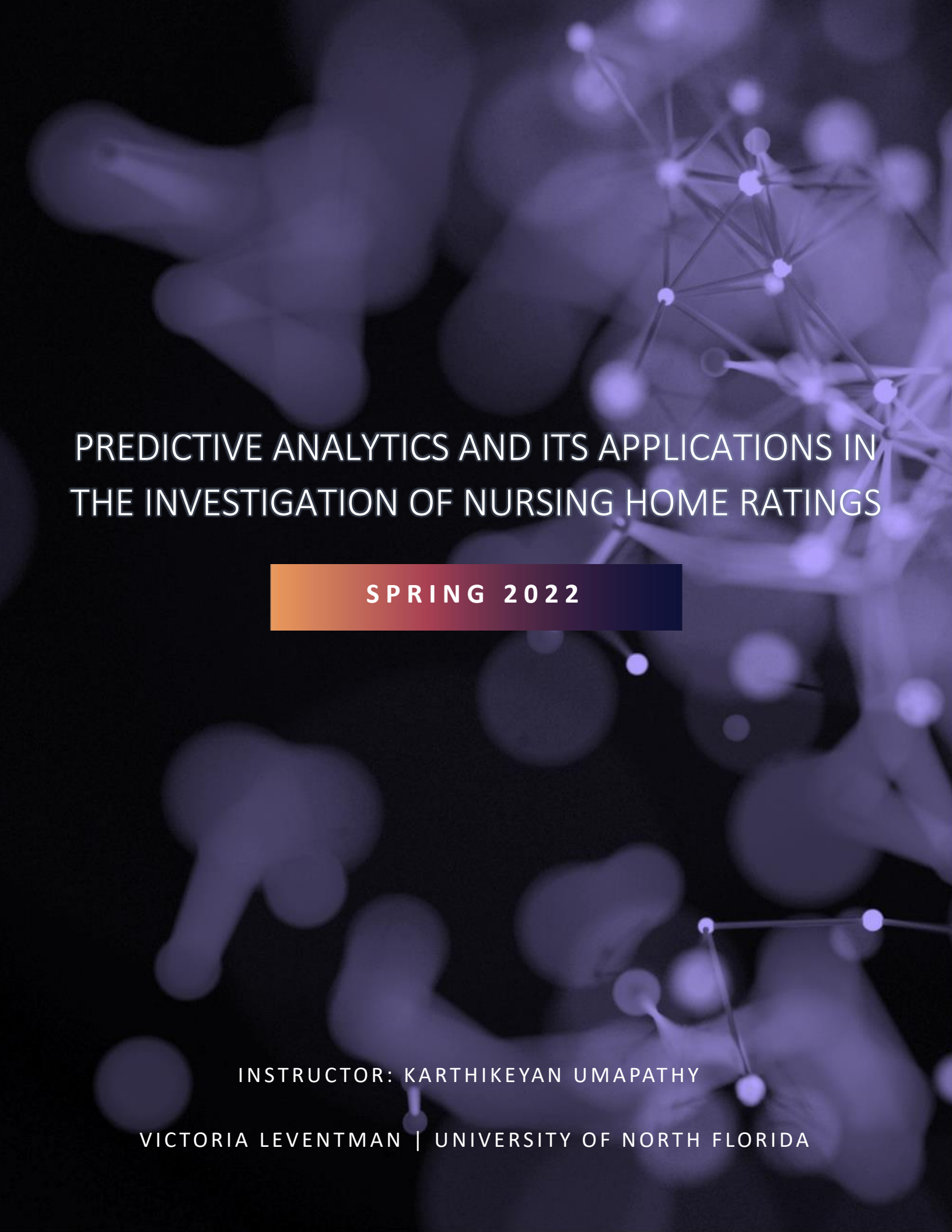


PREDICTIVE ANALYTICS AND ITS APPLICATIONS IN THE INVESTIGATION OF NURSING HOME RATINGS

SPRING 2022

INSTRUCTOR: KARTHIKEYAN UMAPATHY

VICTORIA LEVENTMAN | UNIVERSITY OF NORTH FLORIDA

TABLE OF CONTENTS

Introduction

In the United States of America, family members of the elderly often look to place their loved ones in nursing homes when the family members cannot be the primary caregiver. Families can research online for multiple nursing homes in their local area before deciding on the final one. The primary website to conduct this search would be the Nursing Home Care Compare website. In their decision-making process, family members of the seniors will consider the overall ratings as one of the factors when choosing a nursing home. With overall ratings expressed as a composite star rating from one to five, one can easily scan through a list of local nursing homes and see which ones have a higher number of stars versus a lower number of stars. More specifically, in the Five-Star Quality Rating System, there are nursing homes listed as either much above average (five stars), above average (four stars), average (three stars), below average (two stars), or much below average (one star).

As stated on the Nursing Home Care Compare website, the overall rating is calculated from three other ratings, which are the health inspections rating, quality measures (QM) rating, and staffing rating. As defined by the Centers for Medicare and Medicaid Services (CMS), the calculation is below [11]:

Step 1: **Start with the health inspections rating.**
Step 2: **Add 1 star if the staffing rating is 4 or 5 stars and greater than the health inspections rating. Subtract 1 star if the staffing rating is 1 star.**
Step 3: **Add 1 star if the quality measures rating is 5 stars; subtract 1 star if the quality measures rating is 1 star.**
Step 4: **If the health inspections rating is 1 star, then the overall rating cannot be upgraded by more than 1 star based on the staffing and quality measures ratings.**
Step 5: **If a nursing home is a special focus facility, all star ratings are suppressed.**

Fig. 1 CMS's methodology for overall star rating calculation.

For those viewing the listings of nursing home facilities on the Nursing Home Care Compare website, it can be discernable to them that a higher star rating is a more positive indicator of residential health and quality of life in contrast to a much lower star rating. For each nursing home, the inspection details that were considered in the calculations of the three individual ratings (health inspections, staffing, quality measures) can be further examined on the website. Alongside this information, it would be helpful for the user to have a description under "Overall Rating" that highlights which subset of the collected nursing home data (comprised of factors) was most important in determining the final rating. Here, the importance of a factor shall be defined as the magnitude of its effect on the outcome, with the outcome being a nursing home's overall rating.

The Problem

The Five-Star Quality Rating System takes into account three domains and their corresponding ratings: health inspections, staffing, and quality measures (QM) [5]. Within each of the three domains, numerous factors are measured and considered in the computation of those ratings. For instance, consider the following: Resident Rights deficiencies, the number of hospitalizations per 1000 long-stay resident days, and reported RN staffing hours per resident per day. These features are taken into account for the respective ratings of health inspections, QMs, and staffing (which ultimately result in the assignment of an overall rating), but we would like to drill down and harness the power of machine learning. The

primary goal of this investigation is to see if a machine learning (ML) model can find patterns in the data, which excludes the calculated ratings from those three domains, and correctly predict a nursing home's overall rating. The model can provide us insights on what type of factors most differentiate nursing homes with 1- or 2-star overall ratings from nursing homes with 4 or 5 stars; these factors would be the ones that have the greatest feature importances in the determination of a rating. The motivation behind this problem is that if a subset of collected data has the most importance as determined by the ML model, nursing homes can then be proactive in terms of not possessing those deficiencies for (unannounced) future inspections. After inputting the data collected during inspections and the data self-reported by nursing homes into our ML model to obtain predictions, we could then deliver focused information back to the inspectors. Accordingly, the inspectors can notify and advise a nursing home on what deficiencies they should correct to obtain a higher overall rating next time.

The Data

Original Datasets

Five datasets from the Centers for Medicare and Medicaid Services (CMS) were considered as part of this study. These original datasets are all in CSV format, and they are titled as follows: *Provider Information*, *Health Deficiencies*, *Survey Summary*, *MDS Quality Measures*, and *Medicare Claims Quality Measures*. It is important to note that the nursing home data from CMS is comprised only of the nursing home facilities that participate in Medicare and Medicaid programs. For each of the mentioned datasets, their data quantity characteristics are provided below.

Dataset name	*Provider Information*	*Health Deficiencies*	*Survey Summary*	*MDS Quality Measures*	*Medicare Claims Quality Measures*
Number of rows	15,216	372,926	45,260	273,888	60,864
Number of columns	96	21	41	23	17

Fig. 2 Data quantities for the original datasets.

The primary distinction among the five datasets at a high level is the number of records found for each nursing home. The *Provider Information* dataset contains one record per nursing home, and thus there are 15,216 total nursing homes. Those 15,216 nursing homes can appear as more than one record in the remaining four datasets, which explains the increased number of rows.

For *MDS Quality Measures*, each nursing home has eighteen records for each of the eighteen possible quality measures (QMs) that encompass the Minimum Data Set (MDS) assessments. Similarly, for *Medicare Claims Quality Measures*, four records are present for each nursing home as there are only four QMs that encompass the Medicare claims data. For *Health Deficiencies*, the number of records per nursing home is dependent on how many individual health-related citations they received in the past three years for the three most recent inspections. Additionally, *Survey Summary* contains nursing home data for the three most recent inspections, and hence there are at most three records per nursing home in this particular dataset. It is worth noting that newly opened or newly approved nursing homes for Medicare/Medicaid services may have had only one health inspection documented thus far by CMS. Moreover, a newer nursing home could have received zero citations for health deficiencies thus far, and hence, they do not appear in the *Health Deficiencies* dataset.

Data Preparation

Feature Selection

Our main goal during feature selection was to include only features/attributes that relate to nursing home quality of life and care and the residents within those facilities. The motivation behind this is that nursing homes can prioritize their resources on what is most important when taking action on the findings provided by the ML model.

For *Provider Information*, we selected a subset of fifteen features. This subset of features was deliberately chosen after reading the Technical Users' Guide for the Five-Star Quality Rating System of nursing homes [5] and understanding better which factors contribute to the rating calculations of the three domains (health inspection, staffing, quality measure). The list of the fifteen features and their data types are available in the Appendix. We list five of those features here: Reported RN Staffing Hours per Resident per Day, Registered Nurse hours per resident per day on the weekend, Total nursing staff turnover, Number of Facility Reported Incidents, and Number of Substantiated Complaints.

It is relevant to note that we excluded fire safety data from *Provider Information* since it is not accounted for in the overall rating. Furthermore, we excluded any attributes related to penalties (i.e., fines and payment denials) since they focus more on outcomes rather than being descriptive characteristics of the nursing home facility itself. Since this is an investigation into which nursing home-specific features most affect the overall rating, we excluded the columns for the ratings (Health Inspection Rating, QM Rating, Long-stay QM Rating, Short-Stay QM Rating, Staffing Rating, RN Staffing Rating) that are involved in the composite measure calculation of Overall Rating.

For *Survey Summary*, we selected to have the numeric counts for ten health deficiency categories. The full list of the ten features and their data types are available in the Appendix. We list five of those features here: Count of Freedom from Abuse and Neglect and Exploitation Deficiencies, Count of Quality of Life and Care Deficiencies, Count of Resident Assessment and Care Planning Deficiencies, Count of Nursing and Physician Services Deficiencies, and Count of Resident Rights Deficiencies.

For *Health Deficiencies*, we only selected one categorical feature: Scope Severity Code. The possible codes for categorizing a deficiency are denoted by letters A through L, which range from least severe to most severe in regards to its effect on one or more nursing home residents. While we have the counts for different deficiency categories for a nursing home in *Survey Summary*, the Scope Severity Code column allows us to analyze in greater detail the frequencies of the severities for those deficiencies.

For *MDS Quality Measures*, we selected two features: Measure Description and Four Quarter Average Score. Measure Description is a categorical feature containing descriptions for eighteen different quality measures, and Four Quarter Average Score is a numeric feature. Out of the eighteen unique measure descriptions, five of them are listed here: Percentage of long-stay residents whose need for help with daily activities has increased, Percentage of long-stay residents who lose too much weight, Percentage of long-stay residents who have depressive symptoms, Percentage of long-stay

residents who received an antipsychotic medication, and Percentage of short-stay residents who made improvements in function. For MDS assessments, nine of the quality measures are used in the QM rating calculation, while the remaining nine measures are not. All flu and pneumonia prevention measures [9] are not used in the QM rating calculation. The list of the MDS-based measures (long-stay and short-stay) is available in the Appendix. The reasoning behind also including QMs that are not accounted for in the QM rating is so that we can use machine learning to explore and examine any possible hidden correlations between these measures and the overall rating.

Similarly, for *Medicare Claims Quality Measures*, we selected two features: Measure Description and Adjusted Score. Measure Description is a categorical feature containing descriptions for four distinct quality measures from claims data, and Adjusted Score is a numeric feature. The four unique measure descriptions are listed here: Percentage of short-stay residents who were re-hospitalized after a nursing home admission, Percentage of short-stay residents who had an outpatient emergency department visit, Number of hospitalizations per 1000 long-stay resident days, and Number of outpatient emergency department visits per 1000 long-stay resident days. All four claims-based measures (long-stay and short-stay) are included in the QM rating calculation.

For each dataset, here are the first five rows of data for key features that we chose.

Provider Information

Reported RN Staffing Hours per Resident per Day	Registered Nurse hours per resident per day on the weekend	Total nursing staff turnover	Number of Facility Reported Incidents	Number of Substantiated Complaints
NaN	NaN	NaN	0	0
0.89834	0.52021	52.4	0	0
NaN	NaN	NaN	0	0
0.45553	0.31108	NaN	0	0
0.35595	0.27210	51.3	0	1

Survey Summary

Count of Freedom from Abuse and Neglect and Exploitation Deficiencies	Count of Quality of Life and Care Deficiencies	Count of Resident Assessment and Care Planning Deficiencies	Count of Nursing and Physician Services Deficiencies	Count of Resident Rights Deficiencies
0	0	0	0	1
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	1	1	0	0

Health Deficiencies

Scope Severity Code
D
D
D
D
D

MDS Quality Measures

Measure Description	Four Quarter Average Score
Percentage of long-stay residents whose need for help with daily activities has increased	14.792900
Percentage of long-stay residents who lose too much weight	6.249999
Percentage of low risk long-stay residents who lose control of their bowels or bladder	70.000000
Percentage of long-stay residents with a catheter inserted and left in their bladder	0.408016
Percentage of long-stay residents with a urinary tract infection	0.540540

Medicare Claims Quality Measures

Measure Description	Adjusted Score
Percentage of short-stay residents who were rehospitalized after a nursing home admission	19.653524
Percentage of short-stay residents who had an outpatient emergency department visit	2.159487
Number of hospitalizations per 1000 long-stay resident days	0.888825
Number of outpatient emergency department visits per 1000 long-stay resident days	0.298808
Percentage of short-stay residents who were rehospitalized after a nursing home admission	16.512389

Data Quality

Given that we aimed to predict the overall rating for a nursing home based on a set of collected and reported features, we examined the data quality characteristics of the selected features (mentioned above) for the five datasets. 214 nursing home instances had no overall rating in the *Provider Information* dataset, and since they are not valuable to our investigation, we discarded them. It results in 15,002 facilities remaining for us to work with and analyze. The meaning behind such missing values (for overall rating) is that they were either new nursing homes with not enough gathered data to perform the rating calculation or there were severe and persistent quality problems at the facilities for at least three years. For the latter, these nursing home facilities are placed into a separate, dedicated SFF program and thus not considered in the Five-Star Quality Rating System [4]. From

now on, when discussing the *Provider Information* dataset, we will be referring to the "updated" *Provider Information* containing data on 15,002 facilities.

