

6-26-2019

Development of a Recommender System for Dental Care Using Machine Learning

Man Hung
Roseman University of Health Sciences

Julie Xu
University of Utah

Evelyn Lauren
University of Utah

Maren Wright Voss
Utah State University, maren.voss@usu.edu

Megan N. Rosales
University of Utah

Weicong Su
University of Utah

~~See next page for additional authors~~
Follow this and additional works at: https://digitalcommons.usu.edu/extension_research

Recommended Citation

Hung, M., Xu, J., Lauren, E., Voss, M.W., Rosales, M., Su, W., Ruiz-Negron, B., He, Y., Li, W., Licari, F. (2019). Development of a recommender system for dental care using machine learning. *SN Applied Sciences*, 1(7), 785.

This Article is brought to you for free and open access by the Extension at DigitalCommons@USU. It has been accepted for inclusion in Extension Research by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.

Authors


Man Hung, Julie Xu, Evelyn Lauren, Maren Wright Voss, Megan N. Rosales, Weicong Su, Bianca Ruiz-Negrón, Yao He, Wei Li, and Frank W. Licari



Research Article

Development of a recommender system for dental care using machine learning



Man Hung^{1,2,3,4,5}  · Julie Xu² · Evelyn Lauren^{2,3,6} · Maren W. Voss^{2,7} · Megan N. Rosales^{2,3} · Weicong Su^{2,3} · Bianca Ruiz-Negrón² · Yao He^{2,8} · Wei Li² · Frank W. Licari¹

© Springer Nature Switzerland AG 2019

Abstract

Resource mismanagement along with the underutilization of dental care has led to serious health and economic consequences. Artificial intelligence was applied to a national health database to develop recommendations for dental care. The data were obtained from the 2013–2014 National Health and Nutrition Examination Survey to perform machine learning. Feature selection was done using LASSO in R to determine the best regression model. Prediction models were developed using several supervised machine learning algorithms, including logistic regression, support vector machine, random forest, and classification and regression tree. Feature selection by LASSO along with the inclusion of additional clinically relevant variables identified 8 top features associated with recommendation for dental care. The top 3 features include gum health, number of prescription medications taken, and race. Gum health shows a significantly higher relative importance compared to other features. Demographics, healthcare access, and general health variables were identified as top features related to receiving additional dental care, consistent with prior research. Practicing dentists and other healthcare professionals can follow this model to enable precision dentistry through the incorporation of our algorithms into computerized screening tool or decision tree diagram to achieve more efficient and personalized preventive strategies and treatment protocols in dental care.

Keywords Machine learning · Predictive analytics · Dental care · Artificial intelligence · Oral health · NHANES · Preventive dental medicine

1 Introduction

Over the last decade, there has been an increase in the application of machine learning techniques in medical research. The medical field have benefited from using machine learning to perform various functions such as improving upon patient risk score systems, predicting

the onset of disease, and streamlining hospital operations [1]. The increase in popularity of machine learning approaches is partially related to its non-parametric nature, which does not rely on assumptions of a traditional statistical approach [2], along with its ability to fine-tune model parameters such as bias and accuracy and choose its best predictors. A suitable target field for

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s42452-019-0795-7>) contains supplementary material, which is available to authorized users.

✉ Man Hung, mhung@roseman.edu | ¹College of Dental Medicine, Roseman University of Health Sciences, 10849 S. River Front Parkway, South Jordan, UT 84095, USA. ²Department of Orthopaedic Surgery Operations, University of Utah, Salt Lake City, USA. ³Department of Mathematics, University of Utah, Salt Lake City, USA. ⁴Utah Center for Clinical and Translational Science, Salt Lake City, USA. ⁵Huntsman Cancer Institute, Salt Lake City, USA. ⁶Department of Economics, University of Utah, Salt Lake City, USA. ⁷Utah State University, Salt Lake City, USA. ⁸University of Utah Alzheimer's Center, Salt Lake City, USA.



