AIT – 580: Data Analytics Research Project

Prediction of Overall Rating of a Nursing Home using Machine Learning

1. Abstract:

The Center for Medicare and Medicaid Services publishes a set of quality ratings for each nursing home in the country that participates in the Medicare or Medicaid program. These ratings help families in understanding the differences between different nursing homes and their quality. They also help in making important decisions like choosing the nursing home of their choice (Centers for Medicare and Medicaid Services, 2020). The current paper aims to answer certain research questions related to the dataset. It uses Machine Learning techniques to model the data and understand the relationships between different features. Given a set of attributes, a Machine Learning model is trained to predict the overall rating of a nursing home. The performance of different models is evaluated and the answers to the research questions are provided.

2. Introduction:

The Center for Medicare and Medicaid Services has introduced a 'star' based rating system that quantifies the quality of a nursing home. The 'Overall Rating' of a nursing home is calculated based on its performance on three domains which in turn is a rating.

Health inspections: Annual inspection surveys are conducted and the deficiencies in the nursing homes are noted. A rating is assigned by taking into consideration, the severity, and the number of deficiencies identified during these inspections for the past three years. It also includes the number of revisits that were required by the department to check whether a nursing home has corrected the faults identified during the inspections.

Staffing: RN is an abbreviation for Registered Nurse and LPN is an abbreviation of licensed practical nurse. The rating is based on the total RN number of hours per resident per day and the total number of hours of nurse staffing per resident per day.

Quality Measures: There are 15 Quality Measures present on the Nursing Home Compare website which comprise 9 long stay and 9 short stay evaluations.

As there are multiple features and it is hard to comprehend and analyze all of them, there exists a single feature called the 'Overall Rating' that rates a nursing home on a scale of 1 to 5 (lowest to highest) respectively. Explicit details about the three domains are found in (Centers for Medicare and Medicaid Services, 2020).

3. Literature Review:

In (Lee et al., 2020), Machine Learning methods are applied in Nursing Home Research to perform regression analysis on nurse staffing and falls in nursing homes. They have built six Machine Learning models and compared their performance against each other. They have concluded that Random Forest Regressor works best for the data followed by Logistic Regression. The model can be used to prevent the decrease in the number of nurses in nursing homes thereby improving the quality of life and decrease in the cost of health care. Very little research is being performed in this domain and many papers in relation to the current task were not found.

4. Dataset Description:

The dataset is extracted from the DATA.GOV website (Centers for Medicare and Medicaid Services, 2019). It consists of 86 columns and 15,437 records. Each record is associated with a nursing home. Some of the important columns describing the nursing home are,

Federal Provider Number – A unique number provided by the federal government to a nursing home.

Provider Name – The name of the nursing home.

Provider Address – The address of the nursing home.

Provider City – The city where the nursing home exists.

Provider State – The state where the nursing home exists.

Provider Zip Code – The zip code associated with the nursing home.

Provider Phone Number – The phone number of the nursing home.

Provider SSA County – The county in which the nursing home exists.

Provider County Name – The name of the county the nursing home belongs to.

Ownership Type – Describes whether it is a for-profit or government or non-profit entity.

Provider Type – Describes whether the nursing home is Medicare or Medicaid or Medicare and Medicaid.

Provider Resides in Hospital – A Boolean value indicating TRUE or FALSE to denote whether the provider resides in the hospital or not.

Legal Business Name – The legal business of the nursing home.

There are other features related to the measures described in the introduction section of the paper. For example, some of the features related to **'Health inspections'** include 'Rating Cycle 1 Total Number of Health Deficiencies', 'Rating Cycle 1 Number of Standard Health Deficiencies', 'Rating Cycle 1 Health Deficiency Score', etc. The features related to **'Staffing'** include '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', etc. The features related to '**Quality Measures**' include 'QM Rating', 'QM Rating Footnote', 'Long-Stay QM Rating', etc.

The summary statistics of the data are shown in Table 1. Since there are many columns, only a few of them are displayed in the table.