Additionally, the (updated) *Provider Information* dataset contained the following unique values in the Rating Cycle 3 Number of Health Revisits feature: ".", 0, 1, 2, 3, and 4. After further inspection, we found that 130 nursing homes have this dot value for the mentioned feature. The dot represents that a nursing home has not yet had a Rating cycle 3 Standard Health Survey, but it is upcoming. This explains why those particular nursing homes with the "." also have null values for Rating Cycle 3 Standard Health Survey Date.

Furthermore, in *Provider Information*, the three features concerning turnover data (i.e., Total nursing staff turnover, Registered Nurse turnover, and Number of administrators who have left the nursing home) had more missing values when compared to the other features we selected. There were 2382, 3193, and 3096 missing values, respectively. According to the Nursing Home (NH) Data Dictionary (NH_Primary_Data_Dictionary.xlsx) from CMS, the meaning behind those missing values is that "this facility did not submit staffing data, or submitted data that did not meet the criteria required to calculate a staffing measure." In the Appendix, we have the full list of data quality details and the meaning behind missing values for the selected features in *Provider Information*.

For the remaining datasets, their data quality details on the selected features are stated below.

Dataset name	*MDS Quality Measures*
Number of missing values in "Four Quarter Average Score" column	12,901
Number of nursing homes that have values for all 18 QMs	11,679

Based on the NH Data Dictionary, the meaning behind such missing values for *MDS Quality Measures* is that either the number of residents falling under that QM is a very small number or the data needed for that QM was not submitted or was missing. It is important to note that "MDS assessments are completed by staff at the nursing facility" [8]. Hence, the lack of submission is plausible in situations where information for some QMs was not properly recorded by the staff during these assessments.

Dataset name	*Medicare Claims Quality Measures*
Number of missing values in "Adjusted Score" column	10,350
Number of nursing homes that have values for all 4 QMs	10,899

The meaning behind such missing values in *Medicare Claims Quality Measures* is the same as what we discussed earlier with the *MDS Quality Measures* dataset. Either the percentage of residents falling under that QM is such a small number or the data needed for that QM was not submitted or was missing.

Given that we are mainly interested in using machine learning to investigate feature importances rather than deploying a predictive model, we concluded that it was acceptable to remove the nursing homes that do not have values for all QMs in each of the two respective datasets, *MDS Quality Measures* and *Medicare Claims Quality Measures*. 10,000+ nursing homes is an appropriate dataset size for such ML applications. Furthermore, *Health Deficiencies* did not have any missing values for the "Scope Severity Code" column, and *Survey Summary* did not have any missing values for the ten health deficiency categories that we selected.

Dataset Restructuring and Data Transformations

The rest of our discussion is focused on the same five datasets, but we now refer to the "updated" versions of *Provider Information*, *MDS Quality Measures*, and *Medicare Claims Quality Measures*. "Updated" denotes how we removed nursing homes from the original set of 15,216 nursing home facilities. In preparation for machine learning, the five datasets should eventually be merged into a single dataset.

First, we needed to transform the key data features in *Survey Summary*, *Health Deficiencies*, *MDS Quality Measures*, and *Medicare Claims Quality Measures* into a format that can later be merged by Federal Provider Number, the primary key, to create our machine learning dataset. The final machine learning dataset would contain one record (row) per nursing home. Thus, the nursing homes are the instances in the dataset, where each nursing home has multiple features describing it.

For both *MDS Quality Measures* and *Medicare Claims Quality Measures*, we reshaped the datasets by pivoting from long to wide (i.e., performing a pivot-wider operation). By performing this pivot-wider operation, we obtained one record per nursing home since the QMs and their corresponding percentages became separate columns. This data restructuring is demonstrated below for the nursing home facility 015009 in the *Medicare Claims Quality Measures* dataset.

Federal Provider Number	Measure Description	Adjusted Score
015009	Percentage of short-stay residents who were rehospitalized after a nursing home admission	19.653524
015009	Percentage of short-stay residents who had an outpatient emergency department visit	2.159487
015009	Number of hospitalizations per 1000 long-stay resident days	0.888825
015009	Number of outpatient emergency department visits per 1000 long-stay resident days	0.298808

Performing pivot-wider operation

Measure Description	Number of hospitalizations per 1000 long-stay resident days	Number of outpatient emergency department visits per 1000 long-stay resident days	Percentage of short-stay residents who had an outpatient emergency department visit	Percentage of short-stay residents who were rehospitalized after a nursing home admission
Federal Provider Number				
015009	0.888825	0.298808	2.159487	19.653524

Overall, for the *MDS Quality Measures* transformed dataset, it has 11,679 rows and 18 columns (which represent the 18 MDS-based QMs). The *Medicare Claims Quality Measures* transformed dataset has 10,899 rows and 4 columns (which represent the four claims-based QMs). In both datasets, each instance is an individual nursing home.

For *Health Deficiencies*, we grouped by Federal Provider Number and Scope Severity Code in order to obtain the value counts per code, per nursing home. Then, with this aggregated data, we had to pivot from long to wide to get one unique record per facility. The transformed dataset has 15,087 rows and 11 columns. Below shows a small sample of this aggregated, reshaped dataset.

Scope Severity Code	Code B	Code C	Code D	Code E	Code F	Code G	Code H	Code I	Code J	Code K	Code L
Federal Provider Number											
015009	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
015010	0.0	1.0	8.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0
015012	0.0	0.0	4.0	3.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0

Fig. 3 First three rows of the restructured *Health Deficiencies* dataset.

For *Survey Summary*, we grouped by Federal Provider Number and then carried out aggregation with the sum operation to obtain the total counts for each of the health deficiency categories per nursing home facility. The transformed dataset has one record per facility, and it contains 15,216 rows and 10 columns. Below shows a small sample of this aggregated dataset.

	Count of Freedom from Abuse and Neglect and Exploitation Deficiencies	Count of Quality of Life and Care Deficiencies	Count of Resident Assessment and Care Planning Deficiencies	Count of Nursing and Physician Services Deficiencies	Count of Resident Rights Deficiencies
Federal Provider Number					
15009	0	0	0	0	1
15010	0	2	3	1	1
15012	0	1	0	0	1

Fig. 4 First three rows of the restructured *Survey Summary* dataset.

For *Provider Information*, rather than having three separate features for the Number of Health Revisits, one for each rating cycle (1-3), we decided that it would be more appropriate to sum across the cycles to create a new feature named "Total Health Revisits." This summation across rows still preserves the overall dataset structure of having one record per nursing home. The restructured *Provider Information* dataset now contains 14 columns in total rather than 17. Below displays a small sample of this dataset.

	Registered Nurse turnover	Number of administrators who have left the nursing home	Total Health Revisits	Number of Facility Reported Incidents	Number of Substantiated Complaints	Number of Citations from Infection Control Inspections
Federal Provider Number						
015009	NaN	NaN	2	0	0	0.0
015010	21.4	1.0	3	0	0	0.0
015012	NaN	NaN	3	0	0	0.0

Fig. 5 First three rows of the transformed *Provider Information* dataset.

Data Merging

The remaining discussion of merging multiple datasets will focus on the transformed datasets. For the merging process, we first considered *Survey Summary* and *Health Deficiencies*. We performed a left-join operation, where the left DataFrame is *Survey Summary*, and the right DataFrame is *Health Deficiencies*. The reason behind choosing left join is because *Survey Summary* has inspection data of counts for various health deficiencies, while some nursing homes did not receive any deficiency citations. Hence, these nursing homes would be present in *Survey Summary* as a record (with all counts as 0) but not in *Health Deficiencies* since the latter dataset only contains facilities that received one or more citations during Health Inspection Surveys. For the 129 non-matching rows returned by the left join, we simply filled in the null values with 0.

Then, with the 15,216 rows from that merged dataset (of two DataFrames), we performed an inner-join operation with Provider Information, which contains only 15,002 rows, to retrieve only the matching rows. On a separate note, we did a left join on *MDS Quality Measures* (with 11,679 rows) and *Medicare Claims Quality Measures* (with 10,899 rows) to retrieve all matching and non-matching rows. Any null values that resulted from the merging will later be replaced via an imputation technique inside a Scikit-learn Pipeline.

Lastly, we performed an inner-join operation to combine the merged dataset containing 15,002 rows with another merged dataset, generated from the left join on *MDS Quality Measures* and *Medicare Claims Quality Measures*, containing 11,679 rows. During the data integration, there were no merging issues to be documented. Most importantly, this whole merging process helped create our final machine learning dataset, which is discussed in more detail in the next section. Since the overall star rating is based on health inspections, quality measures, and staffing, we took note of how many features we have in the ML dataset corresponding to each of the three domains. These details are

indicated below, and the full list of features and their corresponding color-coding based on the domains can be found in the Appendix.

Domain	Health Inspections	Quality Measures	Staffing
Number of features	25	22	9

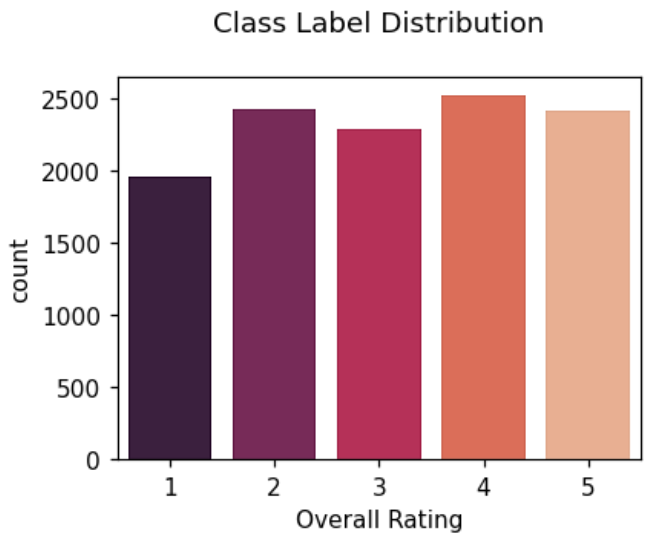
Machine Learning Dataset

In our ML dataset, there are 11,585 nursing home instances, 56 features, and the target variable. All 56 features are numeric. Due to the curse of dimensionality that severely impacts model performance, further feature selection would be carried out within a Scikit-learn pipeline. The overall rating column is our target variable, which contains five unique class labels in total, and they are the following: 1, 2, 3, 4, and 5. An overall rating of 1 is the lowest rating that a nursing home can receive, while a rating of 5 is the highest. The average overall rating across the facilities is 3.088, with the median rating being 3. Since precisely one label (out of many classes) is used per nursing home instance, this is a multi-class classification problem.

Overall Rating	1	2	3	4	5
Instances per Label	1951	2421	2283	2520	2410

Fig. 6 Number of nursing home instances per class label.

The following count plot depicts the class label distribution in the ML dataset. Based on the plot, the classes are not severely imbalanced. While there are fewer nursing home facilities with label 1, it will not negatively impact model performance, and thus no balancing needs to be done.



Data Exploration

It was beneficial to perform exploratory data analysis on our ML dataset to gain further insight into the features/attributes before creating any machine learning models and training them. Hence, we first

examined the descriptive statistics for the 56 numeric features; the full list is available in the Appendix. After observing the minimum and maximum values for each of the features, we concluded that there were no erroneous data points, such as negative counts and percentages being higher than 100% or less than 0%. The table below shows descriptive statistics such as mean, median, standard deviation, minimum value, and maximum value for a subset of attributes.

	Mean	Median	Standard Deviation	Min	Max
Count of Quality of Life and Care Deficiencies	4.931204	4.000000	4.200128	0.000000	36.000000
Count of Resident Rights Deficiencies	3.068968	2.000000	2.930242	0.000000	24.000000
Count of Code D	16.135175	13.000000	12.478989	0.000000	96.000000
Count of Code E	6.013897	4.000000	6.491061	0.000000	80.000000
Total Health Revisits	2.831679	3.000000	0.689163	0.000000	7.000000
Percentage of long-stay residents whose ability to move independently worsened	22.371544	21.906851	10.224248	0.000000	74.994635
Percentage of short-stay residents who had an outpatient emergency department visit	10.496168	9.868402	5.338787	0.000000	41.813369

Fig. 7 Summary statistics for seven features in the ML dataset.

Then, as part of the data exploration, we further examined the features which had the biggest differences between their mean and median value since this often indicates a skewed distribution and possibly the presence of outliers. Based on the box plot visualizations we created, two features had one or more noticeable data points located significantly far away from the rest of the observations. These two features were the following: Count of Code E and Number of Substantiated Complaints.