SN Applied Sciences

(2019) 1:785

| <https://doi.org/10.1007/s42452-019-0795-7>

Received: 9 February 2019 / Accepted: 17 June 2019

Published online: 26 June 2019

SN Applied Sciences
A SPRINGER NATURE journal

machine learning is in the direction of health services to areas of greatest need. Resource mismanagement in primary oral health care places an undue burden on general health service delivery and would be an ideal area of research using machine learning. For example, when emergency departments address non-traumatic dental problems, patients are generally treated by providers without specialized dental training, and extensive post-discharge follow up by a dental practitioner is typically still required [3]. The result is added health system burden without any improvements to the patient experience or outcomes in the long term.

There have been some successful examples of using statistical models to improve oral health delivery and outcomes. Northern Germany used a statistical model to assess supply and demand, which allowed them to improve dental care recommendations to their citizens [4]. Developed and developing nations around the world have done similar analyses and instituted primary oral health care changes to great effect, such as distributing workflow from dental practitioners to dental therapists and employing more community dental health coordinators and dental technologists [5]. The United States could similarly benefit from a streamlined approach to assessing unmet need in supply and demand through the development of a predictive model for recommendations in dental care, given the diverse scope of needs in the country and the imperative to decrease costs of health care delivery.

Certainly, there is an economic and individual benefit to streamlining dental care access. Perhaps more profound than the global economic costs of dental diseases—with \$298 billion in direct treatment costs and another \$144 billion in indirect costs associated with productivity loss—are the costs of failing to treat dental disease. Untreated dental diseases can cause serious health problems such as the development of dentofacial anomalies, severe toothache, dental abscess, destruction of bone, and spread of infection via bloodstream [6]. Individuals face losses in productivity as they must take time off from work or caring for others to address these easily preventable health problems. Untreated dental disease can also lead to tooth loss, which has a highly negative impact on quality of life [7, 8]. Poor oral health is further associated with lifestyle-related comorbidities such as diabetes mellitus, osteoporosis, and rheumatoid arthritis [9–11]. Accessing necessary care is not as simple as finding a dentist as many patients experience barriers to access [12], including cost related barriers [13, 14], anxiety related to dental procedures [13, 15, 16], and socioeconomic limitations such as lack of dental insurance coverage [12, 17, 18].

A steady increase in the prevalence of dental caries [6, 19–21] and the varying implications of poor oral

health [22] have prompted the passage of public policy and oral health imperatives that aim to increase access to dental care for all people. To date, policy changes have addressed issues of access for people of different socioeconomic statuses, provided for the creation of school-based dental clinics, promoted interventions at crucial early stages of development, and trained dental care providers to service areas of increased need [21]. What has not been addressed in these improvements is a method to streamline identification of those with the greatest need for dental care. As access to dental care is recognized as a critical public health concern, it is important to be able to devise some kinds of mechanisms for identifying those in greatest need. Machine-learning application that capitalizes on information extracted from large stores of public health data is an opportunity to provide resources in a way that is targeted and personalized.

Dentistry is one area of medicine that can benefit from artificial intelligence (the field of data science) through the use of machine learning approach. Bringing data science to the field of oral health can efficiently address both its public health burden and the financial impact, as cost-utility analyses help identify the impact of evidence-based community interventions [23]. Accurately determining the characteristics of people who need the most assistance procuring dental care is another area where data science can shed light on effective implementations for oral health interventions. A triage system could recommend appropriate care for groups of cases based on the predictive model identifying those who are most likely to fail to access needed dental services. Given the spectrum and breadth of the unmet need demonstrated in previous research, a machine learning approach would be appropriate to create a data-based triage system. The goal of this study is to use machine learning techniques via artificial intelligence to create a useful, predictive model for dental care recommendations based on individualized need.

2 Methods

2.1 Data source

The data set was obtained from the CDC's National Health and Nutrition Examination Survey (NHANES) 2013–2014 cycle, an annual survey designed to measure diet, health, and nutrition of the U.S. population. The target population for the NHANES were the noninstitutionalized civilian residents of the United States. For the 2013–2014 survey, 14,332 individuals from 30 different survey locations were sampled; of which, 10,175 completed the interview

and 9813 were examined. To produce a reliable estimate across different groups within the population, the data included the oversampling of racial minorities, persons at or below 130 percent of the Department of Health and Human Services poverty guidelines, and persons aged 80 years and older. NHANES collected data through a series of interviews, questionnaires, and health examinations to gain information on lifestyle, diet, overall health status, socioeconomic status, and demographics. While the initial release portions of the data from the more recent NHANES surveys and exams were available, they were not complete. At this time, only the NHANES 2013–2014 survey contained the complete set of data files available for study.