Features in the dataset	count	mean	std	min	25%	50%	75%	max
Total Amount of Fines in Dollars	15436.000	14540.740	48767.953	0.000	0.000	0.000	8125.000	1258368.000
Number of Certified Beds	15436.000	106.243	60.781	1.000	64.000	99.000	127.000	1389.000
Average Number of Residents Per Day	15367.000	85.854	52.669	1.000	51.200	78.100	107.500	751.000
Rating Cycle 2 Total Health Score	15341.000	62.820	102.687	0.000	16.000	36.000	76.000	3399.000
Total Weighted Health Survey Score	15341.000	61.024	67.106	0.000	20.667	41.333	77.333	1408.670
Rating Cycle 2 Health Deficiency Score	15341.000	59.811	87.580	0.000	16.000	36.000	72.000	2266.000
Rating Cycle 1 Total Health Score	15341.000	59.445	84.115	0.000	16.000	36.000	72.000	2784.000
Rating Cycle 1 Health Deficiency Score	15341.000	56.779	72.646	0.000	16.000	36.000	72.000	1505.000
Staffing Rating Footnote	1285.000	10.899	4.120	1.000	12.000	12.000	12.000	18.000
RN Staffing Rating Footnote	1285.000	10.899	4.120	1.000	12.000	12.000	12.000	18.000
Overall Rating Footnote	180.000	9.028	8.511	1.000	1.000	1.000	18.000	18.000
Health Inspection Rating Footnote	180.000	9.028	8.511	1.000	1.000	1.000	18.000	18.000
Rating Cycle 1 Total Number of Health Deficiencies	15341.000	8.198	7.363	0.000	3.000	6.000	11.000	79.000
QM Rating Footnote	210.000	8.024	8.253	1.000	1.000	2.000	18.000	18.000

Table 1. Summary Statistics for Features in the Dataset

The number of values for the feature 'Total Amount of Fines in Dollars' and 'Number of Certified Beds' are the highest. The mean, standard deviation, and 75% confidence interval values are maximum for the 'Total Amount of Fines in Dollars' feature. The 25% and 50% confidence interval values are maximum for 'Number of Certified Beds' and 'Average Number of Residents Per Day' columns.

5. Exploratory Data Analysis:

Figure 1 depicts the 'Total Weighted Health Survey Score' by the State. It can be observed that California and Texas have a higher total weighted health survey score. Places like Vermont, Puerto Rico, and Guam rank have a lower total weighted health survey score.

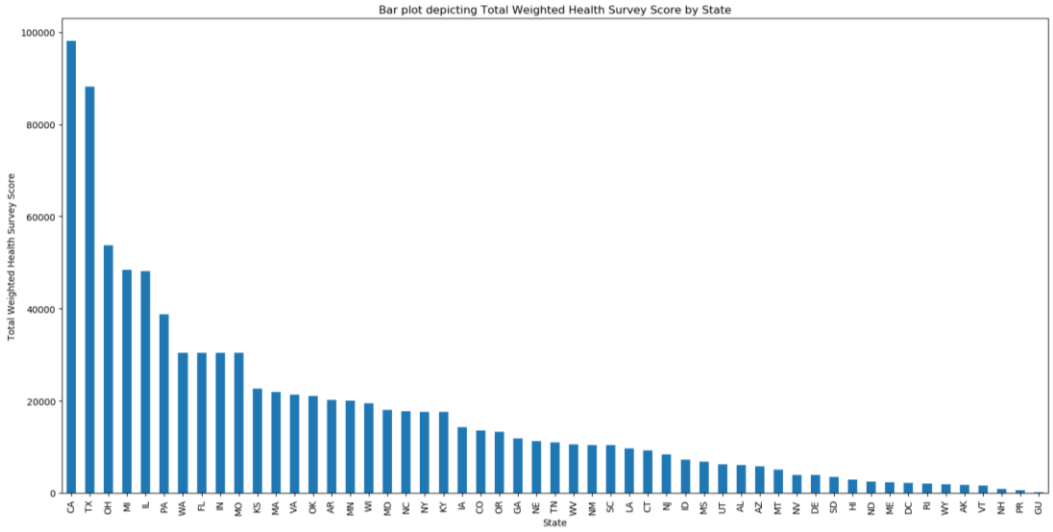

Fig. 1. Bar plot depicting Total Weighted Health Survey Score by State

Figure 2 shows a heatmap visualizing the sum of the 'Overall Rating' of nursing homes by state. It can be observed that California, Texas, Ohio, and Florida have the highest overall rating for nursing homes by state. Considering only 50 states, Alaska, Vermont, Delaware, Wyoming have lower 'Overall Rating' for nursing homes by state.