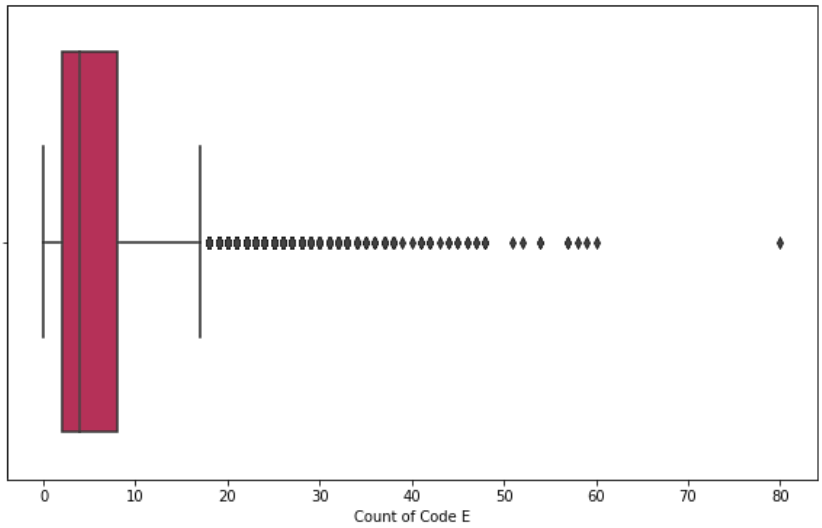


Fig. 8 Box plot for the "Count of Code E" feature.

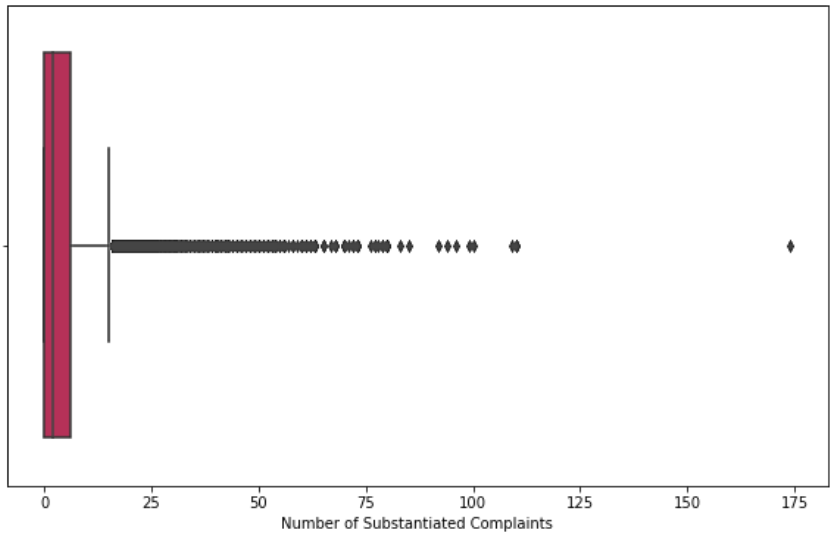
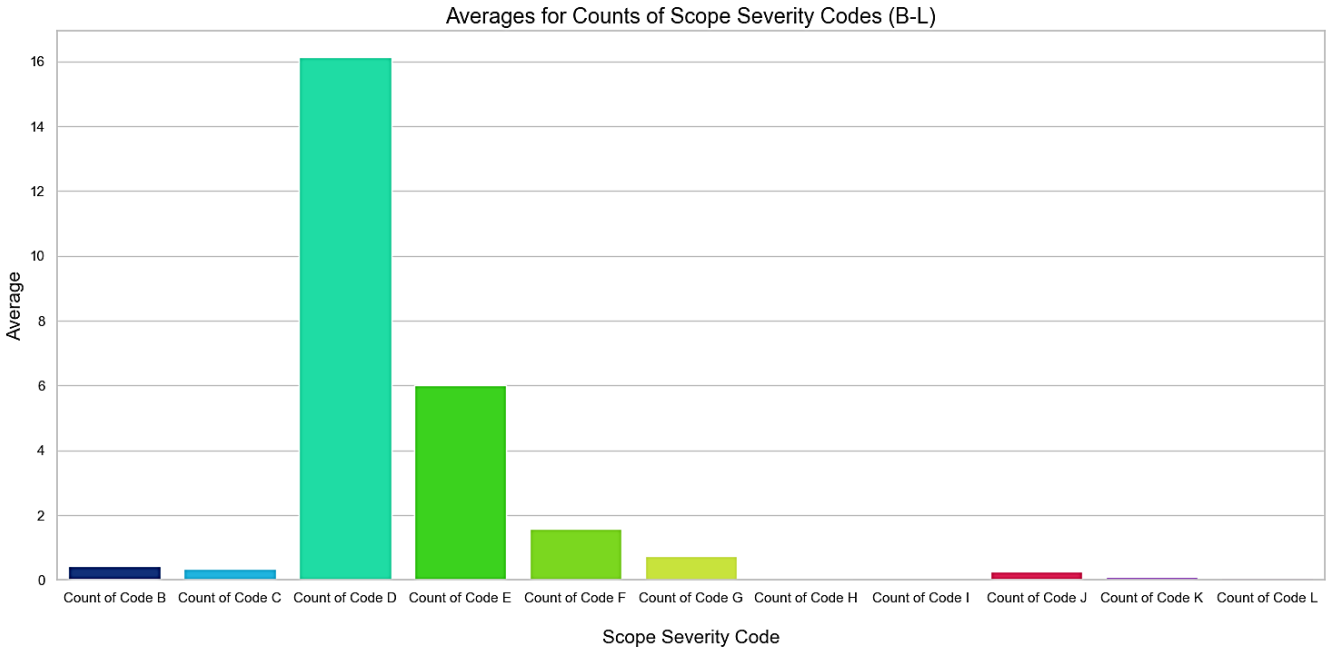


Fig. 9 Box plot for the "Number of Substantiated Complaints" feature.

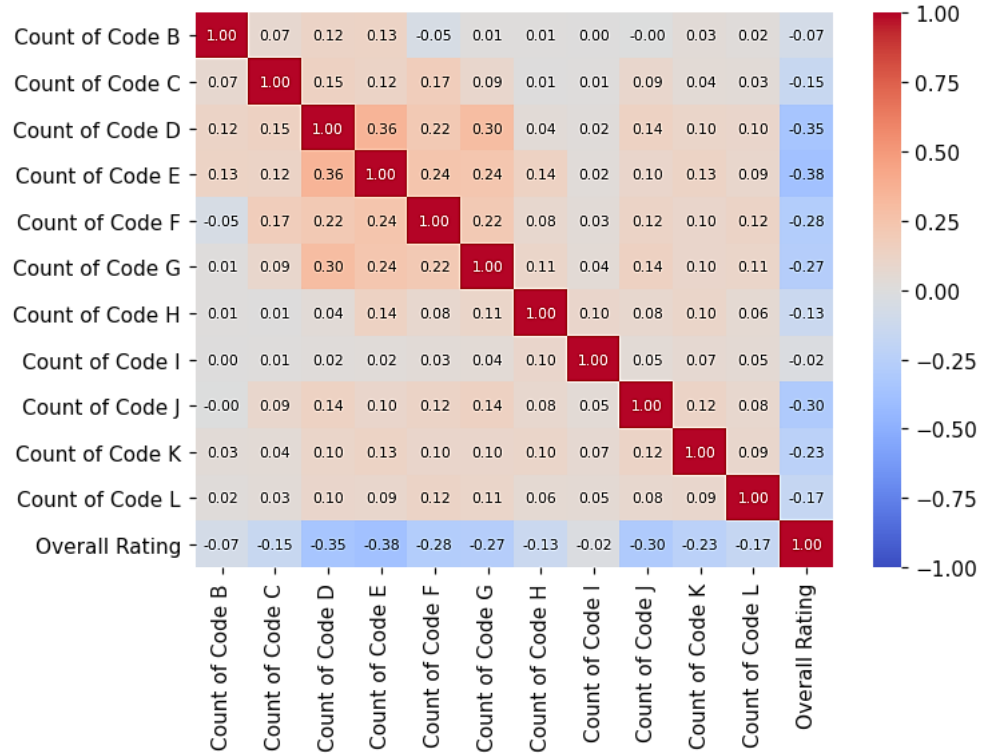
Although there are these outliers in the data, we are not concerned about them because when a "sample size is large enough, you are bound to obtain unusual values…but they are a normal part of the data distribution" [7]. Given that the sample size in the ML dataset is 11,585 nursing homes, we decided not to remove the outliers because this is natural variation in the data points.

Furthermore, with the eleven scope severity codes that nursing homes can receive, we examined which codes were more prevalent among the facilities. Based on the average counts for the 11,585 facilities, we noticed that the codes D, E, F, and G occurred most frequently compared to the other codes.

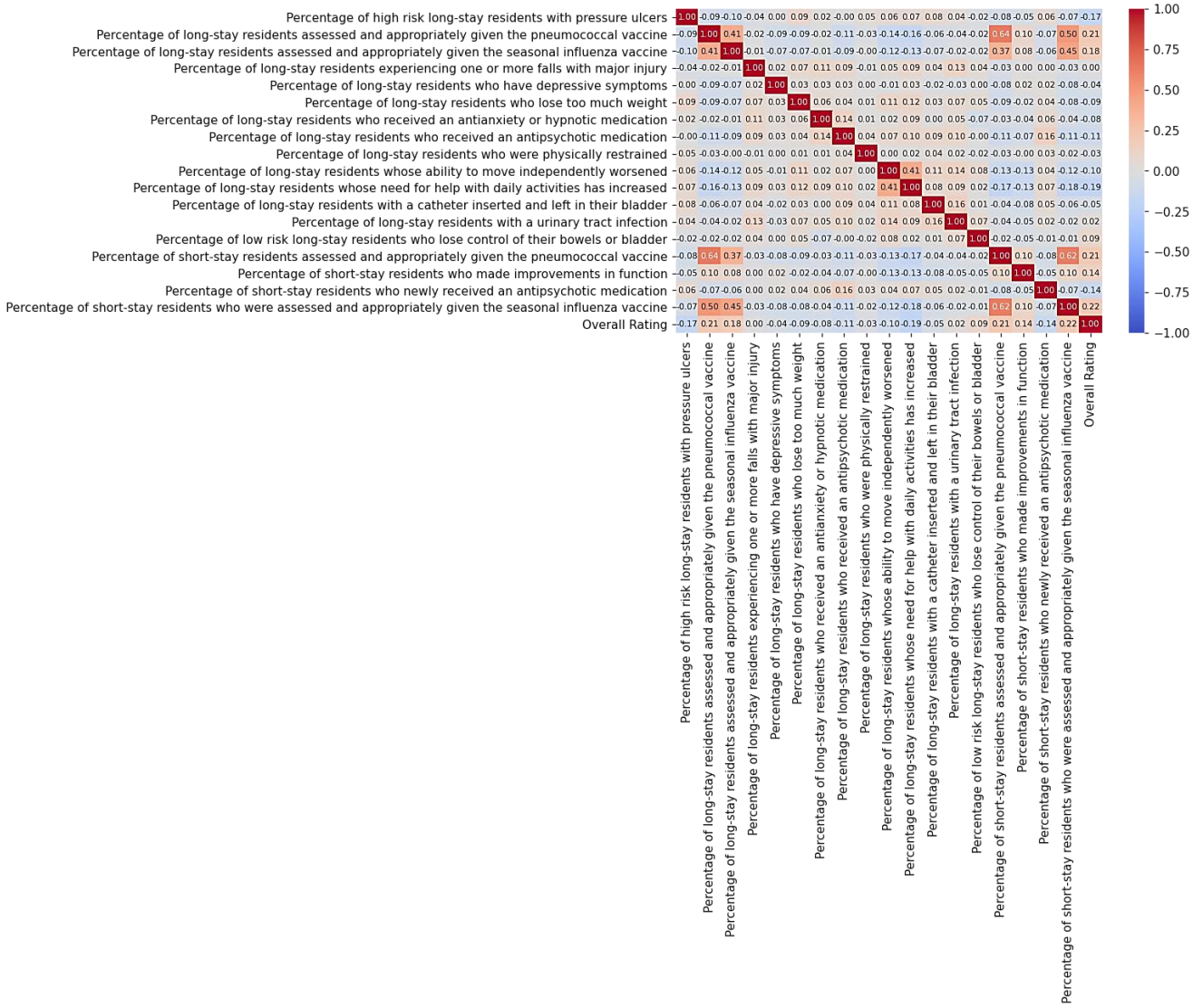


	Code B	Code D	Code E	Code F	Code G	Code J
Average	0.415969	16.135175	6.013897	1.588865	0.746828	0.266120

Fig. 10 Averages for a subset of six scope severity codes.



Then, we inspected the above correlation heatmap containing Kendall's tau correlation coefficients on the severity code features and the overall rating of a nursing home. It was interesting to observe how Scope Severity Code J had a correlation coefficient of -0.30 with overall rating, while this code had a low average count at 0.26612. This coefficient value is close to the correlation coefficients for codes D, E, F, and G, which all had higher average counts. The relationship strength between code J and overall rating is promising given that code B had a higher average (of 0.415969) compared to code J, but code B had the second lowest correlation coefficient of -0.07, which signifies a negligible relationship between it and the overall rating.



From this other correlation heatmap (above), we discovered interesting interactions between data attributes, such as how overall rating is more positively correlated with the short-stay and long-stay quality measures (QMs) relating to pneumococcal and influenza vaccines compared to several of the other non-vaccine related QMs. Additionally, two of the most notable QMs that were found to be negatively correlated with overall rating include the "Percentage of high risk long-stay residents with pressure ulcers" and the "Percentage of long-stay residents whose need for help with daily activities has

increased." Overall, from the correlation heatmaps we created during the exploratory data analysis, the correlation coefficients of +/- 0.21 or more between features and overall rating were promising to observe before we dealt with any ML models. The remaining three correlation heatmaps are in the Appendix.

The Methodology

For partitioning the ML dataset into train/test sets, we decided on a 70:30 split, where 70% of the data is for training the machine learning models, while the remaining 30% is for testing. In the training set, there are 8109 nursing home instances. Then, in the test set, there are 3476 nursing home instances. It is important to note that this train-test split is stratified and thus preserves the distribution of the class labels of the original dataset (before splitting).

Overall Rating	Dataset before splitting	Training set	Test set
1	1951	1366	585
2	2421	1694	727
3	2283	1598	685
4	2520	1764	756
5	2410	1687	723

Fig. 11 Number of instances per label, in each dataset (before and after splitting).

Models

For our investigation into nursing home overall ratings, we considered four machine learning algorithms: k-Nearest Neighbors (kNN), Random Forest, Multi-class Logistic Regression, and AdaBoost. We believed it was suitable to use default hyperparameter values for the classifiers. Additionally, to easily interpret the performance of the four machine learning models, we created a DummyClassifier as our baseline model. Since the DummyClassifier simply predicts the majority class for all instances and thus does not use any data features to do so, this baseline would help us see if the kNN, Random Forest, Multi-class Logistic Regression, and AdaBoost classifiers truly learned from the set of features.