2.2 Outcome

For this study, we used “Overall Recommendation for Care” as the outcome of interest, which was an assessment by a qualified oral health clinician of the need and status of a participant’s current state of oral health. The recommendation for care was assessed for all participants aged one year and older after an oral health examination is conducted by a licensed dentist. After the data set was processed to include the final relevant features, the sample size for this outcome variable was 2141 people. The “Overall Recommendation for Care” is an ordinal variable consisting of 4 options based on the urgency of which a person should see a dentist. In the 2013–2014 data, 2 people were recommended to “See a dentist immediately”, 128 people were recommended to “See a dentist within the next 2 weeks”, 1068 people were recommended to “See a dentist at your earliest convenience”, and 945 people were recommended to “Continue your regular routine care.” The option “See a dentist immediately” was not utilized due to insufficient number of observed cases.

2.3 Approaches

Variables in the NHANES data set, including features of racial demographics, socioeconomic status, lifestyle, and clinical characteristics, were used as inputs to determine their relationships with an overall recommendation for dental care. Machine learning algorithms were used to develop predictive models of the outcome. In machine learning, large amounts of data are inputted into an algorithm, where complex patterns could be recognized to come up with predictions for future outcomes for the variable. Machine learning is useful for clinical diagnostic applications because of its ability to produce cost-effective and efficient tools.

There was a total of 10,175 cases and 2365 variables present in the complete dataset. Cases with missing data of over 50 percent were excluded. A visual inspection and content review of the complete variable list were conducted to additionally exclude variables (e.g., subject sequence number) irrelevant to recommendation for care. Additionally, redundant variables (i.e., multiple ways of identifying participant age) were excluded. This resulted in a sample size of 8936 and 359 variables for input into an initial model for variable selection. The selected variables were evaluated using Chi square tests with 14 variables showing a significant relationship with recommendation for care ($p < 0.001$). Additional testing and culling revealed no significant decrease in accuracy when the features selected were further narrowed down to 8 relevant variables for inclusion in the final machine learning application. The final effective sample size was 2141 cases.

2.4 Classification methods

Classification methods were used to develop the prediction models. In this study, we used several different supervised machine learning classification algorithms, including logistic regression, support vector machine (SVM), random forest, k-nearest neighbors, and Classification and Regression Tree (CART).

The Least Absolute Shrinkage and Selection Operator (LASSO) method was used to determine the best regression model for the variables in a large dataset. LASSO conducts variable selection while also performing a form of regularization. Minimizing the noise from variables with very small coefficients maintains accuracy without over-fitting. LASSO offers automatic variable selection and compact parametric estimation to simplify our model [24]. Research has shown that LASSO outperforms other feature selection methods such as Random Forest, two-sample t test and CART with respect to selection performance and prediction error rates [25]. In particular, LASSO is able to select the true variables more accurately and with fewer fake predictors than Random Forest, two-sample t -test, and CART [25]. LASSO is generally able to aggressively exclude correlated variables, thus resulting in a more parsimonious set of predictors and models.

Feature selection with LASSO was implemented using the `glmnet` library in RStudio 1.1.456 in this study. Python 3.7.0 was used to program the Logistic Regression and Support Vector Machine algorithms. Python 3.6.4 was used to program the CART and Random Forest algorithms. SAS 9.4 and SPSS 25 used for data processing and cleaning. Graphviz 2.38 was used to provide visualization for the

figures and decision trees. All statistical tests were considered significant at $p < 0.05$ two-tailed.

In order to handle the imbalanced data to improve the classification algorithms and reduce classification bias in determining model accuracy, random over-sampling resampling technique was applied to the minority classes of the “Overall Recommendation for Care” class to increase the frequency of the minority classes. The accuracies of logistic regression, support vector machine (SVM), random forest, k-nearest neighbors, and Classification and Regression Tree (CART) were evaluated on the balanced data set with 20,400 cases for each of the 3 “Overall Recommendation for Care” classes. It is worth to note that SVM is generally sufficient to handle imbalanced data; however, there is no one-size-fits-all methods in model building. SVM may not always be guaranteed to have the best performance with imbalanced data. Therefore, it is necessary to explore other methods in addition to SVM in order to obtain the best results.