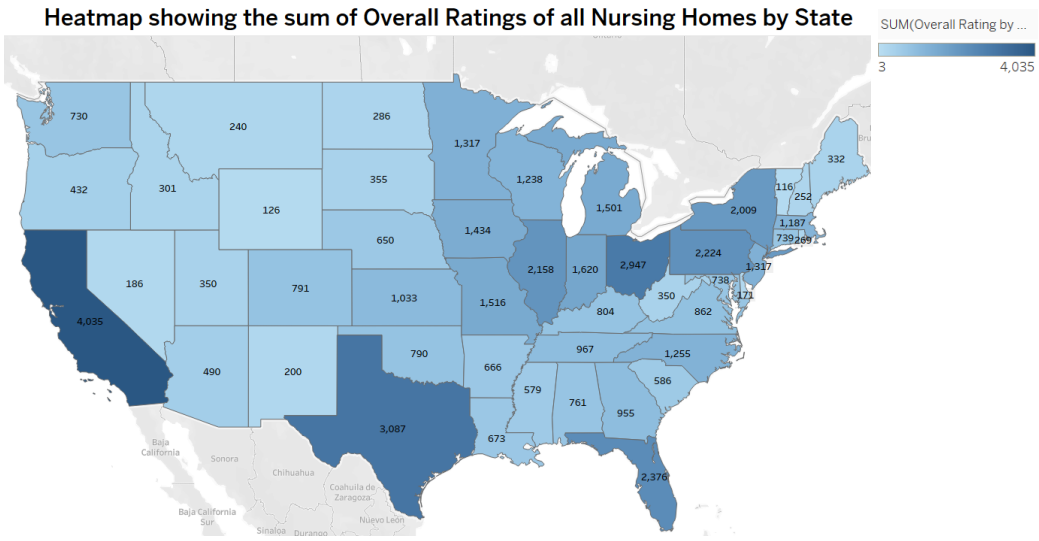

Fig. 2. Heatmap showing the sum of Overall Ratings of all Nursing Homes by State

Figure 3 depicts the box plots for the total amount of fines in dollars given to Nursing Homes by State. It's interesting to note that California is not in the top five states that have incurred the highest amount in the total number of fines, although Texas, Michigan, and Florida are present. A lot of outliers are observed in the box plot as many fines incurred by the nursing homes in each of the states is zero. The median for many states is also close to zero. The Interquartile range is highest for Michigan. The nursing homes in Texas and North Carolina have incurred the highest number of fines. Texas bears the highest value with a total amount greater than 1.2 Million dollars in fines.

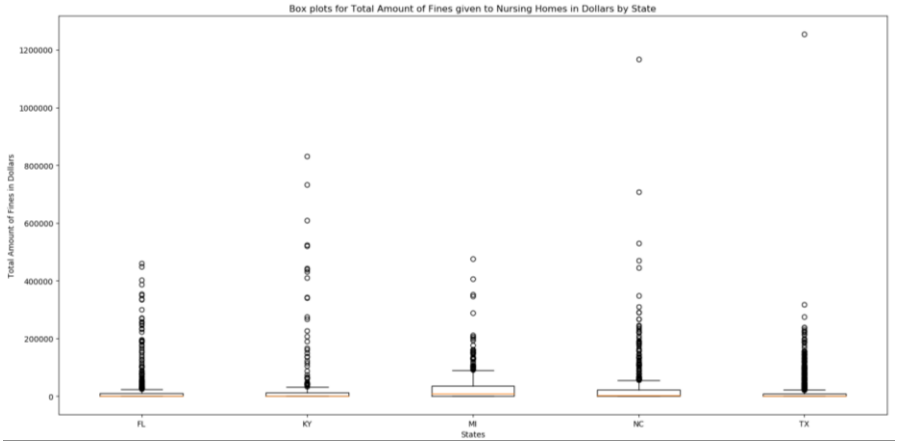

Fig. 3. Box plots for Total Amount of Fines given to Nursing Homes in Dollars by State

The correlation matrix for features with Pearson's correlation coefficient greater than 0.7 is visualized in Figure 4. The highest correlation is observed between 'Rating Cycle 2 Total Health Score' and 'Rating Cycle 2 Health Deficiency Score' along with 'Rating Cycle 1 Health Deficiency Score' and 'Rating Cycle 1 Health Total Health Score' and 'Rating Cycle 1 Health Deficiency Score'. Features with Pearson correlation coefficient values greater than 0.8, can be considered as highly correlated. One of the highly correlated columns from each set can be removed from the dataset as they do not provide any additional value.