Model Selection

To find the best-performing model when evaluating the algorithms on our training set during the model selection phase, we decided on 10-fold stratified cross-validation technique. After selecting this best model, it would then be evaluated on the unseen test data.

Machine Learning Pipelines

Before applying the four chosen machine learning algorithms to the training set, several data transformations (which are chained together) needed to be made via Transformers within the ML pipelines. For rule-based algorithms having trees as the base learner, such as Random Forest and AdaBoost, these do not require feature scaling in the form of standardization. On the contrary, it is necessary to standardize the numerical features for distance-based algorithms such as Logistic Regression and k-Nearest Neighbors. Although, since a later preprocessing step in the pipeline involved handling missing data values via nearest neighbor imputation by Scikit-learn's KNNImputer, where kNN is still distance-based, we had to perform feature scaling first regardless of the classifiers we chose. The standardization of the data values was completed with StandardScaler. The KNNImputer will use a k-

Nearest Neighbors algorithm to fill in the missing values with averages from the instances' nearest neighbors. Given the 50+ features in our ML dataset, feature selection was the last data preprocessing step in the pipeline. Feature selection helps reduce the feature space to make it more manageable so that the model training phase is not slowed down while still providing reliable results from cross-validation. SelectKBest, a type of univariate feature selection in Scikit-learn, was used to grab the k best features, and we carried out hyperparameter tuning to find the optimal k-value out of four possible values (k = 10, 15, 20, or 25 features). Hence, this was a one-dimensional search space for GridSearchCV. Most importantly, one must keep in mind that the specific set of k features that get chosen is "independent of the model that you might want to apply after the feature selection" [6].

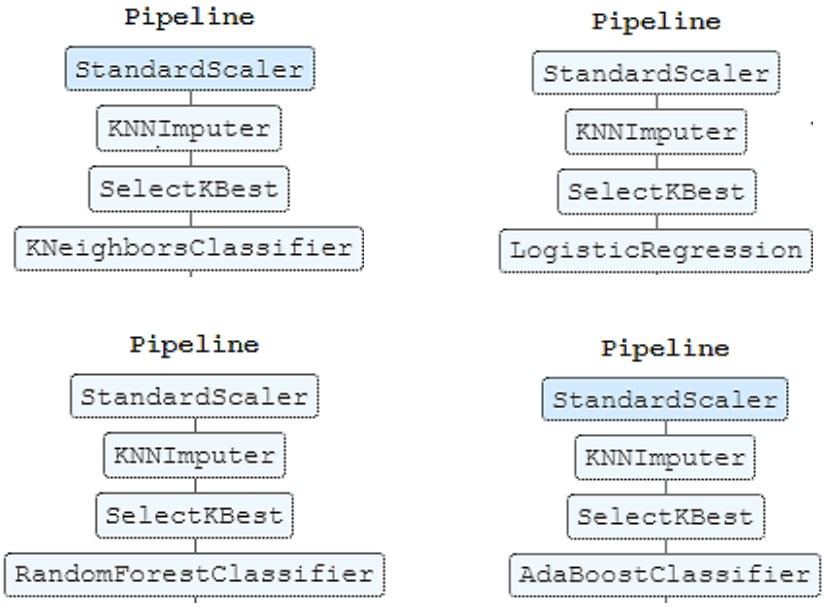


Fig. 12 The four ML pipelines.

The Results

Results Interpretation

The classification accuracies on the *training set* for the five classifiers are shown below.

	1. **Dummy Classifier**	2. **k-Nearest Neighbors**	3. **Multi-class Logistic Regression**	4. **Random Forest**	5. **AdaBoost**
Accuracy	21.75%	42.35%	47.64%	47.40%	45.65%

For the Dummy classifier, we expected its accuracy value of 21.75% because the majority class, which is overall rating 4, contains 1764 instances out of 8109. The two best-performing models for the given data were Logistic Regression and Random Forest, with the former having a slightly better performance. Regarding our main models 2-5, their classification accuracies are better than the baseline. Thus, it shows that the ML models learned some useful patterns from the features. Since Logistic Regression and Random Forest had fairly similar accuracies, we decided that it would be beneficial to run both of the respective models on the test set in order to identify whether or not one model considerably outperforms the other. During hyperparameter tuning, k = 25 (rather than 10, 15, or 20) for SelectKBest provided the

highest performances for the models kNN, Multi-class Logistic Regression, Random Forest, and AdaBoost.

Count of Quality of Life and Care Deficiencies	Count of Resident Assessment and Care Planning Deficiencies	Count of Nursing and Physician Services Deficiencies	Count of Resident Rights Deficiencies	Count of Nutrition and Dietary Deficiencies
Count of Pharmacy Service Deficiencies	Count of Administration Deficiencies	Count of Infection Control Deficiencies	Count of Code D	Count of Code E
Count of Code F	Count of Code G	Count of Code J	Reported Nurse Aide Staffing Hours per Resident per Day	Reported RN Staffing Hours per Resident per Day
Total number of nurse staff hours per resident per day on the weekend	Registered Nurse hours per resident per day on the weekend	Total nursing staff turnover	Registered Nurse turnover	Total Health Revisits
Number of Substantiated Complaints	Number of Citations from Infection Control Inspections	Percentage of long-stay residents whose need for help with daily activities has increased	Percentage of short-stay residents assessed and appropriately given the pneumococcal vaccine	Percentage of short-stay residents who were assessed and appropriately given the seasonal influenza vaccine

Fig. 13 The 25 selected features from the SelectKBest method.

From these 25 features, 16 of them are related to health inspections, 6 of them are related to staffing, and 3 of them are related to quality measures, as indicated in the table. It is important to note that the two vaccine-related quality measures (QMs) are not utilized to calculate the QM rating for a nursing home, but SelectKBest's scoring function f_classif chose them out of all other possible QMs.

Feature Importances

When considering feature importances of various models, "the feature importances of the input data depend on the corresponding classification model and that a feature important for one model may be unimportant for another model" [10]. Hence, we will not be making any pairwise comparisons for the feature importances among the models. Instead, we examine and discuss the top-10 most important features and bottom-5 least important features for each model individually.

By reviewing the feature importances of the 25 selected features, we detected which features had the most impact on a model's predictions for the overall ratings of nursing homes. From the top-10 kNN feature importances, we noticed that a majority of the features were related to the health inspections rating. For the top-10 Logistic Regression feature importances, we saw that a majority of the features were related to the health inspections rating. Based on the top-10 feature importances for Random Forest, we saw that staffing-related features were the most important in predicting the overall rating. From the top-10 AdaBoost feature importances, we noticed that the features determined to have the most impact were a mix of health inspection-related features and staffing-related features.

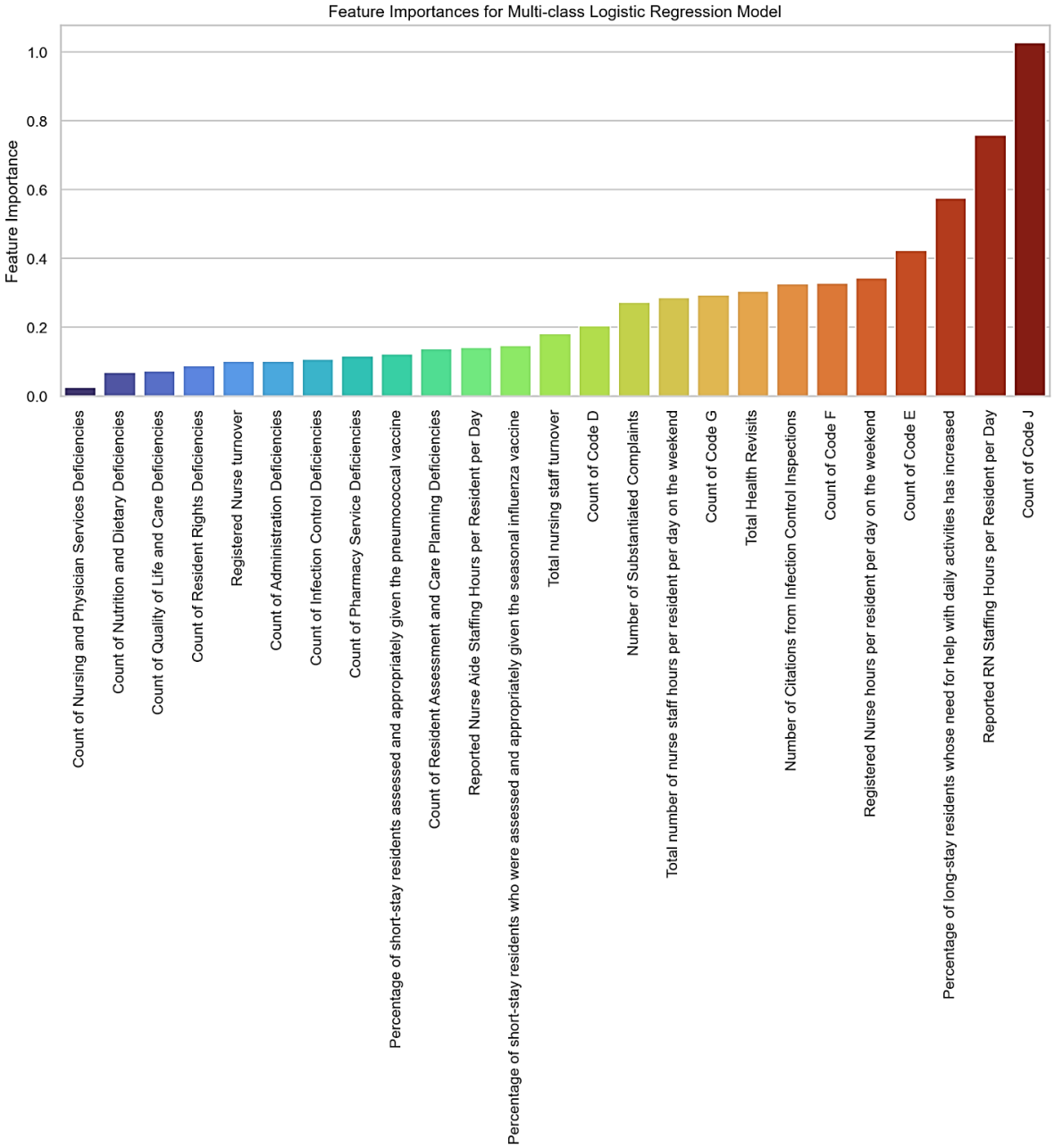
On another note, when reviewing the bottom-5 feature importances for all four models, we observed a pattern where four of the features (or all five in the case of the Random Forest model) were relevant to

health inspections. To add, each of the models ranked the feature "Count of Nursing and Physician Services Deficiencies" with low importance for the class label predictions.

The plots and tables for the feature importances of our two best-performing models are displayed below. The remaining plots and tables (for kNN and AdaBoost) are located in the Appendix.

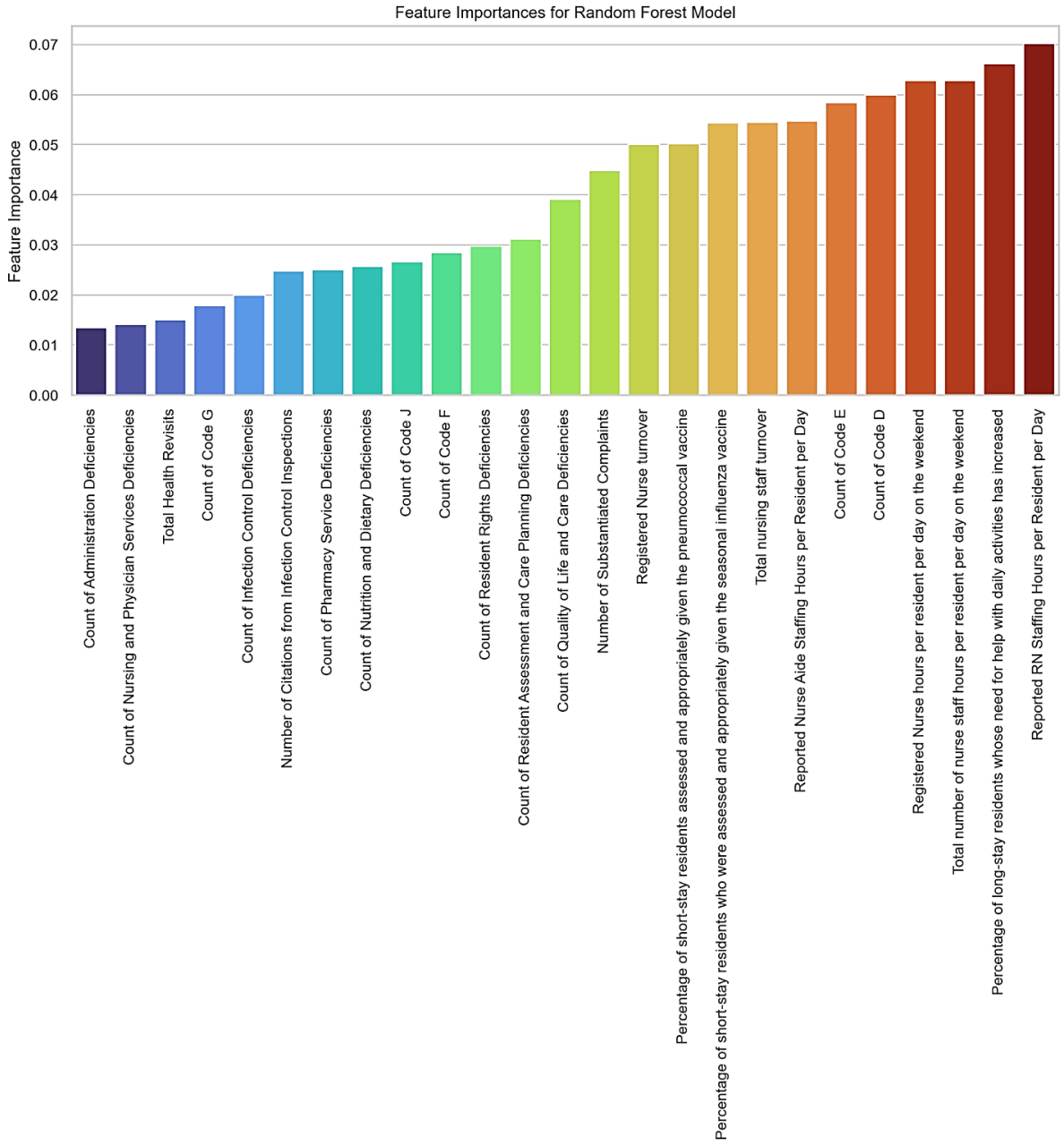
Name	**Feature Importance** (sorted from greatest to least)
Count of Code J	1.027342
Reported RN Staffing Hours per Resident per Day	0.758477
Percentage of long-stay residents whose need for help with daily activities has increased	0.576447
Count of Code E	0.423583
Registered Nurse hours per resident per day on the weekend	0.343422
Count of Code F	0.328882
Number of Citations from Infection Control Inspections	0.325980
Total Health Revisits	0.305161
Count of Code G	0.294118
Total number of nurse staff hours per resident per day on the weekend	0.285727
Number of Substantiated Complaints	0.271958
Count of Code D	0.203602
Total nursing staff turnover	0.181098
Percentage of short-stay residents who were assessed and appropriately given the seasonal influenza vaccine	0.147536
Reported Nurse Aide Staffing Hours per Resident per Day	0.141704
Count of Resident Assessment and Care Planning Deficiencies	0.137960
Percentage of short-stay residents assessed and appropriately given the pneumococcal vaccine	0.122240
Count of Pharmacy Service Deficiencies	0.115843
Count of Infection Control Deficiencies	0.106381
Count of Administration Deficiencies	0.101363
Registered Nurse turnover	0.100814
Count of Resident Rights Deficiencies	0.087395
Count of Quality of Life and Care Deficiencies	0.073351
Count of Nutrition and Dietary Deficiencies	0.068503
Count of Nursing and Physician Services Deficiencies	0.025448

Fig. 14 Feature importances for Multi-class Logistic Regression model.


