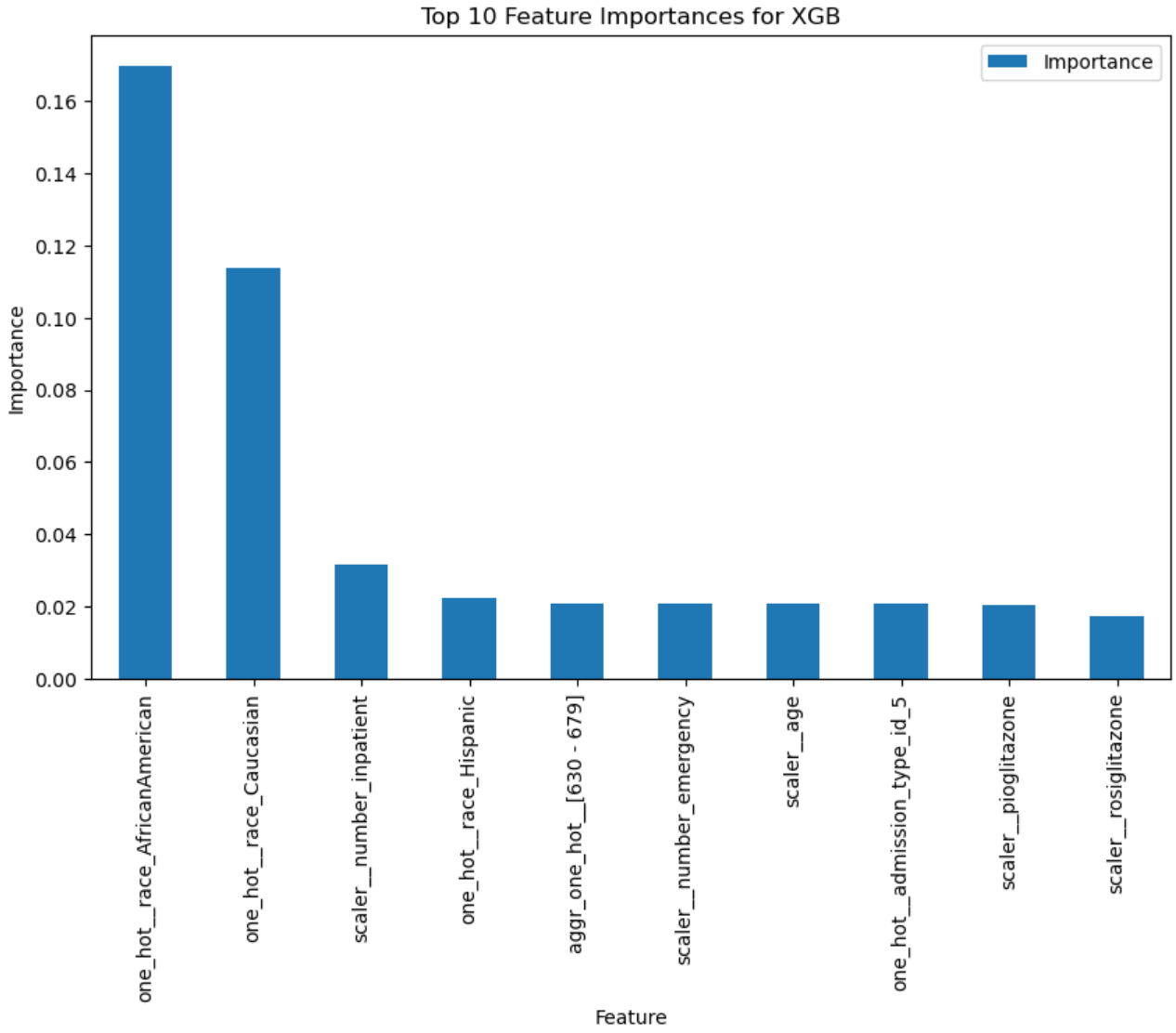

Figure 17: **Top 10 features by importance for XGB:** These features are broadly reasonable, except for the high prominence of race which is discussed in 4. The presence of diagnosis codes is of interest.