Because the large size of the resulting balanced data set resulted in a decision tree that was unmanageable, a smaller balanced data set with 1700 cases for each “Overall Recommendation for Care” class was used to construct a simpler decision tree with CART. This purpose of this tree was to give dental providers a practical visual tool for using the model constructed in this study. For the sake of simplicity and to ensure the best accuracy possible, the following restrictions were applied to the decision tree. Please refer to the Python code in the Appendix for further details of the variable tuning for this tree.

1. The decision tree could have at most 8 levels, excluding the root and bottom levels.
2. The decision tree could have at most 300 leaf nodes.
3. Each node should have at least 100 samples to split
4. In order to be considered a node, each node must have a sample size of at least 20.

The practical decision tree was generated with Graphviz 2.38.

3 Results

There were 2141 individuals in the NHANES 2013–2014 used for this study. The average age was 61.44 (SD = 11.80) with a range from 40 to 80. The sample had 44.8% male (N = 959) and 55.2% female (N = 1182). 58.2% of the participants were married (N = 1246). Non-Hispanic White participants comprised the majority (50.8%, N = 1087),

followed by Non-Hispanic Black (22.2%, N = 476) and Mexican–American (10.7%, N = 230) (See Table 1 for more details on demographics, education, and income). The LASSO coefficients of the highest magnitude are displayed in Table 2, which includes 11 variables classified as significant by LASSO and 3 variables chosen for their clinical relevance. Further testing narrowed down the variables into the top 8 features.

Among the other machine learning models applied using the top 8 features, logistic regression performed the worst with accuracy, precision, and sensitivity below 0.54 and specificity of 0.77 (see Table 3). The CART algorithms and Random Forest both performed well, with accuracy, precision, and sensitivity each over 0.84 and specificity of 0.92 for both methods.

Figure 1 displays the relative importance of the top 8 features. The figure shows gum health as having a significantly higher relative importance towards overall recommendation for dental care compared to other features. The next most important features include number of prescription medications taken and race. Other variables with high importance include general health, health insurance, and country of birth.

Supplementary Figure 2 displays a diagram of decision tree. This tree diagram utilized information such as patient demographics, health status, and healthcare access from the top 8 features in order to provide a visual aid that facilitates the recommendation for care in a dental practice setting. Since the decision tree diagram shown in supplementary Fig. 2 was relatively large, we divided it into three separate supplementary figures (Figs. 3a, 3b, 3c) for readability. Due to the dimensions of the figures and limited margin size, Supplementary Figs. 2, 3a, 3b, and 3c will be made available upon request.

4 Discussion

Underutilization of oral care has led to serious health consequences associated with a decreased quality of life and increased economic costs and public health burden. Machine learning, or artificial intelligence in general provides an important tool in identifying the barriers of receiving necessary dental care and can be used in identifying “at risk” populations. Identification of the features which suggest recommendation for additional dental care is a valuable tool in the public health toolbox for finding, screening, and providing necessary oral health interventions. The aim of this study was to create an accurate, computerized oral care recommender system

Table 1 Demographic characteristics (n = 2141)

Variables	Mean (SD)	Median	Min	Max	n	%
Age (years)	61.44 (11.80)	62	40	80	2141	
<i>Gender</i>						
Male					959	44.8
Female					1182	55.2
<i>Race/ethnicity</i>						
Mexican–American					230	10.7
Other Hispanic					162	7.6
Non-Hispanic White					1087	50.8
Non-Hispanic Black					476	22.2
Non-Hispanic Asian					186	8.7
<i>Marital status</i>						
Married					1246	58.2
Widowed					280	13.1
Divorced					317	14.8
Separated					60	2.8
Never married					171	8.0
Living with partner					66	3.1
Missing					1	< 1.0
<i>Education level—adults 20+</i>						
Less than 9th grade					191	8.9
9th–11th grade					290	13.5
High school graduate/GED or equivalent					484	22.6
Some college or AA degree					629	29.4
College graduate or above					544	25.4
Missing					3	< 1.0
<i>Annual household income</i>						
\$0 to \$4999					33	1.5
\$5000 to \$9999					86	4.0
\$10,000 to \$14,999					164	7.7
\$15,000 to \$19,999					143	6.7
\$20,000 to \$24,999					187	8.7
\$25,000 to \$34,999					200	9.3
\$35,000 to \$44,999					165	7.7
\$45,000 to \$54,999					87	4.1
\$55,000 to \$64,999					69	3.2
\$65,000 to \$74,999					24	1.1
\$75,000 to \$99,999					182	8.5
\$100,000 and over					397	18.5
Missing					404	18.9