Feature Importances for Multi-class Logistic Regression Model
Feature Importance
0.0
0.2
0.4
0.6
0.8
1.0
Count of Nursing and Physician Services Deficiencies
Count of Nutrition and Dietary Deficiencies
Count of Quality of Life and Care Deficiencies
Count of Resident Rights Deficiencies
Registered Nurse turnover
Count of Administration Deficiencies
Count of Infection Control Deficiencies
Count of Pharmacy Service Deficiencies
Percentage of short-stay residents assessed and appropriately given the pneumococcal vaccine
Count of Resident Assessment and Care Planning Deficiencies
Reported Nurse Aide Staffing Hours per Resident per Day
Percentage of short-stay residents who were assessed and appropriately given the seasonal influenza vaccine
Total nursing staff turnover
Count of Code D
Number of Substantiated Complaints
Total number of nurse staff hours per resident per day on the weekend
Count of Code G
Total Health Revisits
Number of Citations from Infection Control Inspections
Count of Code F
Registered Nurse hours per resident per day on the weekend
Count of Code E
Percentage of long-stay residents whose need for help with daily activities has increased
Reported RN Staffing Hours per Resident per Day
Count of Code J

Name	**Feature Importance** (sorted from greatest to least)
Reported RN Staffing Hours per Resident per Day	0.070260
Percentage of long-stay residents whose need for help with daily activities has increased	0.066263
Total number of nurse staff hours per resident per day on the weekend	0.062828
Registered Nurse hours per resident per day on the weekend	0.062778
Count of Code D	0.059985
Count of Code E	0.058369
Reported Nurse Aide Staffing Hours per Resident per Day	0.054702
Total nursing staff turnover	0.054445
Percentage of short-stay residents who were assessed and appropriately given the seasonal influenza vaccine	0.054409
Percentage of short-stay residents assessed and appropriately given the pneumococcal vaccine	0.050133
Registered Nurse turnover	0.050011
Number of Substantiated Complaints	0.044866
Count of Quality of Life and Care Deficiencies	0.039068
Count of Resident Assessment and Care Planning Deficiencies	0.031177
Count of Resident Rights Deficiencies	0.029724
Count of Code F	0.028387
Count of Code J	0.026661
Count of Nutrition and Dietary Deficiencies	0.025726
Count of Pharmacy Service Deficiencies	0.025012
Number of Citations from Infection Control Inspections	0.024835
Count of Infection Control Deficiencies	0.019903
Count of Code G	0.017922
Total Health Revisits	0.014994
Count of Nursing and Physician Services Deficiencies	0.014070
Count of Administration Deficiencies	0.013472

Fig. 15 Feature importances for Random Forest model.



Evaluation on Test Set

We evaluated the Multi-class Logistic Regression model and the Random Forest model on our test set containing 3476 nursing home instances, and we obtained their respective classification accuracies. The Multi-class Logistic Regression model had 47.01% accuracy for the predictions on the test set, and the Random Forest model had 46.58% accuracy. Since these classification accuracies are not drastically different from the model performances obtained on the training set, this signifies that there was no extreme overfitting to the data.

To better understand which class labels (1, 2, 3, 4, 5) the two models were performing poorly on and what the misclassifications were, we inspected their confusion matrix plots. From the plots, we found that

most of the misclassified nursing homes with a true overall rating of 2, 3, or 4 often had (incorrectly) predicted labels of +/- 1 for the rating. Additionally, the misclassified nursing homes with a true overall rating of 1 or 5 most often had predicted labels of 2 or 4, respectively.

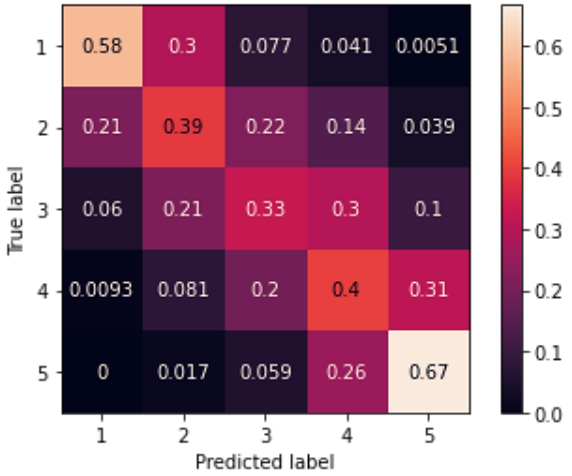


Fig. 16 Confusion matrix plot for Multi-class Logistic Regression.

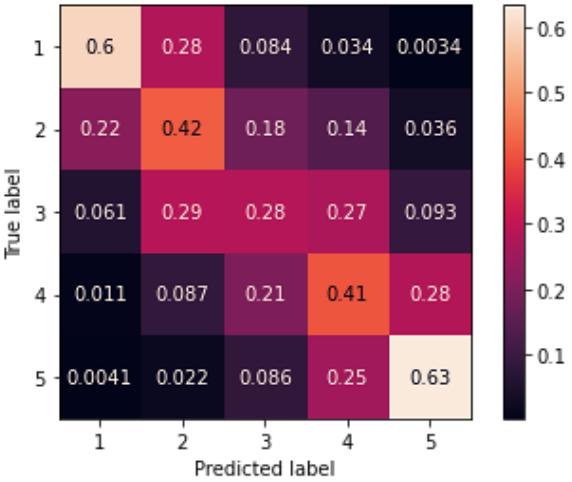


Fig. 17 Confusion matrix plot for Random Forest.

Conclusions and Future Work

To conclude, the high-level takeaways from our investigation was that Multi-class Logistic regression (a distance-based algorithm) performed slightly better than Random Forest (a rule-based algorithm) when predicting the overall ratings of nursing homes in the test data. The findings from the two confusion matrices indicate that the respective decision boundaries (used by the models) are not clearly defined, which leads to more misclassifications. Although, when considering the highest-performing Logistic Regression model and its number of correct predictions, the model found patterns in the data, even when the calculated ratings from the three domains (health inspections, staffing, QM) were excluded. On another note, we discovered two QMs (the vaccine-related ones) that are not officially used in calculating the QM rating, but they both

had a significant statistical relationship with the target variable according to SelectKBest. This indicates that it would be beneficial to have the flu and pneumonia prevention measures also accounted for in the quality measures rating, which is a component of the overall star rating.

When provided a particular dataset, knowing which combinations of hyperparameter values to specify for searching requires more in-depth research and expertise about the individual ML algorithms. Hence, due to the time constraints, only default parameters were used for the four chosen classifiers, but hyperparameter optimizations can help in increasing model performance. If we had more time available during this investigation of nursing home ratings, some improvements that could be made involve doing further exploration of the different hyperparameter tuning options available for the machine learning models. This hyperparameter tuning can be performed with GridSearchCV such that the parameter grid will now also contain specific parameter values per model. Furthermore, more values for k in SelectKBest can be tried during hyperparameter tuning to see if model performance improves or not. The reasoning behind doing so is due to Hughes Phenomenon. Hughes Phenomenon states that as the number of features increases, a classifier's performance will continue increasing until there is a certain k number of features where for any value greater than k, the classifier's performance begins to worsen. So, within our parameter grid for GridSearchCV, it would be worthwhile to test k = 26, 27, 28, 29, 30, 31, 32, 33, 34, 35 features.

References

[1] Brownlee, J. (2020). How to Calculate Feature Importance With Python. https://machinelearningmastery.com/calculate-feature-importance-with-python/

[2] Brownlee, J. (2019). Your First Machine Learning Project in Python Step-By-Step in Python Machine Learning. https://machinelearningmastery.com/machine-learning-in-python-step-by-step/

[3] Beuzen, T. (2020). Simultaneous feature preprocessing, feature selection, model selection, and hyperparameter tuning in scikit-learn with Pipeline and GridSearchCV. https://www.tomasbeuzen.com/post/scikit-learn-gridsearch-pipelines/

[4] CMS Releases Latest List of Special Focus Facilities and Candidates. (2022). https://theconsumervoice.org/news/detail/latest/updated-sff-list-january-2022

[5] Design for Care Compare Nursing Home Five-Star Quality Rating System: Technical Users' Guide. (2022). *The Centers for Medicare & Medicaid Services*. https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/CertificationandComplianc/downloads/usersguide.pdf

[6] Froehlich, J. E. (n.d.). Feature Selection and Hyperparameter Tuning. https://makeabilitylab.github.io/physcomp/signals/FeatureSelectionAndHyperparameterTuning/index.html

[7] Frost, J. (n.d.). Guidelines for Removing and Handling Outliers in Data. https://statisticsbyjim.com/basics/remove-outliers/

[8] Minimum Data Set (MDS) Data. (n.d.). *UCSF Pepper Center*. https://peppercenter.ucsf.edu/minimum-data-set-mds-data

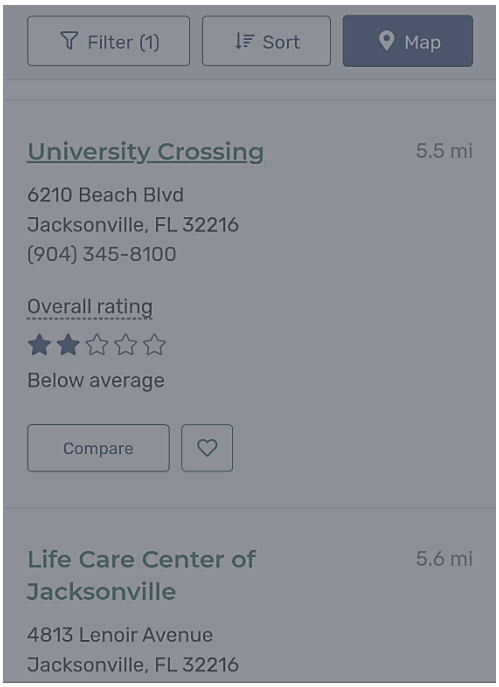
[9] Quality measures. (n.d.). Nursing homes including rehab services, *The Centers for Medicare & Medicaid Services*. https://data.cms.gov/provider-data/topics/nursing-homes/quality-of-resident-care

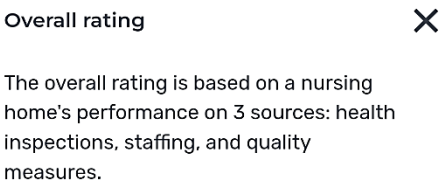
[10] Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3(272). https://doi.org/10.1007/s42452-021-04148-9

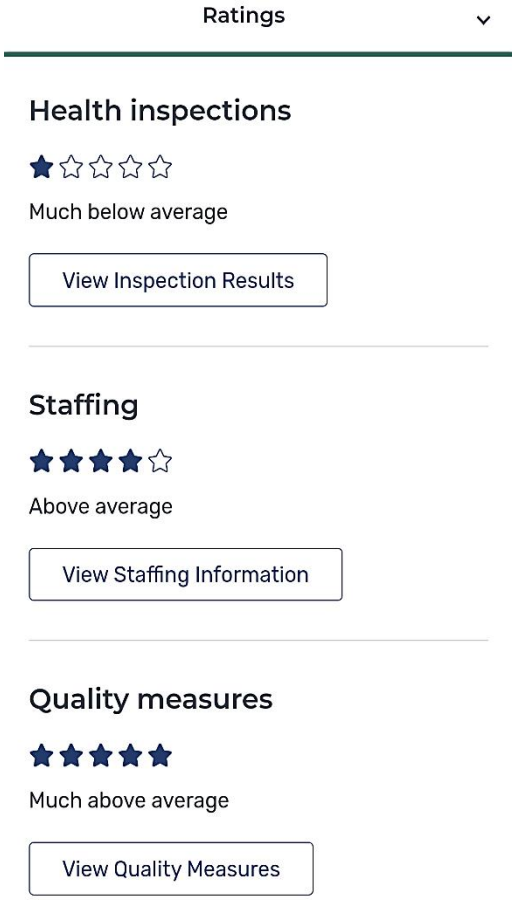
[11] Technical details. (n.d.). Nursing homes including rehab services, *The Centers for Medicare & Medicaid Services*. https://data.cms.gov/provider-data/topics/nursing-homes/technical-details

Appendix

Nursing Home Care Compare website







Feature Selection for ***Provider Information*** dataset

* All columns below are numeric data types.