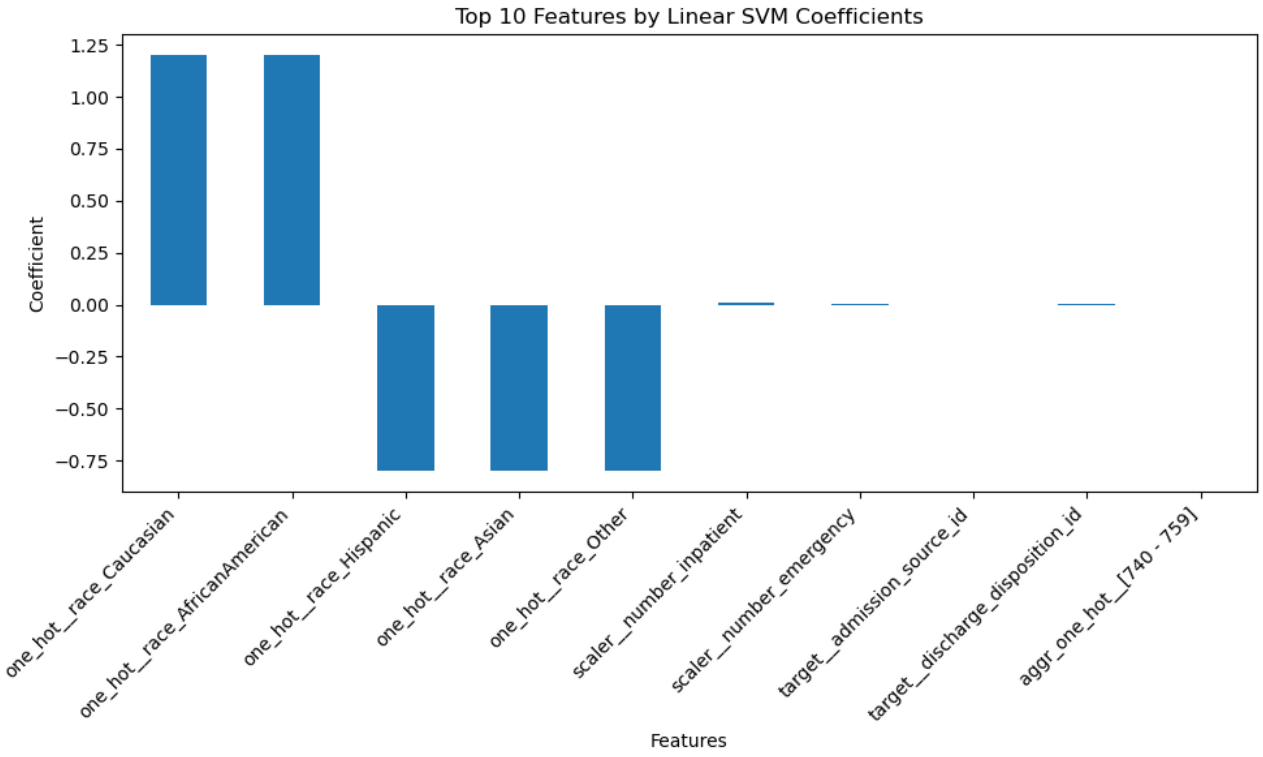

Figure 18: **Top 10 features by importance for Linear SVM:** This displays similar behaviour to the XGB feature importance in Figure 17 and permutation importances in 19. It is interesting that the non-race features all contribute a small amount to the decision.

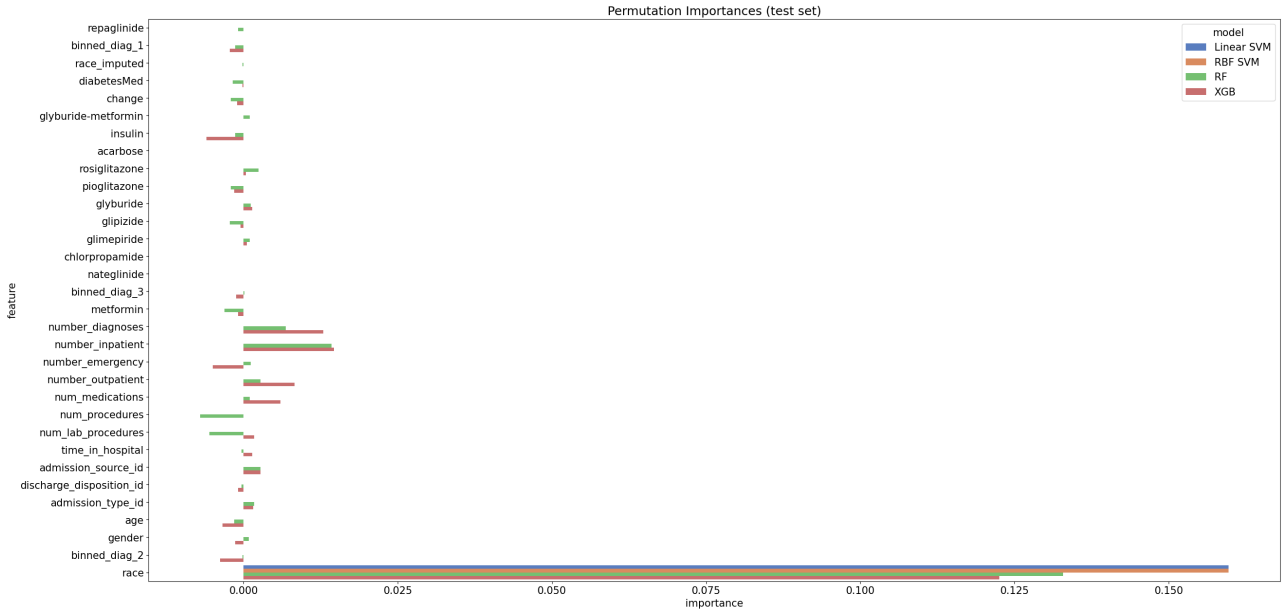

Figure 19: **Mean permutation importances of each model applied to the test set:** This indicates much similarity in the relevant importances ascribed to different features, although the exact extent of their predictive importance (and contribution towards a positive or negative result) often differs.