through pattern recognition from a large database consisting of demographics, lifestyle, and oral health variables. The implementation of computer assisted referral systems can contribute to improvements in quality of life and reduction in excessive oral and general healthcare related costs.

Prior research has identified socioeconomic barriers such as healthcare costs [13] and racial disparities [26], along with psychological barriers such as dental anxiety [27] as reasons for not seeking appropriate dental care. According to the results from this study, two demographic variables and two healthcare access variables

Table 2 Top features included in the machine learning models

Variable Name	Variable Description	Total (N)	See Dentist			p-value	LASSO Coefficient
			2 weeks (n)	Convenience (n)	Continue reg. care (n)		
DMDBORN4	Country of birth	2141	128	1068	945	< .001	-0.5389
HUQ030	Routine place to go for healthcare	2141	128	1068	945	< .001	0.3689
RIDRETH3	Race/Hispanic origin w/NH Asian	2141	128	1068	945	< .001	0.3045
DMDHHSZA	# of children 5 years or younger in HH	2141	128	1068	945	< .001	-0.2654
RIAGENDR	Gender	2141	128	1068	945	< .001	-0.1994
HUQ010	General health condition	2141	128	1068	945	< .001	0.1835
HIQ011	Covered by health insurance	2141	128	1068	945	< .001	0.1832
DSD010AN	Any Antacids Taken?	2141	128	1068	945	< .001	0.0689
HUQ071	Overnight hospital patient in last year	2141	128	1068	945	< .001	-0.0787
OHQ845	Rate the health of your teeth and gums	2141	128	1068	945	< .001	-0.0787
DMDFMSIZ	Total number of people in the Family	2141	128	1068	945	< .001	0.1396
CSQ202*	Had persistent dry mouth in past 12 month	2141	128	1068	945		
RXDUSE*	Taken prescription medicine in past month	2141	128	1068	945		
RXDOUNT*	Number of prescription medicines taken	2141	128	1068	945		

*Did not perform LASSO; variables were selected due to its clinical relevance

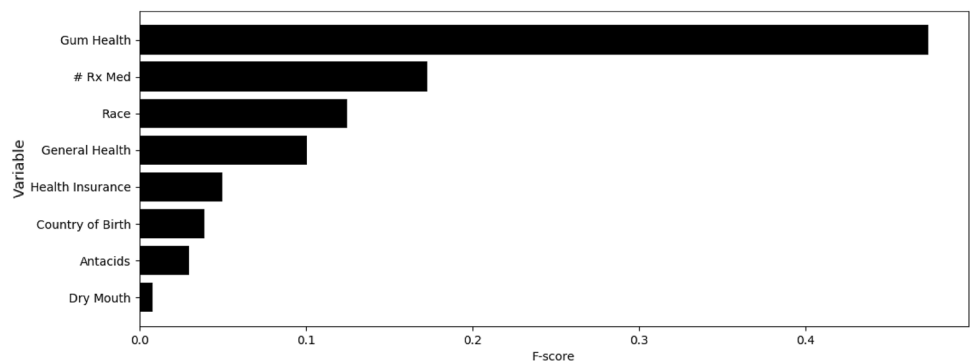
Table 3 Performance metrics of machine learning models with the final 8 features