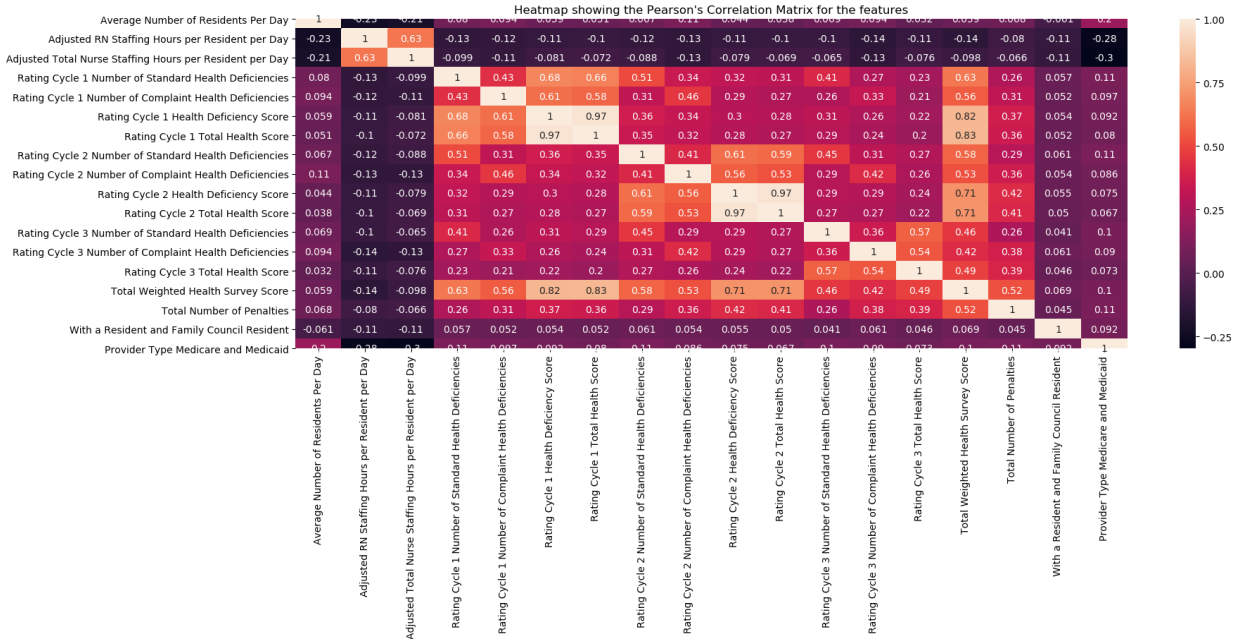

Fig. 4. Heat Map depicting the Pearson Correlation Coefficient for the most Correlated Features

Figure 5 shows a scatterplot between 'Overall Rating' and 'Total Weighted Health Survey Score' columns. Regression analysis is performed by fitting Linear regression to identify if there is any correlation between the two features.

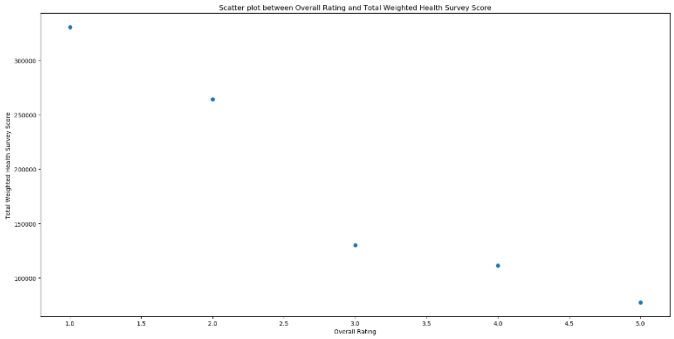

Fig. 5. Scatterplot between Overall Rating and Total Weighted Health Survey Score