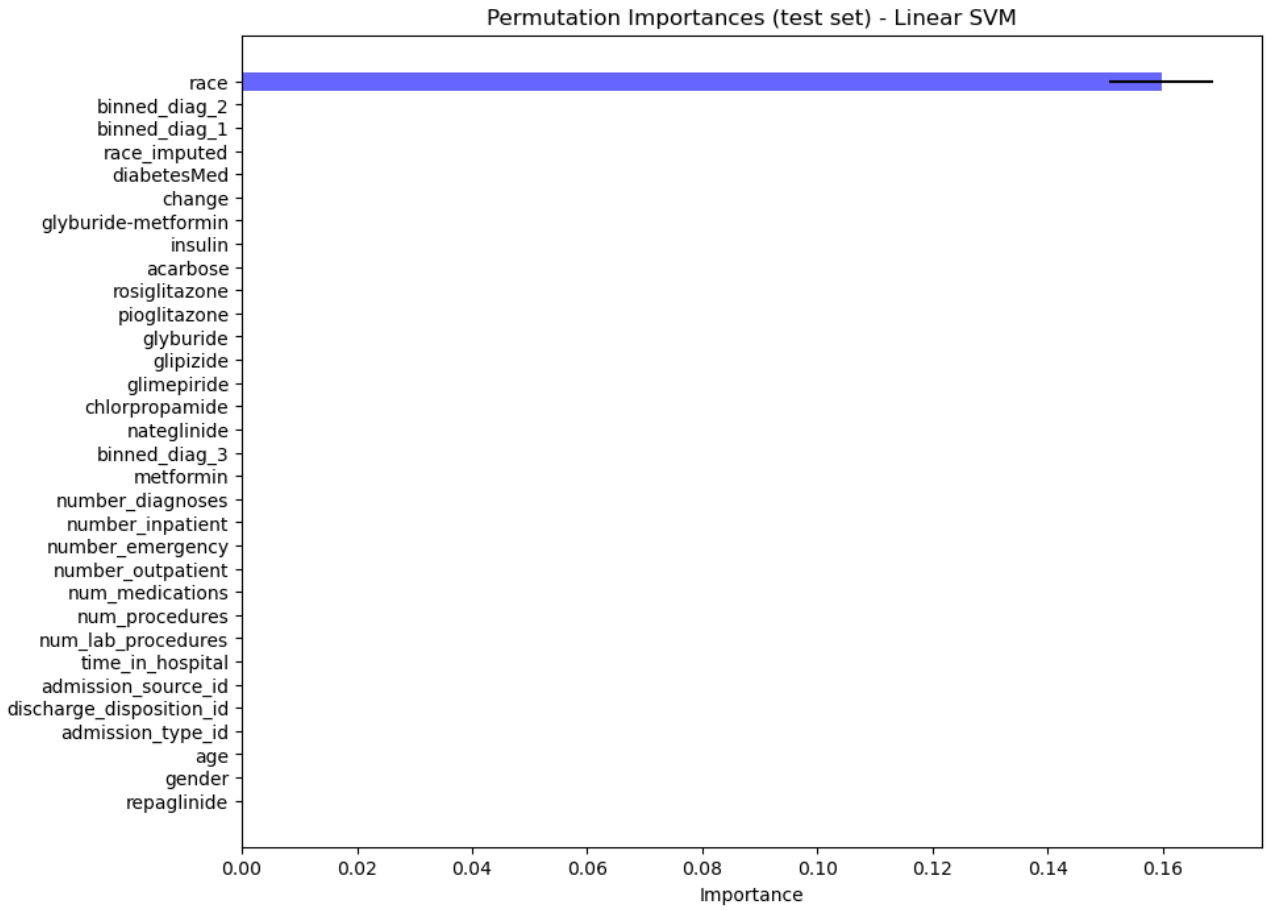

Figure 20: **Mean and standard deviation permutation importances of Linear SVM:** As can be seen in the aggregate version, both SVMs (see **??** prioritize race to make their predictions. This is consistent with our analysis of the Linear SVM's coefficient weights in Figure 18.