Classifier	Accuracy	Precision	Sensitivity	Specificity
Random forest	0.841	0.840	0.841	0.920
CART	0.840	0.840	0.840	0.920
k-nearest neighbors	0.824	0.823	0.824	0.912
Support vector machine	0.719	0.714	0.719	0.860
Logistic regression	0.534	0.522	0.534	0.767

were identified as top features in predicting needs for individualized dental treatment. The top identifiers included race. This is consistent with previous research

indicating that demographics are an important factor related to receiving appropriate dental care [28, 29]. The LASSO algorithm in this study also identified general health as a factor most related to recommendations for dental care, consistent with prior literature on the relationship between general health and dental disease [9–11]. While a simple knowledge of these factors may be helpful for general screening purposes, an even more relevant application of this research would be the creation of computer assisted screening tools. The specific characteristics of the individual can then be considered and weighted against their relative risk for untreated dental disease, allowing referrals for oral health screening and treatment when necessary that are more targeted, more precise and more efficient.

Fig. 1 Relative importance of variables



We believe that our model can have many practical applications in the era of evidence-based dentistry. For practicing dentists, following our model using a computerized tool or simply the decision tree diagram can lead to efficient resource utilization based on visit examination and lead to increased time efficiency. On a larger scale, our model can help informing useful implementations through public health programs or application of tools targeted toward individuals with the greatest needs, leading to value-based dental care and precision dentistry.

Despite the many benefits of our model, there are limitations to note. Using LASSO compromised some model complexity in exchange for the ability to select the most relevant predictors; however, this was an efficient approach for cross-sectional study. A longitudinal analysis in the future may benefit from ridge regression or Elastic Net, which incorporates penalties of ridge regression and LASSO, to see if the variable coefficients of the model change.

5 Conclusion

This study created a computerized oral care recommender model using machine learning techniques and a large, national database. Our results reinforce the

importance of public policy incentives geared toward helping at-risk populations and inform public policy regarding improvements of oral health delivery. Not only can the model be useful for practicing dentists, it can become a tool for other healthcare professionals for accurate dental care referral as well. It encourages the effectiveness of preventative oral care strategies that can eventually reduce treatment costs and set the groundwork for future developments of computerized tool that can pave the way toward individualized healthcare.