1. Reported Nurse Aide Staffing Hours per Resident per Day
2. Reported LPN Staffing Hours per Resident per Day
3. Reported RN Staffing Hours per Resident per Day
4. Total number of nurse staff hours per resident per day on the weekend
5. Registered Nurse hours per resident per day on the weekend
6. Reported Physical Therapist Staffing Hours per Resident Per Day
7. Total nursing staff turnover
8. Registered Nurse turnover
9. Number of administrators who have left the nursing home
10. Rating Cycle 1 Number of Health Revisits
11. Rating Cycle 2 Number of Health Revisits
12. Rating Cycle 3 Number of Health Revisits
13. Number of Facility Reported Incidents
14. Number of Substantiated Complaints
15. Number of Citations from Infection Control Inspections

Feature Selection for ***Survey Summary*** dataset
* All columns below are numeric data types.

1. Count of Freedom from Abuse and Neglect and Exploitation Deficiencies
2. Count of Quality of Life and Care Deficiencies
3. Count of Resident Assessment and Care Planning Deficiencies
4. Count of Nursing and Physician Services Deficiencies
5. Count of Resident Rights Deficiencies.
6. Count of Nutrition and Dietary Deficiencies
7. Count of Pharmacy Service Deficiencies
8. Count of Environmental Deficiencies
9. Count of Administration Deficiencies
10. Count of Infection Control Deficiencies

MDS-based measures within the ***MDS Quality Measures*** dataset:
***Orange coloring** denotes that this measure is used in the QM rating calculation.

1. Percentage of long-stay residents whose need for help with daily activities has increased
2. Percentage of long-stay residents who lose too much weight
3. Percentage of low risk long-stay residents who lose control of their bowels or bladder
4. Percentage of long-stay residents with a catheter inserted and left in their bladder
5. Percentage of long-stay residents with a urinary tract infection
6. Percentage of long-stay residents who have depressive symptoms
7. Percentage of long-stay residents who were physically restrained
8. Percentage of long-stay residents experiencing one or more falls with major injury
9. Percentage of long-stay residents assessed and appropriately given the pneumococcal vaccine
10. Percentage of long-stay residents who received an antipsychotic medication
11. Percentage of short-stay residents assessed and appropriately given the pneumococcal vaccine
12. Percentage of short-stay residents who newly received an antipsychotic medication
13. Percentage of long-stay residents whose ability to move independently worsened
14. Percentage of long-stay residents who received an antianxiety or hypnotic medication
15. Percentage of high risk long-stay residents with pressure ulcers
16. Percentage of long-stay residents assessed and appropriately given the seasonal influenza vaccine
17. Percentage of short-stay residents who made improvements in function
18. Percentage of short-stay residents who were assessed and appropriately given the seasonal influenza vaccine

Data Quality details in (updated) *Provider Information*

Feature	Number of missing values	Total percentage that is missing (out of 15002 NH instances)
Reported Nurse Aide Staffing Hours per Resident per Day	439	2.926%
Reported LPN Staffing Hours per Resident per Day	439	2.926%
Reported RN Staffing Hours per Resident per Day	439	2.926%
Total number of nurse staff hours per resident per day on the weekend	440	2.933%
Registered Nurse hours per resident per day on the weekend	440	2.933%
Reported Physical Therapist Staffing Hours per Resident Per Day	439	2.926%
Total nursing staff turnover	2382	**15.878%**
Registered Nurse turnover	3193	**21.284%**
Number of administrators who have left the nursing home	3096	**20.637%**
Rating Cycle 1 Number of Health Revisits	0	0%
Rating Cycle 2 Number of Health Revisits	0	0%
Rating Cycle 3 Number of Health Revisits	0	0%
Number of Facility Reported Incidents	0	0%
Number of Substantiated Complaints	0	0%
Number of Citations from Infection Control Inspections	12	0.08%

The meaning behind the missing values in ***Provider Information***:
*According to the footnotes in NH_Primary_Data_Dictionary.xlsx

Feature name	Footnote description
Reported Nurse Aide Staffing Hours per Resident per Day	This facility did not submit staffing data, or submitted data that did not meet the criteria required to calculate a staffing measure.
Reported LPN Staffing Hours per Resident per Day	This facility did not submit staffing data, or submitted data that did not meet the criteria required to calculate a staffing measure.
Reported RN Staffing Hours per Resident per Day	This facility did not submit staffing data, or submitted data that did not meet the criteria required to calculate a staffing measure.
Total number of nurse staff hours per resident per day on the weekend	This facility did not submit staffing data, or submitted data that did not meet the criteria required to calculate a staffing measure.
Registered Nurse hours per resident per day on the weekend	This facility did not submit staffing data, or submitted data that did not meet the criteria required to calculate a staffing measure.
Reported Physical Therapist Staffing Hours per Resident Per Day	This facility did not submit staffing data, or submitted data that did not meet the criteria required to calculate a staffing measure.

Total nursing staff turnover	This facility did not submit staffing data, or submitted data that did not meet the criteria required to calculate a staffing measure.
Registered Nurse turnover	This facility did not submit staffing data, or submitted data that did not meet the criteria required to calculate a staffing measure.
Number of administrators who have left the nursing home	This facility did not submit staffing data, or submitted data that did not meet the criteria required to calculate a staffing measure.
Number of Citations from Infection Control Inspections	N/A

All features in the **ML dataset** and their corresponding domain:

Health Inspections, Quality Measures, Staffing

Count of Freedom from Abuse and Neglect and Exploitation Deficiencies	Count of Quality of Life and Care Deficiencies	Count of Resident Assessment and Care Planning Deficiencies	Count of Nursing and Physician Services Deficiencies
Count of Resident Rights Deficiencies	Count of Nutrition and Dietary Deficiencies	Count of Pharmacy Service Deficiencies	Count of Environmental Deficiencies
Count of Administration Deficiencies	Count of Infection Control Deficiencies	Count of Code B	Count of Code C
Count of Code D	Count of Code E	Count of Code F	Count of Code G
Count of Code H	Count of Code I	Count of Code J	Count of Code K
Count of Code L	Reported Nurse Aide Staffing Hours per Resident per Day	Reported LPN Staffing Hours per Resident per Day	Reported RN Staffing Hours per Resident per Day
Total number of nurse staff hours per resident per day on the weekend	Registered Nurse hours per resident per day on the weekend	Reported Physical Therapist Staffing Hours per Resident Per Day	Total nursing staff turnover
Registered Nurse turnover	Number of administrators who have left the nursing home	Total Health Revisits	Number of Facility Reported Incidents
Number of Substantiated Complaints	Number of Citations from Infection Control Inspections	Percentage of high risk long-stay residents with pressure ulcers	Percentage of long-stay residents assessed and appropriately given the pneumococcal vaccine
Percentage of long-stay residents assessed and appropriately given the seasonal influenza vaccine	Percentage of long-stay residents experiencing one or more falls with major injury	Percentage of long-stay residents who have depressive symptoms	Percentage of long-stay residents who lose too much weight
Percentage of long-stay residents who received an antianxiety or hypnotic medication	Percentage of long-stay residents who received an antipsychotic medication	Percentage of long-stay residents who were physically restrained	Percentage of long-stay residents whose ability to move independently worsened
Percentage of long-stay residents whose need for help with daily activities has increased	Percentage of long-stay residents with a catheter inserted and left in their bladder	Percentage of long-stay residents with a urinary tract infection	Percentage of low risk long-stay residents who lose control of their bowels or bladder

Percentage of short-stay residents assessed and appropriately given the pneumococcal vaccine	Percentage of short-stay residents who made improvements in function	Percentage of short-stay residents who newly received an antipsychotic medication	Percentage of short-stay residents who were assessed and appropriately given the seasonal influenza vaccine
Number of hospitalizations per 1000 long-stay resident days	Number of outpatient emergency department visits per 1000 long-stay resident days	Percentage of short-stay residents who had an outpatient emergency department visit	Percentage of short-stay residents who were rehospitalized after a nursing home admission

Descriptive statistics for the 56 features in the **ML dataset**:

	count	mean	std	min	25%	50%	75%	max
Count of Freedom from Abuse and Neglect and Exploitation Deficiencies	11585.0	0.808373	1.312347	0.000000	0.000000	0.000000	1.000000	13.000000
Count of Quality of Life and Care Deficiencies	11585.0	4.931204	4.200128	0.000000	2.000000	4.000000	7.000000	36.000000
Count of Resident Assessment and Care Planning Deficiencies	11585.0	3.171342	2.792905	0.000000	1.000000	3.000000	5.000000	20.000000
Count of Nursing and Physician Services Deficiencies	11585.0	0.595339	1.001335	0.000000	0.000000	0.000000	1.000000	9.000000
Count of Resident Rights Deficiencies	11585.0	3.068968	2.930242	0.000000	1.000000	2.000000	4.000000	24.000000
Count of Nutrition and Dietary Deficiencies	11585.0	1.862408	1.841451	0.000000	1.000000	1.000000	3.000000	15.000000
Count of Pharmacy Service Deficiencies	11585.0	2.526716	2.259915	0.000000	1.000000	2.000000	4.000000	17.000000
Count of Environmental Deficiencies	11585.0	0.868019	1.306764	0.000000	0.000000	0.000000	1.000000	12.000000
Count of Administration Deficiencies	11585.0	0.543202	0.998645	0.000000	0.000000	0.000000	1.000000	11.000000

	count	mean	std	min	25%	50%	75%	max
Count of Infection Control Deficiencies	11585.0	1.313940	1.182106	0.000000	0.000000	1.000000	2.000000	8.000000
Count of Code B	11585.0	0.415969	0.998551	0.000000	0.000000	0.000000	0.000000	13.000000
Count of Code C	11585.0	0.350022	0.767164	0.000000	0.000000	0.000000	0.000000	11.000000
Count of Code D	11585.0	16.135175	12.478989	0.000000	7.000000	13.000000	22.000000	96.000000
Count of Code E	11585.0	6.013897	6.491061	0.000000	2.000000	4.000000	8.000000	80.000000
Count of Code F	11585.0	1.588865	2.265348	0.000000	0.000000	1.000000	2.000000	31.000000
Count of Code G	11585.0	0.746828	1.406331	0.000000	0.000000	0.000000	1.000000	18.000000
Count of Code H	11585.0	0.037980	0.301253	0.000000	0.000000	0.000000	0.000000	11.000000
Count of Code I	11585.0	0.002158	0.079355	0.000000	0.000000	0.000000	0.000000	5.000000
Count of Code J	11585.0	0.266120	0.845066	0.000000	0.000000	0.000000	0.000000	20.000000
Count of Code K	11585.0	0.104359	0.501795	0.000000	0.000000	0.000000	0.000000	11.000000
Count of Code L	11585.0	0.051618	0.300370	0.000000	0.000000	0.000000	0.000000	6.000000
Reported Nurse Aide Staffing Hours per Resident per Day	11368.0	2.130226	0.529740	0.000000	1.779530	2.070055	2.445500	6.883690
Reported LPN Staffing Hours per Resident per Day	11368.0	0.870892	0.321046	0.000000	0.662112	0.860965	1.062482	3.793760
Reported RN Staffing Hours per Resident per Day	11368.0	0.636060	0.326672	0.000000	0.413300	0.578110	0.791397	7.106850
Total number of nurse staff hours per resident per day on the weekend	11368.0	3.153788	0.700167	0.003910	2.712862	3.071030	3.548997	8.769690
Registered Nurse hours per resident per day on the weekend	11368.0	0.422200	0.255160	0.000000	0.243655	0.368700	0.544355	5.415000
Reported Physical	11368.0	0.069995	0.057782	0.000000	0.030515	0.058480	0.094493	0.746970

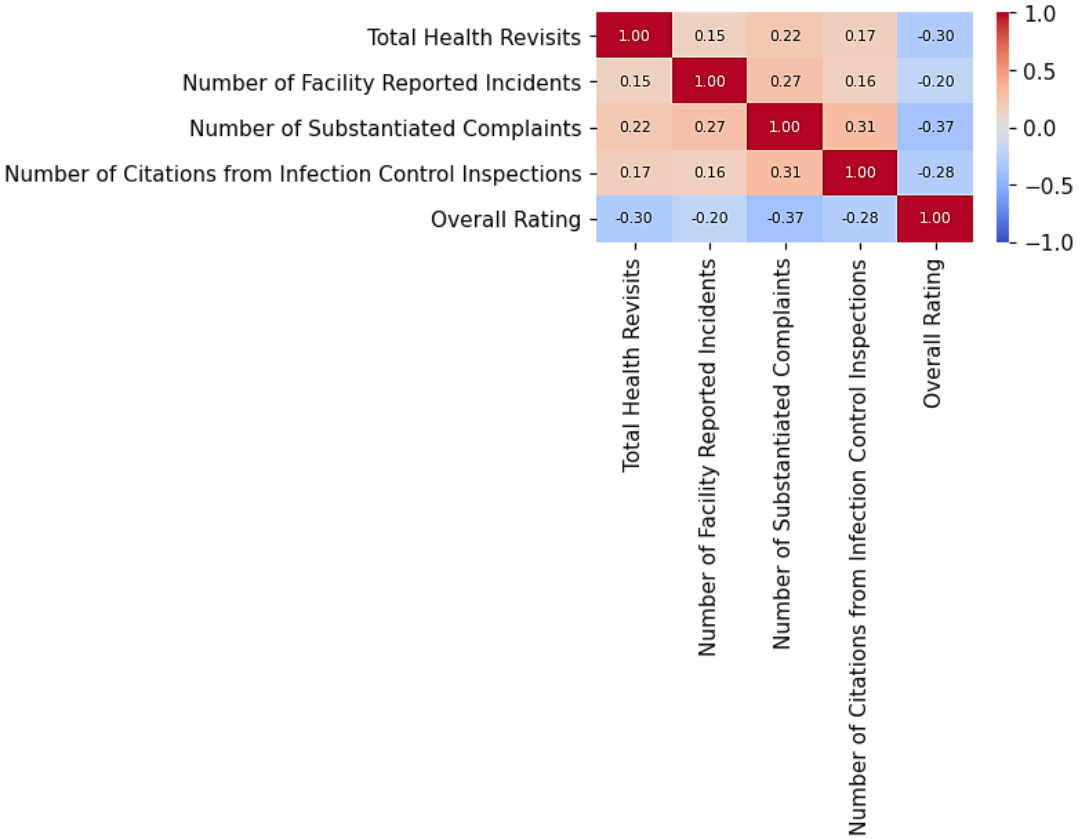
	count	mean	std	min	25%	50%	75%	max
Therapist Staffing Hours per Resident Per Day								
Total nursing staff turnover	10097.0	51.739665	14.824203	6.100000	41.600000	51.200000	61.300000	100.00000
Registered Nurse turnover	9716.0	50.367126	20.786939	0.000000	35.500000	50.000000	64.500000	100.00000
Number of administrators who have left the nursing home	9746.0	1.172071	2.039035	0.000000	0.000000	1.000000	1.000000	31.000000
Total Health Revisits	11585.0	2.831679	0.689163	0.000000	3.000000	3.000000	3.000000	7.000000
Number of Facility Reported Incidents	11585.0	1.543461	3.462885	0.000000	0.000000	0.000000	2.000000	65.000000
Number of Substantiated Complaints	11585.0	5.212171	9.267987	0.000000	0.000000	2.000000	6.000000	174.00000
Number of Citations from Infection Control Inspections	11585.0	1.454467	2.439075	0.000000	0.000000	1.000000	2.000000	43.000000
Percentage of high risk long-stay residents with pressure ulcers	11585.0	8.426922	4.548798	0.000000	5.095544	7.894736	11.038960	36.601308
Percentage of long-stay residents assessed and appropriately given the pneumococcal vaccine	11585.0	93.241861	12.078861	1.126128	92.452830	98.353908	100.00000	100.00000
Percentage of long-stay residents assessed and appropriately given the seasonal influenza vaccine	11585.0	95.757334	6.556284	23.444976	94.736842	98.009950	99.591837	100.00000
Percentage of long-stay residents experiencing	11585.0	3.341688	2.371435	0.000000	1.632651	2.970295	4.605262	21.505379