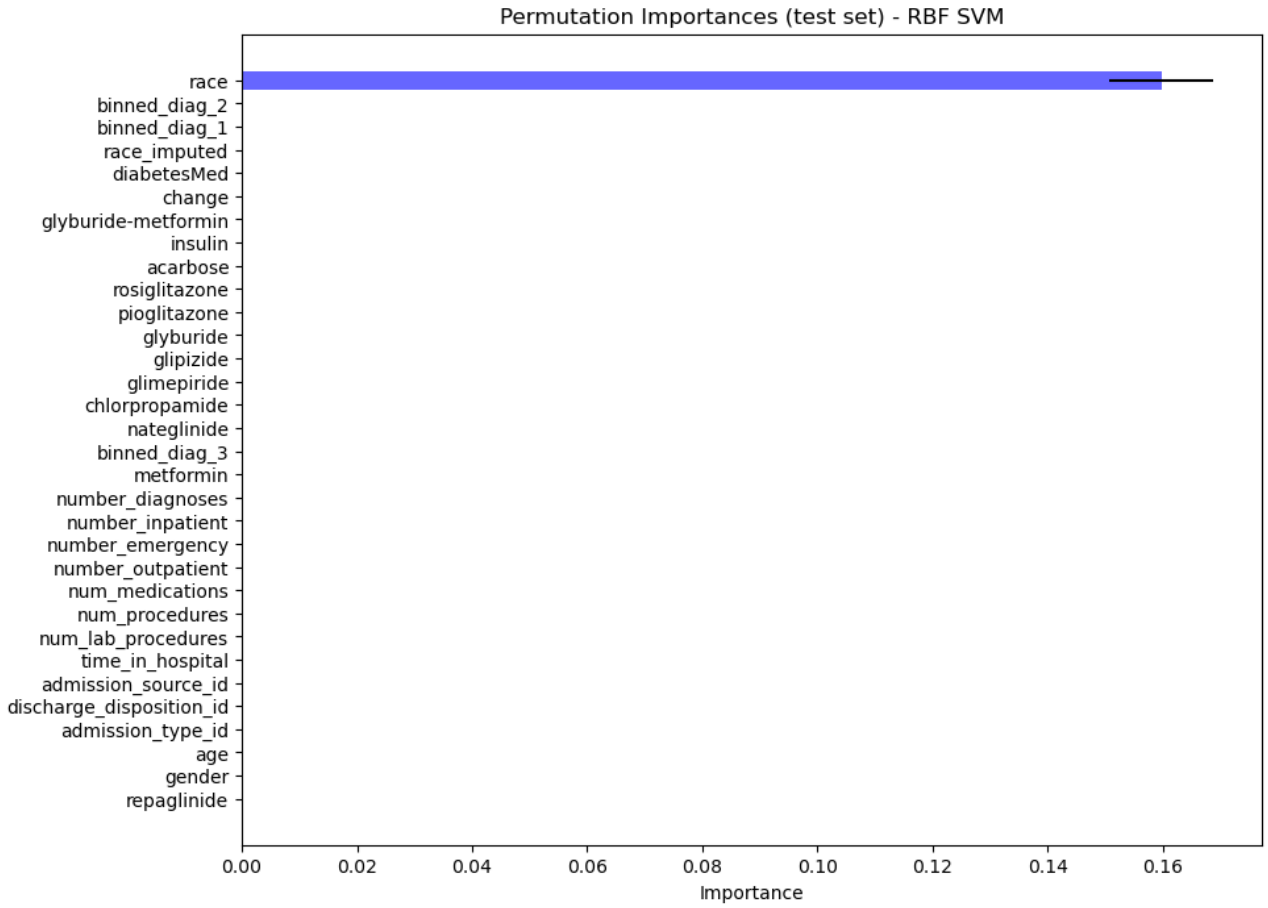

Figure 21: **Mean and standard deviation permutation importances of RBF SVM.**

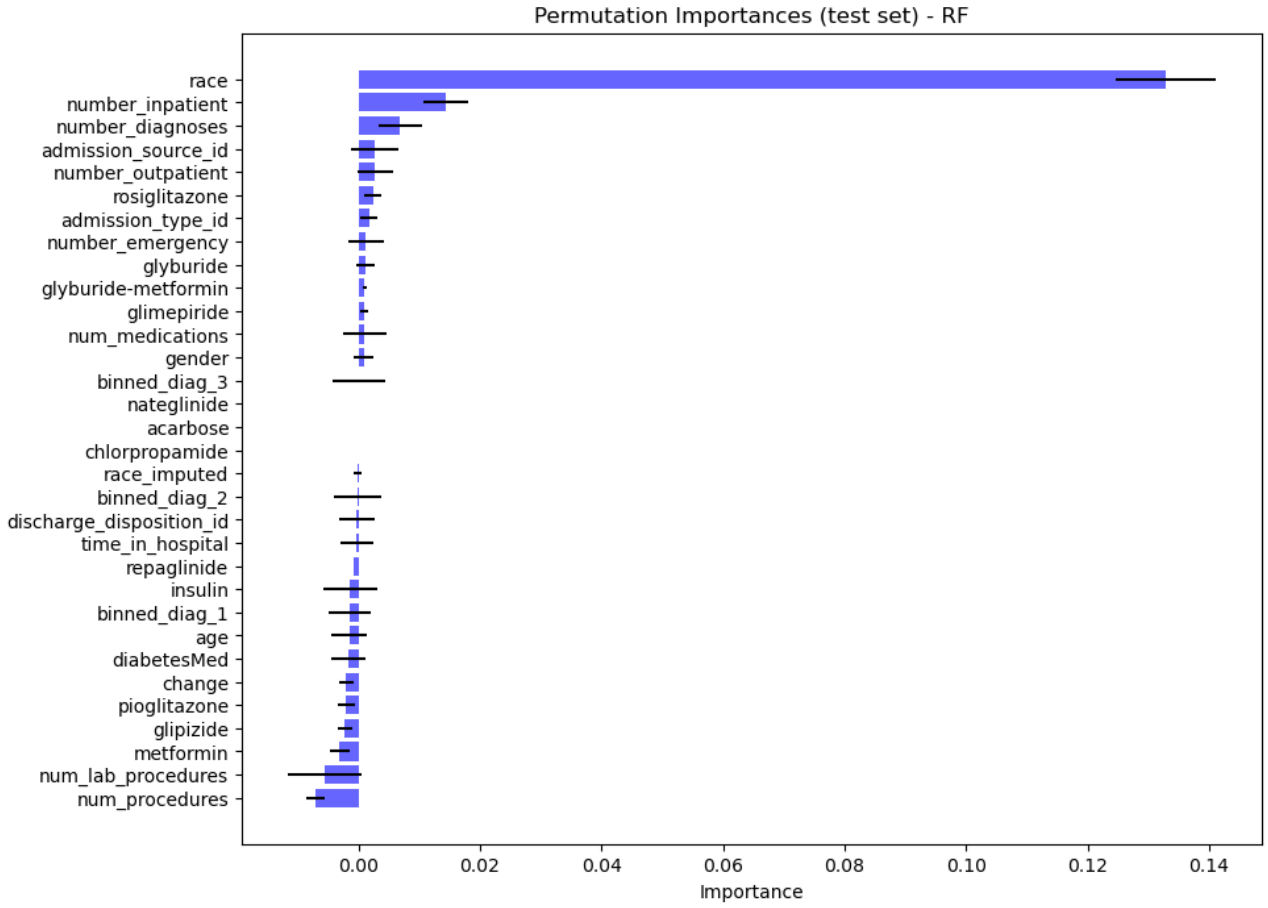

Figure 22: **Mean and standard deviation permutation importances of RF:** Interesting notes include the prominence of some drugs, and of the primary diagnosis as opposed to the secondary/tertiary diagnoses. This indicates it may be interesting to explore our suggestion of retaining the diagnosis order structure in Section 3.

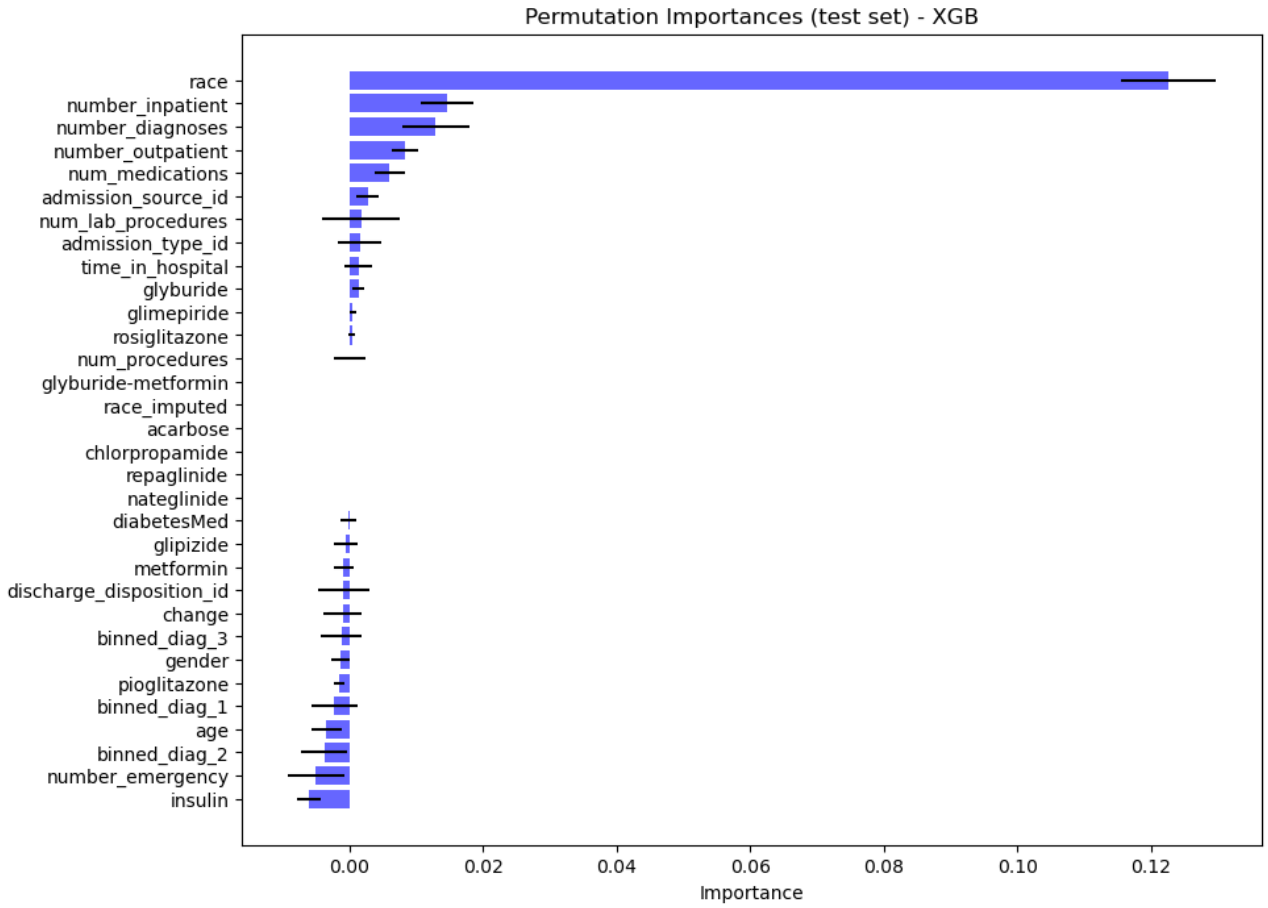

Figure 23: **Mean and standard deviation permutation importances of XGB.**

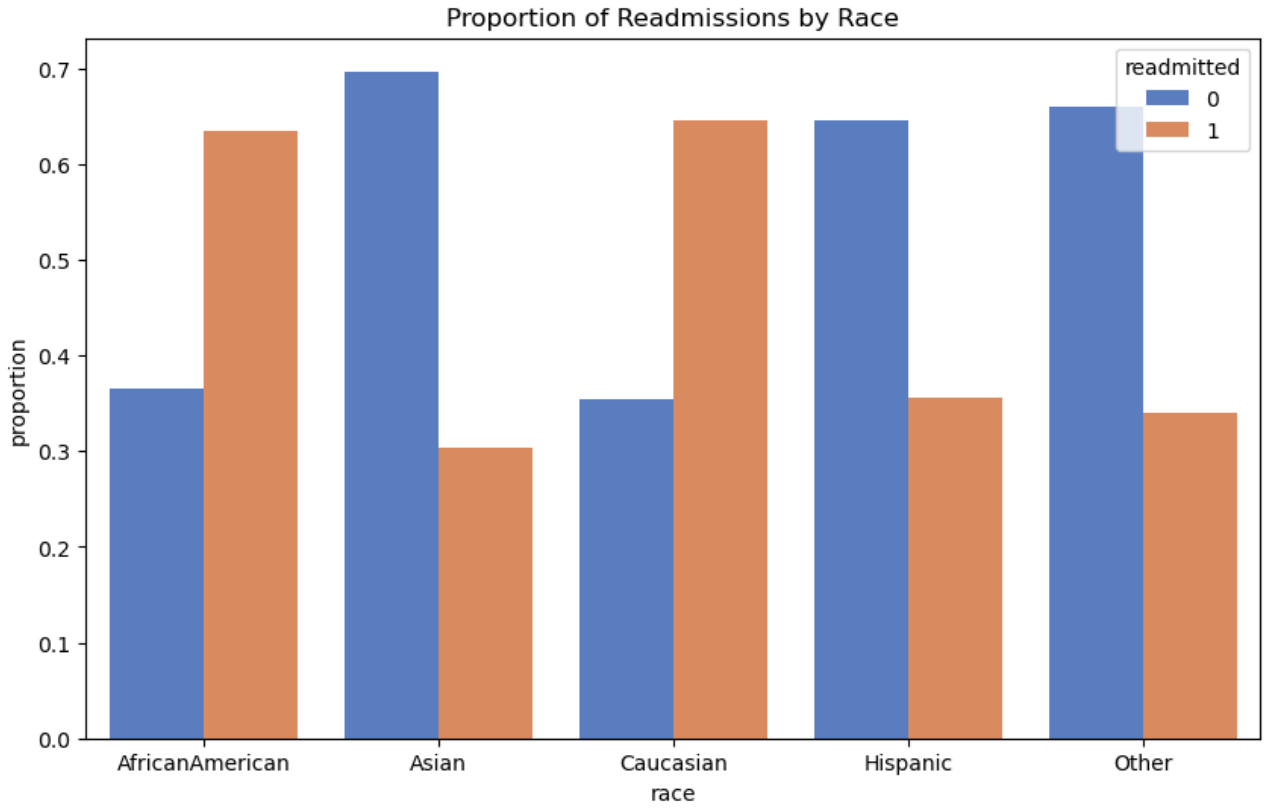

Figure 24: **Bar chart showing proportion of readmitted patients by race in subsample:** As shown here, although our overall subsample balanced our target, there was a significant difference in the proportion of readmitted patients by race. It seems that our models are then heavily relying on this feature to predict readmission, which is undesirable behaviour as discussed in Section 4.

Table 6: Nested cross-validation test scores for each model

Model	**Mean Test F2**	**Std Test F2**	**Mean Test Accuracy**	**Std Test Accuracy**
Linear SVM	0.618	0.012	0.650	0.012
RBF SVM	0.621	0.015	0.653	0.011
RF	0.636	0.012	0.665	0.014
XGB	0.634	0.018	0.656	0.014

Table 7: **Best hyperparameters for each model in nested CV:** For each of the five iterations of the outer loop, GridSearchCV gave one set of best hyperparameters after running on 5 inner cross-validation folds.

Model	**Best Parameters**
Linear SVM	{'classifier__C': 0.01} (all folds)
RBF SVM	{'classifier__C': 100, 'classifier__gamma': 0.01} (all folds)
RF	{'classifier__max_depth': 20} (all folds)
	'classifier__n_estimators': 1000 (two folds)
	'classifier__n_estimators': 100 (one fold)
	'classifier__n_estimators': 200 (one fold)
	'classifier__n_estimators': 500 (one fold)
XGB	{'classifier__learning_rate': 0.1} (all folds)
	{'classifier__max_depth': None, 'classifier__n_estimators': 1000} (two folds)
	{'classifier__max_depth': None, 'classifier__n_estimators': 50} (one fold)
	{'classifier__max_depth': 5, 'classifier__n_estimators': 50} (one fold)
	{'classifier__max_depth': 5, 'classifier__n_estimators': 100} (one fold)

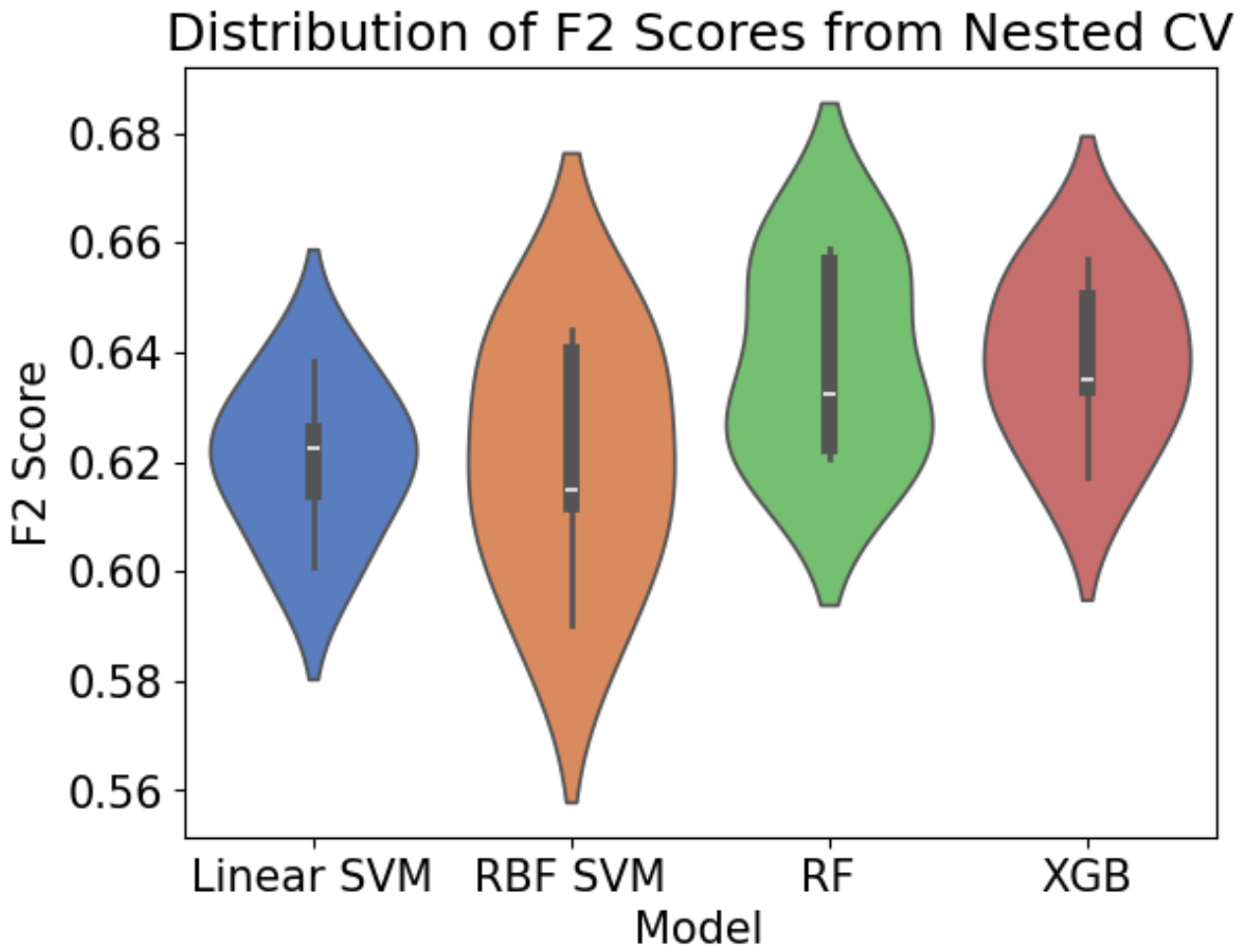

Figure 25: **Distribution of F2 scores on nested CV test sets:** This indicates the F2 score results from each test set in the nested cross-validation folds. The extent of variation here is quite different from 14, which can be attributed to the significant increase in the number of times we ran our models. There is a clearer difference in performance overall, whereas there is a smoother depiction of variance.