Acknowledgements This project was supported by the Roseman University College of Dental Medicine Clinical Outcomes Research and Education, the University of Utah Undergraduate Research Opportunity Program, and the Population Health Research Foundation with funding in part from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant 5UL1TR001067.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix A

```
import pandas as pd
from sklearn import cross_validation
from sklearn.model_selection import train_test_split
import os
os.environ["PATH"] += os.pathsep + 'C:/Program Files/graphviz-2.38/release/bin'

from numpy.random import seed
seed(5)
from tensorflow import set_random_seed
set_random_seed(42)
from sklearn import metrics
from sklearn import tree
from IPython.display import Image
from sklearn.tree import export_graphviz
import pydotplus
from sklearn.ensemble import RandomForestClassifier
import numpy as np
import matplotlib.pyplot as plt

original = pd.read_csv('C:/Users/Weicong Su/Desktop/DentalCare/NHANES2013-
2014_8.6.18_FinalUnbalanced.csv',
                      usecols = ['OHAREC', 'DMDBORN4', 'RIDRETH3', 'HUQ010', 'HIQ011',
' DSD010AN', 'OHQ845', 'CSQ202', 'RXDCOUNT'])
original = original.dropna()

Class2=['2']
Class3=['3']
Class4=['4']

df2 = original.loc[original['OHAREC'].isin(Class2)] # there exist 128 records of Class2
df3 = original.loc[original['OHAREC'].isin(Class3)] # there exist 1068 records of Class3
```

```
Class4=['4']

df2 = original.loc[original['OHAREC'].isin(Class2)] # there exist 128 records of Class2
df3 = original.loc[original['OHAREC'].isin(Class3)] # there exist 1068 records of Class3
df4 = original.loc[original['OHAREC'].isin(Class4)] # there exist 945 records of Class4
df5 = df4.head(89)
print(len(df2))
print(len(df3))
print(len(df4))
original = original.append(df5)
for i in range(0,7): # append 7 copies of Class 2 to make balanced data
    original = original.append(df2,ignore_index=True)

df2 = original.loc[original['OHAREC'].isin(Class2)] # there exist 1024 records of Class2
df3 = original.loc[original['OHAREC'].isin(Class3)] # there exist 1068 records of Class3
df4 = original.loc[original['OHAREC'].isin(Class4)] # there exist 1034 records of Class4
print(len(df2))
print(len(df3))
print(len(df4))

df6=df2.head(676)
df7=df3.head(632)
df8=df4.head(666)
original=original.append(df6)
original=original.append(df7)
original=original.append(df8)

df2 = original.loc[original['OHAREC'].isin(Class2)] # there exist 1700 records of Class2
df3 = original.loc[original['OHAREC'].isin(Class3)] # there exist 1700 records of Class3
df4 = original.loc[original['OHAREC'].isin(Class4)] # there exist 1700 records of Class4
```

```
print(len(df2))
print(len(df3))
print(len(df4))

X = original.drop(columns=['OHAREC']) # This is a matrix.
y = original['OHAREC'] # This is a vector. Do NOT put double brackets.
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
X_train.rename(index=str, columns={"HUQ010": "General Health", "HIQ011": "Health Insurance", "DSD010AN": "Antacids", "DMDBORN4": "Country of Birth", "RIDRETH3": "Race", "RXDCOUNT": "# Rx Med", "OHQ845": "Gum Health", "CSQ202": "Dry Mouth"}, inplace=True)

clfDT =
tree.DecisionTreeClassifier(max_depth=8,max_leaf_nodes=300,min_samples_split=100,min_samples_leaf=20)
clfDT.fit(X_train, y_train)
clfDT_ypred0 = clfDT.predict(X_test)
print('Accuracy = ', metrics.accuracy_score(y_true = y_test, y_pred = clfDT_ypred0))

# Create DOT data
dot_data = tree.export_graphviz(clfDT, feature_names=X_train.columns, class_names='234',
impurity=False, out_file=None)

# Draw graph
graph = pydotplus.graph_from_dot_data(dot_data)

# Show graph
Image(graph.create_png())

# Create PDF
graph.write_pdf("tree.pdf")

# Create PNG
graph.write_png("tree.png")
```

```
# Calculate feature importances
importances = clfDT.feature_importances_

# Sort feature importances in descending order
indices = np.argsort(importances)[0::]

# Rearrange feature names so they match the sorted feature importances
names = [X_train.columns[i] for i in indices]

# Create plot
plt.figure()

# Create plot title
#plt.title("Feature Importance Decision Tree")

# Add bars

plt.barh(range(X_train.shape[1]), importances[indices], color='k')

# Add feature names as x-axis labels
plt.yticks(range(X_train.shape[1]), names, rotation=0)

# Add axis labels
plt.xlabel('F-score')
plt.ylabel('Variable')

# Show plot
plt.show()
```