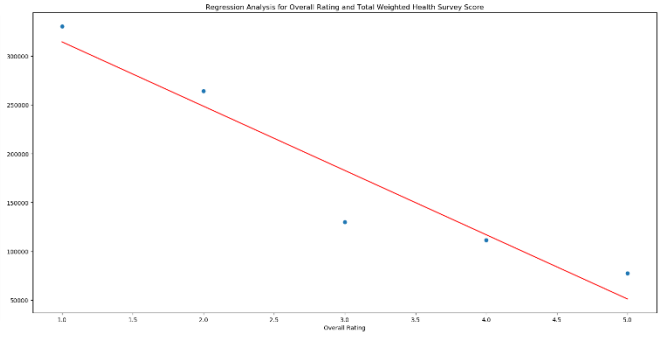

Fig. 6. Regression Analysis for Overall Rating Total Weighted Health Survey Score

A line was fitted to the data by using linear regression for the two features. An inverse relationship is depicted in figure 6. As the Overall rating increases, the Total Weighted Health Survey Score decreases. This observation is supported by the lines "Note that a lower survey score corresponds to fewer deficiencies and revisits, and thus better performance on the health inspection domain." (Centers for Medicare and Medicaid Services, 2019).

Chi-square test of independence was performed for the features, 'Provider Resides in Hospital' and 'Continuing Care Retirement Community'. The null hypothesis assumes that the features are independent. The alternate hypothesis assumes that the features are dependent. A chi-square test reveals the p-value as '7.1246454255565634e-12'. Since the p-value is less than 0.05 (at 95% confidence level), we reject the

null hypothesis and accept the alternate hypothesis concluding that the features are dependent. A similar test between the features, 'Provider Changed Ownership in Last 12 Months' and 'Abuse Icon' reveals that they are independent. Since the p-value at 0.05 level of significance (95% confidence level) was calculated to be 0.975 which is greater than 0.05.

6. Research Questions:

The paper aims to answer three research questions concerning dataset.

1. Can a Machine Learning model be developed that can allocate a class (overall rating) to a nursing home depending on patterns observed in the data?
2. If so, how reliable, and how accurate is it?
3. What are the most important factors that affect a nursing home's overall rating?

Answers to the above questions can help in identifying the most important factors of a nursing home affecting its overall rating. This can help a nursing home in making certain decisions to increase its rating and improve in the areas that affect its rating highly.

7. Data Pre-processing:

A lot of textual data such as Location, Federal Provider Number, Health Survey Dates are not very useful for model fitting and such columns were dropped from the dataset. Additionally, Adjusted features have been considered instead of individual reported features. For example, instead of considering 'Reported Nurse Aide Staffing Hours per Resident per Day', 'Adjusted Nurse Aide Staffing Hours per Resident per Day' was considered. The columns containing TRUE/FALSE (Boolean values) have been replaced with 1/0 respectively. The NaN values for certain columns like, 'QM Rating', 'Long-Stay QM Rating', 'Short-Stay QM Rating', 'Staffing Rating' were filled with zero.

Mean value of the column was used to replace NaN values in 'Adjusted Nurse Aide Staffing Hours per Resident per Day', 'Adjusted LPN Staffing Hours per Resident per Day', 'Adjusted Total Nurse Staffing Hours per Resident per Day', 'Average Number of Residents Per Day', etc. Dummy encoding was performed for certain columns like, 'With a Resident and Family Council' and 'Provider Type' as it contains categorical variables. The columns whose values are all NaN values are dropped from the dataset.

Instead of normalizing the whole dataset, each column has been normalized to a scale of [0,1]. If normalization were performed for the whole dataset, the column containing large numbers would become dominant in decision making. Pearson's Correlation Coefficient was calculated in between all the features of the dataset and one of the highly correlated columns in each set has been dropped from the dataset as they do not provide any additional value to the model. The total number of columns has been reduced from 86 to 52 after pre-processing.

7.1 Class distribution:

The number of records belonging to each class (Overall Rating) is observed. Class 5 has the highest in number of records whereas the lowest belong to class 1. But the difference between them is 1000 records which is tolerable. The distribution is natural, and no sampling methods are needed as no class is greatly under-represented or over-represented.