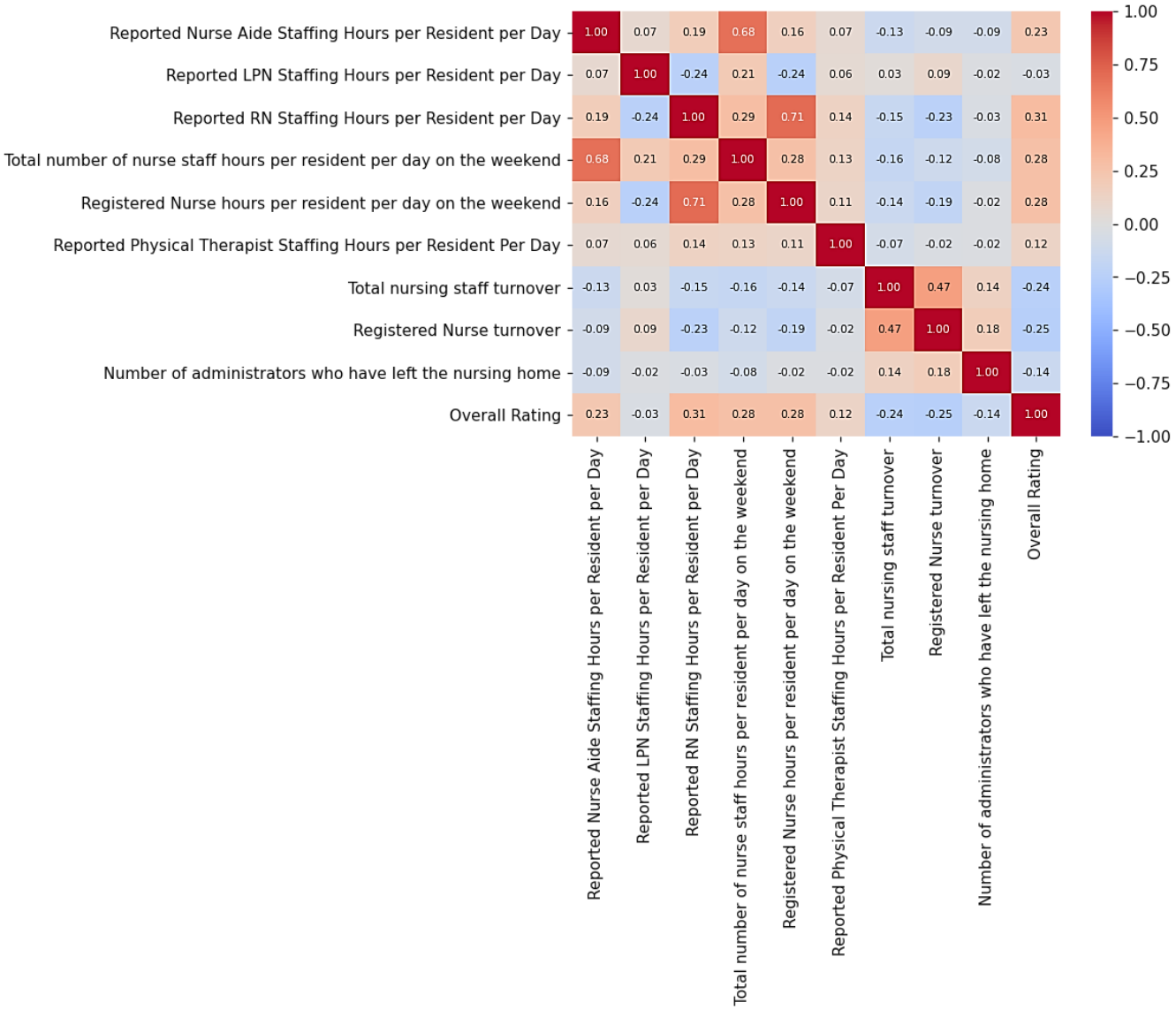
	count	mean	std	min	25%	50%	75%	max
one or more falls with major injury								
Percentage of long-stay residents who have depressive symptoms	11585.0	7.447928	14.228029	0.000000	0.440530	2.380954	7.276996	100.00000
Percentage of long-stay residents who lose too much weight	11585.0	6.863166	3.941857	0.000000	4.089221	6.349208	9.090911	30.952383
Percentage of long-stay residents who received an antianxiety or hypnotic medication	11585.0	18.974704	9.138731	0.000000	12.405064	18.015668	24.500000	74.698795
Percentage of long-stay residents who received an antipsychotic medication	11585.0	13.271343	7.622776	0.000000	7.857145	12.389383	17.730494	59.459462
Percentage of long-stay residents who were physically restrained	11585.0	0.096960	0.642876	0.000000	0.000000	0.000000	0.000000	20.279720
Percentage of long-stay residents whose ability to move independently worsened	11585.0	22.371544	10.224248	0.000000	14.854153	21.906851	28.939174	74.994635
Percentage of long-stay residents whose need for help with daily activities has increased	11585.0	15.434886	6.616084	0.000000	10.833332	15.025909	19.565216	52.606635
Percentage of long-stay residents with a catheter inserted and left in their bladder	11585.0	1.465577	1.596152	0.000000	0.360041	1.018891	2.068995	18.259357
Percentage of long-stay	11585.0	2.269828	2.403816	0.000000	0.591717	1.574802	3.133902	28.368795

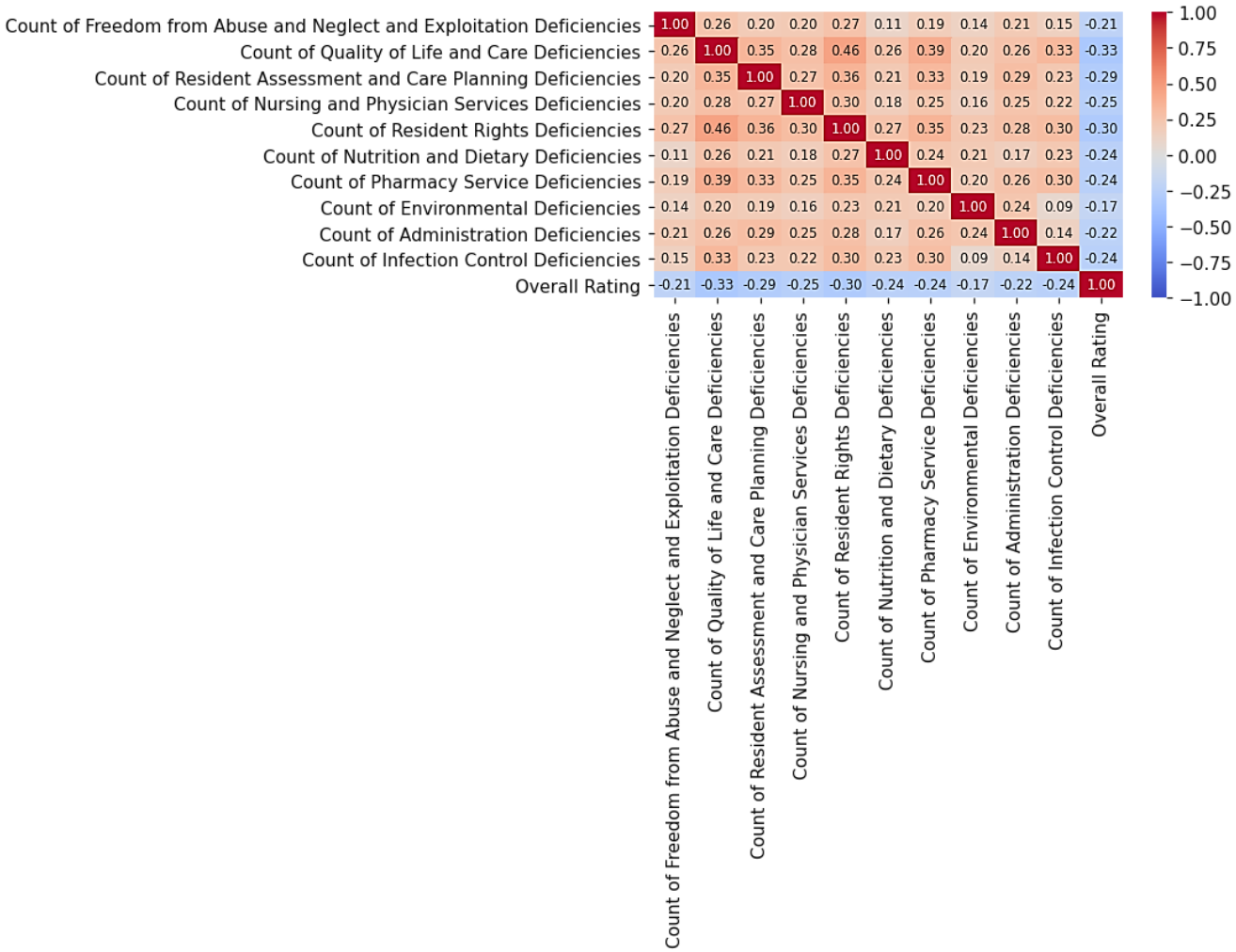
	count	mean	std	min	25%	50%	75%	max
residents with a urinary tract infection								
Percentage of low risk long-stay residents who lose control of their bowels or bladder	11585.0	48.716642	17.266504	0.000000	36.974791	48.837209	60.185184	100.00000
Percentage of short-stay residents assessed and appropriately given the pneumococcal vaccine	11585.0	80.240899	22.008150	1.465201	70.473085	88.727272	96.998798	100.00000
Percentage of short-stay residents who made improvements in function	11585.0	73.378885	12.936190	0.000000	66.504231	74.605587	82.302651	100.00000
Percentage of short-stay residents who newly received an antipsychotic medication	11585.0	1.779542	1.979745	0.000000	0.326798	1.315791	2.469136	22.222221
Percentage of short-stay residents who were assessed and appropriately given the seasonal influenza vaccine	11585.0	79.267221	19.149877	2.816901	70.588235	85.504886	93.617021	100.00000
Number of hospitalizations per 1000 long-stay resident days	9890.0	1.449809	0.648578	0.000000	0.985102	1.379640	1.836078	5.127923
Number of outpatient emergency department visits per 1000 long-stay resident days	9890.0	0.762804	0.504871	0.000000	0.403179	0.652814	1.001233	3.827379

	count	mean	std	min	25%	50%	75%	max
Percentage of short-stay residents who had an outpatient emergency department visit	9890.0	10.496168	5.338787	0.000000	6.780362	9.868402	13.470649	41.813369
Percentage of short-stay residents who were rehospitalized after a nursing home admission	9890.0	22.664678	6.726358	0.000000	18.292198	22.731294	27.103161	51.402290

Exploratory Data Analysis – Correlation heatmaps:



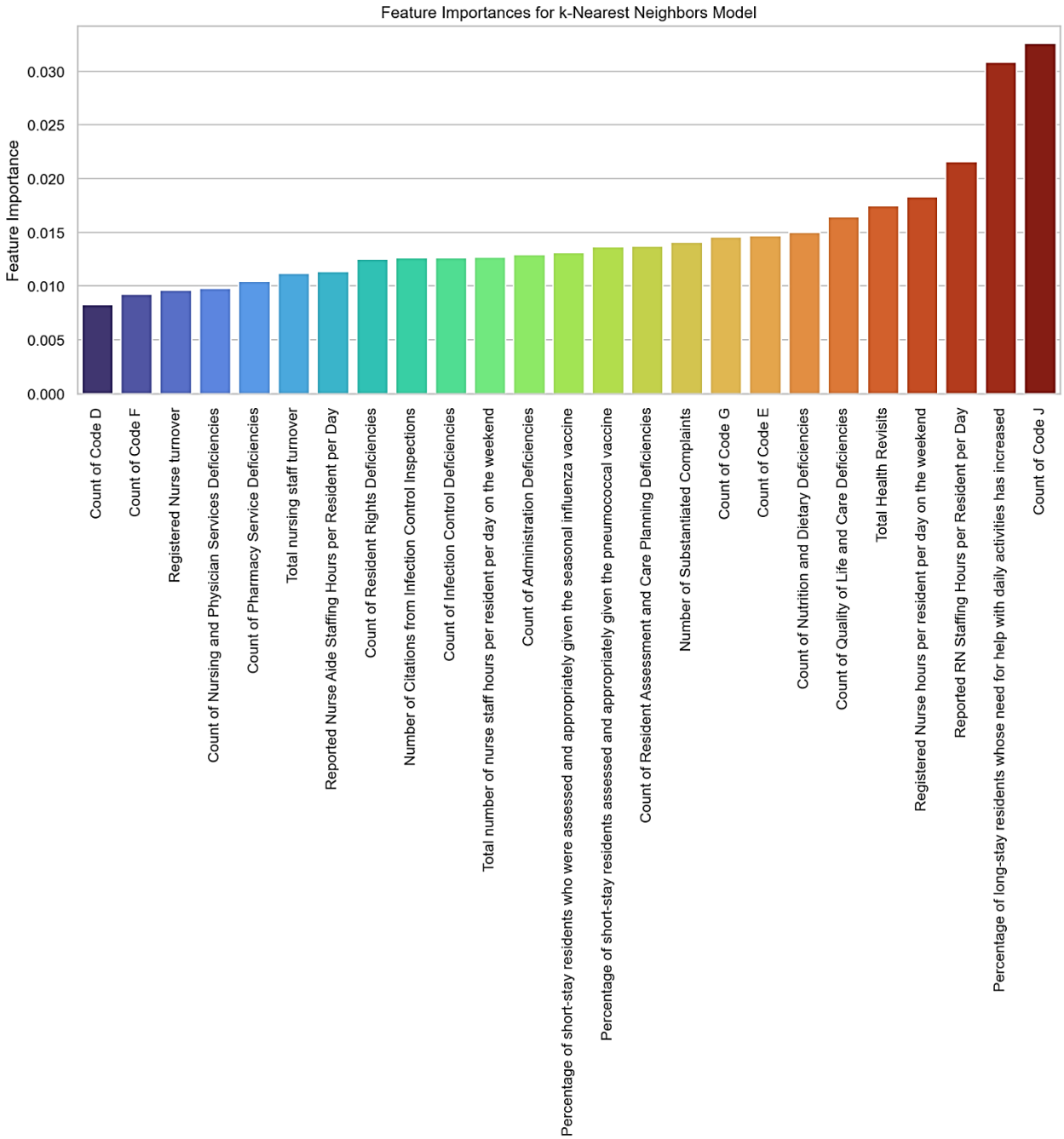

Reported Nurse Aide Staffing Hours per Resident per Day
Reported LPN Staffing Hours per Resident per Day
Reported RN Staffing Hours per Resident per Day
Total number of nurse staff hours per resident per day on the weekend
Registered Nurse hours per resident per day on the weekend
Reported Physical Therapist Staffing Hours per Resident Per Day
Total nursing staff turnover
Registered Nurse turnover
Number of administrators who have left the nursing home
Overall Rating
1.00 0.07 0.19 0.68 0.16 0.07 -0.13 -0.09 -0.09 0.23
0.07 1.00 -0.24 0.21 -0.24 0.06 0.03 0.09 -0.02 -0.03
0.19 -0.24 1.00 0.29 0.71 0.14 -0.15 -0.23 -0.03 0.31
0.68 0.21 0.29 1.00 0.28 0.13 -0.16 -0.12 -0.08 0.28
0.16 -0.24 0.71 0.28 1.00 0.11 -0.14 -0.19 -0.02 0.28
0.07 0.06 0.14 0.13 0.11 1.00 -0.07 -0.02 -0.02 0.12
-0.13 0.03 -0.15 -0.16 -0.14 -0.07 1.00 0.47 0.14 -0.24
-0.09 0.09 -0.23 -0.12 -0.19 -0.02 0.47 1.00 0.18 -0.25
-0.09 -0.02 -0.03 -0.08 -0.02 -0.02 0.14 0.18 1.00 -0.14
0.23 -0.03 0.31 0.28 0.28 0.12 -0.24 -0.25 -0.14 1.00
Reported Nurse Aide Staffing Hours per Resident per Day
Reported LPN Staffing Hours per Resident per Day
Reported RN Staffing Hours per Resident per Day
Total number of nurse staff hours per resident per day on the weekend
Registered Nurse hours per resident per day on the weekend
Reported Physical Therapist Staffing Hours per Resident Per Day
Total nursing staff turnover
Registered Nurse turnover
Number of administrators who have left the nursing home
Overall Rating
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Count of Freedom from Abuse and Neglect and Exploitation Deficiencies
Count of Quality of Life and Care Deficiencies
Count of Resident Assessment and Care Planning Deficiencies
Count of Nursing and Physician Services Deficiencies
Count of Resident Rights Deficiencies
Count of Nutrition and Dietary Deficiencies
Count of Pharmacy Service Deficiencies
Count of Environmental Deficiencies
Count of Administration Deficiencies
Count of Infection Control Deficiencies
Overall Rating
1.00 0.26 0.20 0.20 0.27 0.11 0.19 0.14 0.21 0.15 -0.21
0.26 1.00 0.35 0.28 0.46 0.26 0.39 0.20 0.26 0.33 -0.33
0.20 0.35 1.00 0.27 0.36 0.21 0.33 0.19 0.29 0.23 -0.29
0.20 0.28 0.27 1.00 0.30 0.18 0.25 0.16 0.25 0.22 -0.25
0.27 0.46 0.36 0.30 1.00 0.27 0.35 0.23 0.28 0.30 -0.30
0.11 0.26 0.21 0.18 0.27 1.00 0.24 0.21 0.17 0.23 -0.24
0.19 0.39 0.33 0.25 0.35 0.24 1.00 0.20 0.26 0.30 -0.24
0.14 0.20 0.19 0.16 0.23 0.21 0.20 1.00 0.24 0.09 -0.17
0.21 0.26 0.29 0.25 0.28 0.17 0.26 0.24 1.00 0.14 -0.22
0.15 0.33 0.23 0.22 0.30 0.23 0.30 0.09 0.14 1.00 -0.24
-0.21 -0.33 -0.29 -0.25 -0.30 -0.24 -0.24 -0.17 -0.22 -0.24 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Feature Importances (on the 25 selected features):

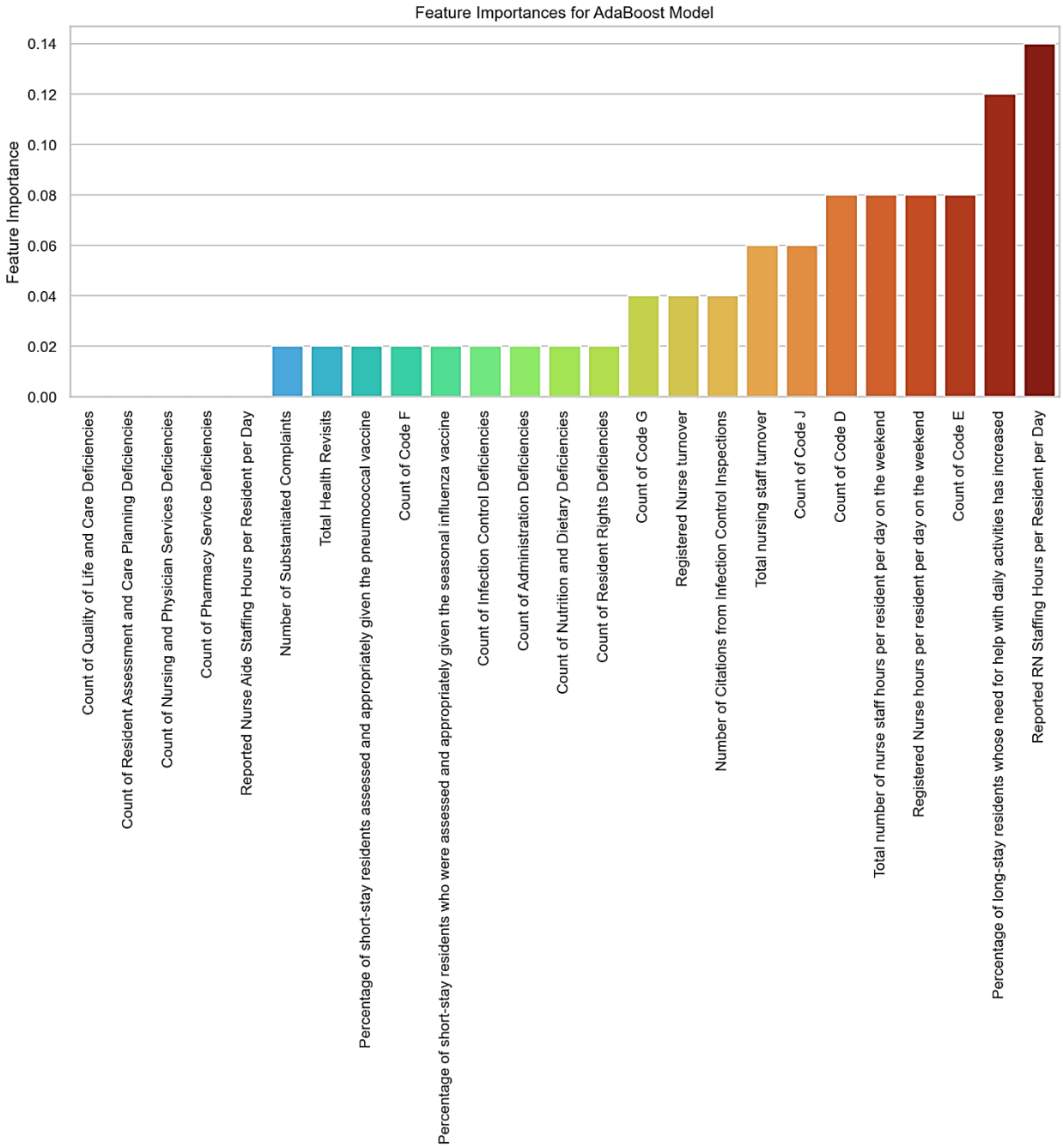
The feature importances for the kNN model.

Name	**Feature Importance** (sorted from greatest to least)
Count of Code J	0.032606
Percentage of long-stay residents whose need for help with daily activities has increased	0.030805
Reported RN Staffing Hours per Resident per Day	0.021606
Registered Nurse hours per resident per day on the weekend	0.018301
Total Health Revisits	0.017462
Count of Quality of Life and Care Deficiencies	0.016426
Count of Nutrition and Dietary Deficiencies	0.014996
Count of Code E	0.014675
Count of Code G	0.014576
Number of Substantiated Complaints	0.014108
Count of Resident Assessment and Care Planning Deficiencies	0.013688
Percentage of short-stay residents assessed and appropriately given the pneumococcal vaccine	0.013664
Percentage of short-stay residents who were assessed and appropriately given the seasonal influenza vaccine	0.013097
Count of Administration Deficiencies	0.012949
Total number of nurse staff hours per resident per day on the weekend	0.012702
Count of Infection Control Deficiencies	0.012653
Number of Citations from Infection Control Inspections	0.012628
Count of Resident Rights Deficiencies	0.012480
Reported Nurse Aide Staffing Hours per Resident per Day	0.011345
Total nursing staff turnover	0.011148
Count of Pharmacy Service Deficiencies	0.010433
Count of Nursing and Physician Services Deficiencies	0.009767
Registered Nurse turnover	0.009594
Count of Code F	0.009224
Count of Code D	0.008262


Feature Importances for k-Nearest Neighbors Model
Feature Importance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Count of Code D
Count of Code F
Registered Nurse turnover
Count of Nursing and Physician Services Deficiencies
Count of Pharmacy Service Deficiencies
Total nursing staff turnover
Reported Nurse Aide Staffing Hours per Resident per Day
Count of Resident Rights Deficiencies
Number of Citations from Infection Control Inspections
Count of Infection Control Deficiencies
Total number of nurse staff hours per resident per day on the weekend
Count of Administration Deficiencies
Percentage of short-stay residents who were assessed and appropriately given the seasonal influenza vaccine
Percentage of short-stay residents assessed and appropriately given the pneumococcal vaccine
Count of Resident Assessment and Care Planning Deficiencies
Number of Substantiated Complaints
Count of Code G
Count of Code E
Count of Nutrition and Dietary Deficiencies
Count of Quality of Life and Care Deficiencies
Total Health Revisits
Registered Nurse hours per resident per day on the weekend
Reported RN Staffing Hours per Resident per Day
Percentage of long-stay residents whose need for help with daily activities has increased
Count of Code J

The feature importances for the AdaBoost model.

Name	**Feature Importance** (sorted from greatest to least)
Reported RN Staffing Hours per Resident per Day	0.14
Percentage of long-stay residents whose need for help with daily activities has increased	0.12
Count of Code D	0.08
Count of Code E	0.08
Registered Nurse hours per resident per day on the weekend	0.08
Total number of nurse staff hours per resident per day on the weekend	0.08
Count of Code J	0.06
Total nursing staff turnover	0.06
Number of Citations from Infection Control Inspections	0.04
Registered Nurse turnover	0.04
Count of Code G	0.04
Percentage of short-stay residents assessed and appropriately given the pneumococcal vaccine	0.02
Number of Substantiated Complaints	0.02
Total Health Revisits	0.02
Percentage of short-stay residents who were assessed and appropriately given the seasonal influenza vaccine	0.02
Count of Code F	0.02
Count of Infection Control Deficiencies	0.02
Count of Administration Deficiencies	0.02
Count of Nutrition and Dietary Deficiencies	0.02
Count of Resident Rights Deficiencies	0.02
Reported Nurse Aide Staffing Hours per Resident per Day	0.00
Count of Resident Assessment and Care Planning Deficiencies	0.00
Count of Pharmacy Service Deficiencies	0.00
Count of Nursing and Physician Services Deficiencies	0.00
Count of Quality of Life and Care Deficiencies	0.00


Feature Importances for AdaBoost Model
Feature Importance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Count of Quality of Life and Care Deficiencies
Count of Resident Assessment and Care Planning Deficiencies
Count of Nursing and Physician Services Deficiencies
Count of Pharmacy Service Deficiencies
Reported Nurse Aide Staffing Hours per Resident per Day
Number of Substantiated Complaints
Total Health Revisits
Percentage of short-stay residents assessed and appropriately given the pneumococcal vaccine
Count of Code F
Percentage of short-stay residents who were assessed and appropriately given the seasonal influenza vaccine
Count of Infection Control Deficiencies
Count of Administration Deficiencies
Count of Nutrition and Dietary Deficiencies
Count of Resident Rights Deficiencies
Count of Code G
Registered Nurse turnover
Number of Citations from Infection Control Inspections
Total nursing staff turnover
Count of Code J
Count of Code D
Total number of nurse staff hours per resident per day on the weekend
Registered Nurse hours per resident per day on the weekend
Count of Code E
Percentage of long-stay residents whose need for help with daily activities has increased
Reported RN Staffing Hours per Resident per Day