References

1. Callahan A, Shah NH (2018) Machine learning in healthcare. Key advances in clinical informatics. Elsevier, Amsterdam, pp 279–291
2. Kuo CY, Yu LC, Chen HC, Chan CL (2018) Comparison of models for the prediction of medical costs of spinal fusion in taiwan diagnosis-related groups by machine learning algorithms. *Healthc Inform Res* 24(1):29–37
3. Mostajer Haqiqi A, Bedos C, Macdonald ME (2016) The emergency department as a 'last resort': why parents seek care for their child's nontraumatic dental problems in the emergency room. *Community Dent Oral Epidemiol* 44(5):493–503
4. Jager R, van den Berg N, Hoffmann W, Jordan RA, Schwendicke F (2016) Estimating future dental services' demand and supply: a model for Northern Germany. *Community Dent Oral Epidemiol* 44(2):169–179
5. Mumghamba EG (2014) Integrating a primary oral health care approach in the dental curriculum: a tanzanian experience. *Med Princ Pract* 23(Suppl 1):69–77
6. Bagramian RA, Garcia-Godoy F, Volpe AR (2009) The global increase in dental caries. A pending public health crisis. *Am J Dent* 22(1):3–8
7. Akhter R, Hassan NM, Aida J, Zaman KU, Morita M (2008) Risk indicators for tooth loss due to caries and periodontal disease in recipients of free dental treatment in an adult population in Bangladesh. *Oral Health Prev Dent* 6(3):199–207
8. Marcenes W, Kassebaum NJ, Bernabe E et al (2013) Global burden of oral conditions in 1990–2010: a systematic analysis. *J Dent Res* 92(7):592–597
9. Lee JH, Lee JS, Park JY et al (2015) Association of lifestyle-related comorbidities with periodontitis: a nationwide cohort study in Korea. *Medicine (Baltimore)* 94(37):e1567
10. Boehm TK, Scannapieco FA (2007) The epidemiology, consequences and management of periodontal disease in older adults. *J Am Dent Assoc* 138(Suppl):26s–33s
11. Griffin SO, Barker LK, Griffin PM, Cleveland JL, Kohn W (2009) Oral health needs among adults in the United States with chronic diseases. *J Am Dent Assoc* 140(10):1266–1274
12. Doty HE, Weech-Maldonado R (2003) Racial/ethnic disparities in adult preventive dental care use. *J Health Care Poor Underserved* 14(4):516–534
13. Armfield J (2012) The avoidance and delaying of dental visits in Australia. *Aust Dent J* 57(2):243–247
14. Bagewitz IC, Soderfeldt B, Palmqvist S, Nilner K (2002) Dental care utilization: a study of 50- to 75-year-olds in southern Sweden. *Acta Odontol Scand* 60(1):20–24
15. Gordon SM, Dionne RA, Snyder J (1998) Dental fear and anxiety as a barrier to accessing oral health care among patients with special health care needs. *Spec Care Dentist* 18(2):88–92
16. Haumschild MS, Haumschild RJ (2009) The importance of oral health in long-term care. *J Am Med Dir Assoc* 10(9):667–671
17. Flores G, Tomany-Korman SC (2008) Racial and ethnic disparities in medical and dental health, access to care, and use of services in US children. *Pediatrics* 121(2):e286–e298
18. Liu J, Probst JC, Martin AB, Wang JY, Salinas CF (2007) Disparities in dental insurance coverage and dental care among US children: the National Survey of Children's Health. *Pediatrics* 119(Suppl 1):S12–S21
19. Denson L, Janitz AE, Brame LS, Campbell JE (2016) Oral cavity and oropharyngeal cancer: changing trends in incidence in the United States and Oklahoma. *J Okla State Med Assoc* 109(7–8):339–345
20. Imazato S, Ikebe K, Nokubi T, Ebisu S, Walls AW (2006) Prevalence of root caries in a selected population of older adults in Japan. *J Oral Rehabil* 33(2):137–143
21. George MC (2013) Public policy and legislation for oral health: a convergence of opportunities. *J Dent Hyg* 87:50–52
22. Petersen PE, Bourgeois D, Ogawa H, Estupinan-Day S, Ndiaye C (2005) The global burden of oral diseases and risks to oral health. *Bull World Health Organ* 83:661–669
23. Neumann PJ, Farquhar M, Wilkinson CL, Lowry M, Gold M (2016) Lack of cost-effectiveness analyses to address healthy people 2020 priority areas. *Am J Public Health* 106(12):2205–2207
24. Wang H, Li G, Tsai C-L (2007) Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J R Stat Soc Ser B (Stat Methodol)* 69(1):63–78
25. Lu F, Petkova E (2014) A comparative study of variable selection methods in the context of developing psychiatric screening instruments. *Stat Med* 33(3):401–421
26. Gupta N, Vujicic M, Yarbrough C, Harrison B (2018) Disparities in untreated caries among children and adults in the U.S., 2011–2014. *BMC Oral Health* 18(1):30
27. Skaret E, Raadal M, Berg E, Kvale G (1999) Dental anxiety and dental avoidance among 12 to 18 year olds in Norway. *Eur J Oral Sci* 107(6):422–428
28. Wamala S, Merlo J, Boström G (2006) Inequity in access to dental care services explains current socioeconomic disparities in oral health: the Swedish National Surveys of Public Health 2004–2005. *J Epidemiol Community Health* 60(12):1027–1033
29. Stella MY, Bellamy HA, Kogan MD et al (2002) Factors that influence receipt of recommended preventive pediatric health and dental care. *Pediatrics* 110(6):e73

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.