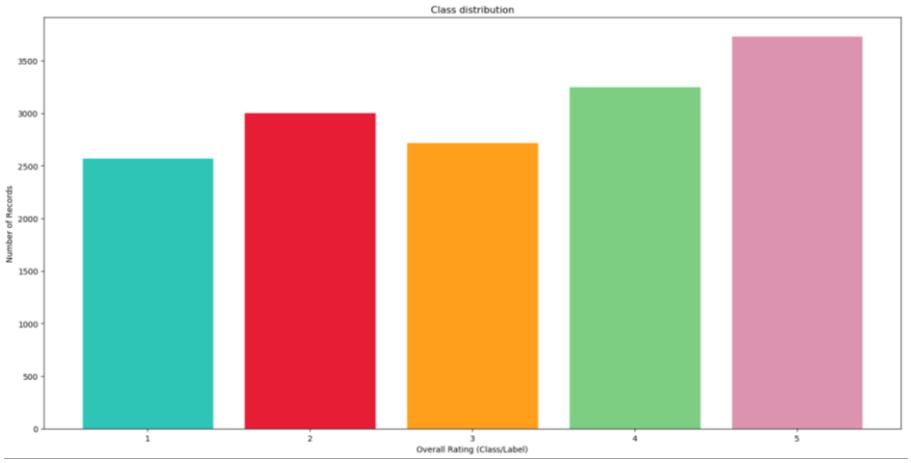

Fig. 7. Distribution of the number of Records by Class (Overall Rating)

8. Model fitting:

The current task of predicting the 'Overall Rating' is a classification task. Machine Learning algorithms such as Random Forest and Support Vector Machines are some of the robust algorithms that can be applied for Classification (Binary or Multi-class).

Random Forest is an ensemble technique that uses multiple decision trees and allocates a class to a given instance depending on its input features. It can perform both regression and classification tasks. Data is sampled with replacement and several decision trees are trained over the extracted samples. During sampling, a percentage of data is separated and labeled as OOB (Out of Bag) data. The trained decision tree is tested over the OOB data and an OOB error estimate is obtained. Best 'n' trees with the least OOB error are finalized and are regarded as a single Random Forest Classifier. Given a test instance, each decision tree inside a Random Forest Classifier predicts a class. Voting like mechanism is performed and the class predicted by most of the trees is generated as the output by the Random Forest Classifier. Due to the ensemble approach, Random Forests can model the data very efficiently (Leo et al., 2020).

Support Vector Machine is a supervised algorithm that can be used for Classification and Regression. It identifies the best hyperplane that can divide the data efficiently. It is also called as a Max – Margin classifier, as it tries to construct a hyperplane whose distance to the nearest element of both the classes is maximum (Cortes et al., 1995). Because of their proven efficiency in many classification tasks, both models were considered to model the current data.

8.1. Training:

The dataset is divided into the ratio of 80:20 where 80% of the data forms the train set and the rest 20% forms the test set. The features on which the models would be trained is termed as X_train. For the current dataset, X_train constitutes all the 51 columns (excluding the 'Overall Training' column) obtained after pre-processing the dataset. The target/column that the dataset would be trained to predict is termed as Y_train and for the current dataset, it would be the 'Overall Rating' column.

During the training process, the machine learning model is fitted to the dataset. However, the efficiency of modeling depends upon a set of inputs given to the model which are often called hyperparameters. Depending on the data, the hyper-parameters need to be adjusted and this process is called hyper-

parameter tuning. The best hyper-parameters for the training the Random Forest and the SVM models have been identified by a technique classed as K-Fold cross-validation. In this technique, the train set is divided into K folds/parts in which K-1 parts are considered as the train set and one part as the test set. A set of hyper-parameters is provided as the input and the model is trained over K-1 parts followed by testing on one part (Refaeilzadeh et al., 2009). This process is performed K times and the metrics are evaluated each time. The average of classification metrics such as average Accuracy, average Precision, average Recall, and average F1-score for the selected hyper-parameters are computed. The best set of hyper-parameters generates the highest test scores and is thus determined.

With respect to the current dataset, 5 – fold cross-validation was performed and the optimal hyperparameters for the Machine Learning models are identified as below.