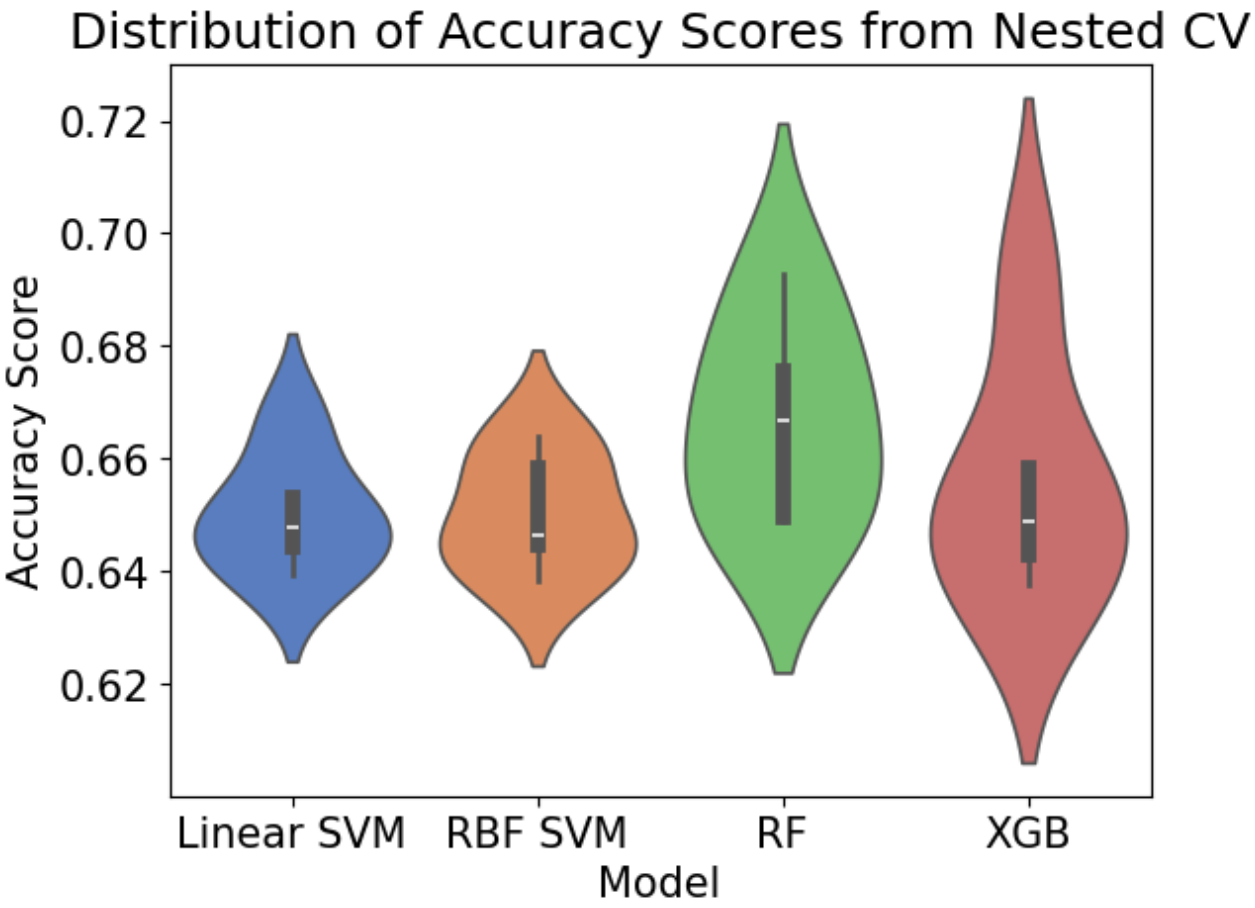

Figure 26: **Distribution of accuracy scores on nested CV test sets:** This indicates the accuracy results from each test set in the nested cross-validation folds. There is similar behaviour here as to that we observed for the F2 score, although there is less variance in the SVM models and overall lower performance of the XGB model.

References

[1] Diagnostic Code Descriptions (ICD-9) - Province of British Columbia. URL: https://www2.gov.bc.ca/gov/content/health/practitioner-professional-resources/msp/physicians/diagnostic-code-descriptions-icd-9.

[2] Richard J. Chen, Judy J. Wang, Drew F.K. Williamson, Tiffany Y. Chen, Jana Lipkova, Ming Y. Lu, Sharifa Sahai, and Faisal Mahmood. Algorithm fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719, 6 2023. URL: /pmc/articles/PMC10632090//pmc/articles/PMC10632090/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC10632090/, doi:10.1038/S41551-023-01056-8.

[3] Joseph Futoma, Jonathan Morris, and Joseph Lucas. A comparison of models for predicting early hospital readmissions. *Journal of biomedical informatics*, 56:229–238, 8 2015. URL: https://pubmed.ncbi.nlm.nih.gov/26044081/, doi:10.1016/J.JBI.2015.05.016.