RandomForestClassifier (bootstrap=True, class_weight=None, criterion='entropy', max_depth=30,

max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0,

min_impurity_split=None, min_samples_leaf=1, min_samples_split=2,

min_weight_fraction_leaf=0.0, n_estimators=500, n_jobs=None,

oob_score=False, random_state=None, verbose=0, warm_start=False)

SVC (C=10, cache_size=200, class_weight=None, coef0=0.0,

decision_function_shape='ovr', degree=3, gamma=1, kernel='rbf', max_iter=-1,

probability=False, random_state=None, shrinking=True, tol=0.001,

verbose=False)

The cross-validation scores for Random Forest and Support Vector Machines are represented in Table 2.

Model	Accuracy	Average Precision	Average Recall	Average F1-score
Random Forest	**0.960**	**0.960**	**0.960**	**0.960**
SVM	0.954	0.952	0.954	0.954

Table 2. Cross-validation scores for Random Forest and Support Vector Machines

At the end of cross-validation, the hyper-parameters are determined and the whole training set is trained as a single entity by the models. The models are ready to be tested on the test set.

8.2. Testing and Evaluation:

Each record in the test set is passed through the train set to predict the 'Overall Rating' for that record. Each record represents a nursing home. So, the model is essentially trained to predict the 'Overall Rating' for a nursing home, given a set of its features. The predictions are evaluated by the classification metrics – Precision, Accuracy, Recall, and F1-score.

Classification Metrics				
Class	**Precision**	**Recall**	**F1-score**	**Support**
1	**1.00**	0.97	0.98	513
2	0.97	0.96	0.96	635
3	0.95	0.94	0.94	540
4	0.98	**0.99**	**0.98**	625
5	0.98	**0.99**	**0.98**	739

Table 3. Class - wise Classification Metrics for Random Forest on the test set

Classification Metrics				
Class	**Precision**	**Recall**	**F1-score**	**Support**
1	0.97	**0.97**	**0.97**	513
2	0.96	0.95	0.95	635
3	0.95	0.95	0.95	540
4	0.94	0.95	0.94	625
5	**0.98**	**0.97**	**0.97**	739

Table 4. Class - wise Classification Metrics for Support Vector Machine on the test set

Model	**Accuracy**	**Average Precision**	**Average Recall**	**Average F1-score**
Random Forest	**0.97**	**0.97**	**0.97**	**0.97**
SVM	0.95	0.95	0.95	0.95

Table 5. Overall comparison of the performance of Machine Learning Models

It can be concluded that the Random Forest classifier is the best performing model for the current dataset. It has dominated SVM on all the classification metrics such as Accuracy, average Precision, average Recall, and average F1-score.

9. Answers to Research Questions:

9. 1. Can a machine learning model be developed that can allocate a class (overall rating) to a nursing home depending on patterns observed in the data?

Answer: Yes. A Machine Learning model can be developed that can categorize a nursing home into one of the five-star ratings present in the 'Overall Rating' feature. It can also be concluded that clear patterns exist in data with respect to the Overall Rating column and the Machine Learning models are being able to distinguish such patterns. So, it is possible to build reliable models over the current dataset and achieve the goal successfully. The procedure for building such a model is also stated above.

9. 2. If so, how reliable, and how accurate is it?

Answer: The performance of the models can be evaluated through different classification metrics. The Accuracy of the Random Forest model is calculated as 97%. The average precision with which the Random Forest model predicts the class is calculated as 97%. In simple terms, Recall is the ratio to what has been predicted correctly to what all should have been predicted correctly. The average Recall value is also 97%. F1-score is the harmonic mean between Precision and Recall. The higher the F1-score, the better the

model. The current Random Forest model achieves an average F1-score of 97% on the test set. With the help of the stated metrics, it can be concluded that the model is stable and very reliable.

9. 3. What are the most important factors that affect a nursing home's overall rating?

Answer: The trained Random Forest model provides a deep insight into the most important features of the dataset. This is because Random Forest builds decision trees that model the data by picking the features that are most important with respect to the target column. They calculate feature importance by using techniques such as 'entropy', 'gini', 'gain-ratio', etc. For the current model, the order of feature importance is represented in Fig. 8.

The top ten features that affect the 'Overall Rating' include, 'Health Inspection Rating', 'QM Rating', 'Staffing Rating', 'Total Weighted Health Survey Score', 'Adjusted RN Staffing Hours per Resident per Day', 'Rating Cycle 1 Total Health Score', 'Long-Stay QM Rating', 'Adjusted Total Nurse Staffing Hours per Resident per Day' and 'Rating Cycle 1 Health deficiency Score' and 'Short stay QM Rating'.

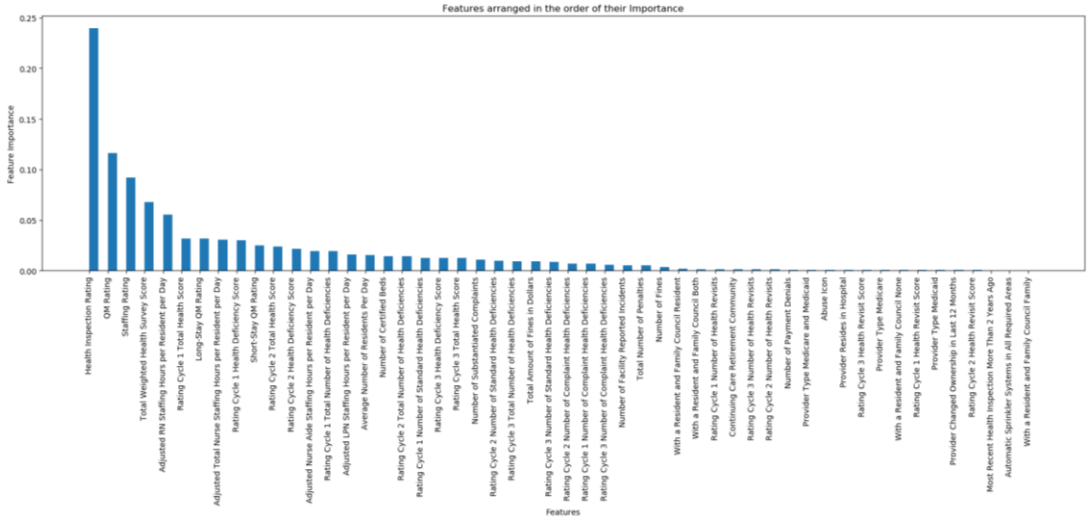

Fig. 8. Order of Feature Importance is determined by the Random Forest Classifier

10. Conclusion:

The current paper provides a detailed description of the methods, tools, techniques, and the procedure required to answer the proposed research questions. Two Machine Learning models (Random Forest and Support Vector Machine) are built, extensive analysis and testing were performed to gain insights from the trained models. Relevant conclusions have been drawn and answers have been provided. The nursing homes can make use of the wealth of information uncovered from the data and make the necessary decisions to improve their Overall Rating.

11. Limitations:

The interpretability of the Random Forest model is very low, as it is a collection of decision trees. When the trees are large, it becomes very difficult to visualize or interpret them. The memory and the computational costs associated with the training and testing of a Random Forest is high. The predictions are not very fast, and it may be a problem for time-sensitive applications (Jansen, 2018).

Though there are definitive rules over which the ratings are provided to the nursing homes, it is possible that the data is not accurate. The model may not give the same results for such inaccurate data records.

12. Recommended future analysis:

The current data defines different attributes of a nursing home and different ratings given to it, along with information related to its deficiencies, penalties, and fines. However, it does not describe the socio-economic factors of the geographical area where the nursing home resides. Such information can be added to the dataset to enrich the quality of the results and find new insights.

As the data increases, models such as Random Forest and SVM do not scale very well. Advanced models such as Neural Networks can be trained over the dataset to overcome this problem.

References

Centers for Medicare and Medicaid Services. (2019). Provider Info [Data file]. Retrieved on 29th March 2020 from https://catalog.data.gov/dataset/provider-info-1de34

Centers for Medicare and Medicaid Services. Design for Nursing Home Compare Five-Star Quality Rating System, 2020.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Jansen, Stefan. Hands-on Machine Learning for Algorithmic Trading: Design and Implement Investment Strategies Based on Smart Algorithms That Learn from Data Using Python. Packt Publishing, 2018.

Lee, Soo-Kyoung, et al. “Application of Machine Learning Methods in Nursing Home Research: Using Fall-Related Data in Nursing Homes.” 2020, doi:10.21203/rs.3.rs-21878/v1.

Leo, Breiman, and Adele Cutler. “Random Forests.” Random Forests - Classification Description. Retrieved on Retrieved on 10th May, 2020 from www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

Refaeilzadeh P., Tang L., Liu H. (2009) Cross-Validation. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA