

## Predictive Analytics Based on the NHANES 1999-2016 Dataset for the Hepatitis an Antibody Prediction: A Python Case Study

Mai Thi Hoang Ta<sup>1</sup>, Jinan Fiaidhi<sup>2</sup>, Sabah Mohammed<sup>3</sup>

<sup>123</sup>*Department of Computer Science, Lakehead University, Ontario, CANADA*

<sup>1</sup>*mta@lakeheadu.ca, <sup>2</sup>jfiaidhi@lakeheadu.ca, <sup>3</sup>mohammed@lakeheadu.ca*

### Abstract

*Predictive analytics aims at building an analytical model in order to predict a target variable. This data science area currently has a lot of applications in many fields, such as in analytical customer relationship management, direct marketing, project risk management, clinical decision support systems, etc. Our research aims at performing predictive analytics on healthcare dataset to search for potentially valuable prediction models that are able to predict health-related target variables based on related input factors such as demographics, diet habit and relevant examination factors such as weight and height, etc. The healthcare data that have been used for our predictive analysis is collected from an important program conducted by the U.S. Centers for Disease Control and Prevention, which consists of 93,702 observations across 961 categories containing both interview and examination data from more than 93,000 participants. We have employed Multi-Linear Regression, Logistic Regression, Support Vector Classification, Support Vector Regression, Random Forest Classification (RFC), and Random Forest Regression algorithms to build various prediction models on the cleaned dataset. The result has shown that we have achieved good models based on the prediction of related social and healthcare factors (AUC ranging from  $\approx 0.76$  to  $\approx 0.87$ ), RFC has outperformed other classification algorithms, Fisher Score was a key feature selection algorithm, and the demographical factors have played a dominant role in the prediction of some questionnaire and laboratory target variables. Finally, based on the result of the best prediction models, we decided to develop a Hepatitis A Antibody prediction web prediction system.*

**Keywords:** NHANES, National Health and Nutrition Examination Survey, Random Forest, Support Vector Machine, Logistic Regression, Python, Machine Learning

### 1. Introduction

Predictive analytics aims at building an analytical model in order to predict a target variable [12]. Results from such systems can be very productive across many application fields, such as customer relationship management, child protection, clinical decision support systems, direct marketing, fraud detection, project risk management, etc. Thanks <sup>1</sup>to the development of information technologies and especially the revolution of machine learning algorithms, data analysis, predictive analytics and other data mining areas are now able to rapidly grab their new faces within a fairly short time period. The explosion of big data can be seen as an initial trigger to this fruitful area. This article focuses on performing predictive analytics on healthcare dataset to search for potentially valuable

---

#### Article history:

Received (March 2, 2018), Review Result (April 6, 2018), Accepted (May 3, 2018)

prediction models that are able to predict health-related target variables based on related input factors such as demographics, diet habit and relevant examination factors such as weight and height, etc. After reviewing some current healthcare datasets, we have focused on the National Health and Nutrition Examination Survey data. The National Health And Nutrition Examination Survey (NHANES) is an important program conducted by the National Center for Health Statistics of U.S. Government. The program aims at assessing health and nutrition status from different population groups across the U.S. This program started in 1960s and became a continuous program in 1999. According to the NHANES 1999-20162, this dataset is unique because it contains both interview and examination data.

We have collected NHANES data from 1999 to 2016 to prepare our analytical work. In order to download the whole dataset that belongs to the NHANES continuous program, we have come to <https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx> and download all the data that belong to each category from demographics, dietary, examination, laboratory and questionnaire. The whole dataset consists of 1,038 data files containing both interview and examination data from more than 93,000 participants. After that, we have converted all these XPT files into CSV files, cleaned and preprocessed the data. Our final dataset contains 93,702 observations across 961 categories. We have been employing this dataset to apply Multi-Linear Regression (MLR), Logistic Regression (LR), Support Vector Classification (SVC), Support Vector Regression (SVR), Random Forest Classification (RFC), and Random Forest Regression (RFR) algorithms to build various prediction models. Our research aims at predicting health-related target variables based on related input factors such as demographics, diet habit and relevant examination factors such as weight, height, etc. After our predictive analytical work, we have achieved some models that have efficiently predicted a number of target variables across some health and social factors, which include home condition, source of income from Social Security or Railroad Retirement, Hepatitis B vaccination, Hepatitis A antibody, and menstrual period regularity.

From our analytical work, it has also been shown that the RFC algorithm has outperformed both LR and SVC on the NHANES dataset, and the demographic data group has played a dominant role in predicting some target variables among questionnaire and laboratory test result factors. Based on the result of our best prediction models, we have also developed a prediction system that is able to predict the Hepatitis A antibody status in a human body based on some of the target participant's demographical factors. The remaining content of this paper is organized as follows: section 2 is a literature review; section 3 is about our environmental setup; data preprocessing and our predictive analytics will be presented and discussed through sections 4, 5 and 6; section 7 presents our Hepatitis A antibody prediction system; and finally, section 8 discusses our findings and future work.

## 2. Literature review

Researchers have employed different subsets of NHANES data, such as NHANES 2003-2010, NHANES 2003-2008, and NHANES 1999-2006. They have achieved some analytical results across some health-related categories, including food consumption, health concern, laboratory tests, and diseases.

---

<sup>2</sup> [https://www.cdc.gov/nchs/nhanes/nhanes\\_questionnaires.htm](https://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm)

There were findings that have shown the association of food consumption with some health-related issues. For example, an article from Environmental Research [2] has analyzed the trends in blood mercury concentrations and fish consumption among U.S. women of reproductive age based on NHANES data 1999-2010. They found that there was an inverse ratio between intake of mercury and fish consumption, which has been aligned with an existent trend in which people, especially pregnant women, who needed to reduce their mercury concentrations usually increase their fish consumption. There was another finding that shown the relation between food consumption and health concern. That was a publication from Nutrition Journal [3] which indicated that apple consumption could be related to a reduced risk of obesity in children. Their population was 13,339 participations that they obtained from NHANES data from 2003 to 2010. The research found that children who consumed whole apples and/or apple products were 25%-30% less likely to be obese than children who did not use any kinds of apple consumption forms.

There were researches and discoveries that tried to demonstrate the relation between laboratory tests and some diseases those have not been commonly found or previously unknown about the relation between them. For example, a research group from University of California Davis, USA [1] has attempted to find whether there were any relations between dyslipidemia, which could be observed from a few laboratory tests for some different kinds of lipidity, and psoriasis, a skin disease. However, after performing some data analysis using logistic regression model on SAS V9.3, they concluded that psoriasis was not necessarily resulted from a changing in lipid levels. In 2013, authors from Department of Electrical Engineering, Standford University discovered an association between urinary triclosan, a chemical that has been commonly found in toothpastes, soaps, and household cleaning supplies, and body mass index [7]. They have based on NHANES 2003-2008 to do their data analysis.

Another group who has also been in the above category was the authors from Boston University School of Medicine. They aimed at an estimation of the ubiquity of gout and hyperuricemia based on an analysis from more than 24,000 participants in NHANES 1988-1994 and NHANES 2007-2008 [8]. The research has found that the prevalence of gout was 3.9% among US adults in 2007-2008, while the prevalence of hyperuricemia was 21.2% among men and 21.6% among women. It has also indicated that those prevalences might have increased over the time, and they might have been associated with the increase of adiposity and hypertension. There were a few groups who have found the relations between health condition and some demographic factors. For instance, Mozumdar and Liguori have compared the prevalence in metabolic syndrome between NHANES 1988-1994 and NHANES 1999-2006 among U.S. adults of different races or ethnicities. Their population included about 6,500 participants from each of the above datasets. Finally, they found there was a persistent increase of the target value over thirteen years. They also specified that this could be a serious health concern because the problem would likely raise the probability of an increase in type 2 diabetes population [5].

Another research group [4] who came from the Division of Toxicology and Human Health Science of Agency for Toxic Substances and Disease Registry of the US also focused on the association of a health condition with a demographic factor specifically that was about the regions where people were living. They discovered that the dichlorodiphenyldichloroethylene concentration levels among people who were living in the West of the America has been persistently higher than those among people who were living in the Midwest, Northeast, and South of America. They also found a steady higher

level of perfluorinated compound concentrations and the sum of 35 polychlorinated biphenyls (PCBs) congeners in some regions compared to others. They have based on NHANES 1999 through 2004 in order to conduct their studies, and their findings has revealed some important cautions about persistent organic pollutants within the United States. In 2012, Befort and his colleagues who based on NHANES 2005-2008 also found a prevalence of obesity among both U.S. rural and urban adults [6].

There were other groups that focused on data prediction and/or improving some analytical techniques in order to enhance the data analysis. They also used NHANES data for their experiments. For example, Khanna et al. have initially employed k-means clustering in order to search for temporal dietary patterns over NHANES 1999-2004 [14]. After that, they have developed a new algorithm called Modified Dynamic Time Warping [17] that has helped improving their analysis performance. The authors from E-health Networking Conference [15] attempted to apply logistic regression model in order to predict coronary heart disease risk and obtained an AUC value of 79% on NHANES data. However, they did not mention how large the dataset they have included into their data analysis, as well as size of their training set and their test set. Some other authors have extended the traditional scatterplot matrix by visualizing clusters of a multivariate dataset by utilizing the upper portion of the matrix [16]. Their system has employed k-mean clustering and has tested on the NHANES dataset.

There were still some other groups who have employed some NHANES image data for their research analysis work [18][19]. However, this project will only focus on text data of the National Health and Nutrition Examination Survey.

From our literature review, it has been shown that most of recent research works have only focused on the trends on some categories over other categories. There were a few research groups who had been working on predictive analytics on NHANES such as the authors from E-health Networking Conference [15]. However, their employment of machine learning algorithms on NHANES data were still very limited.

### **3. Environmental setup**

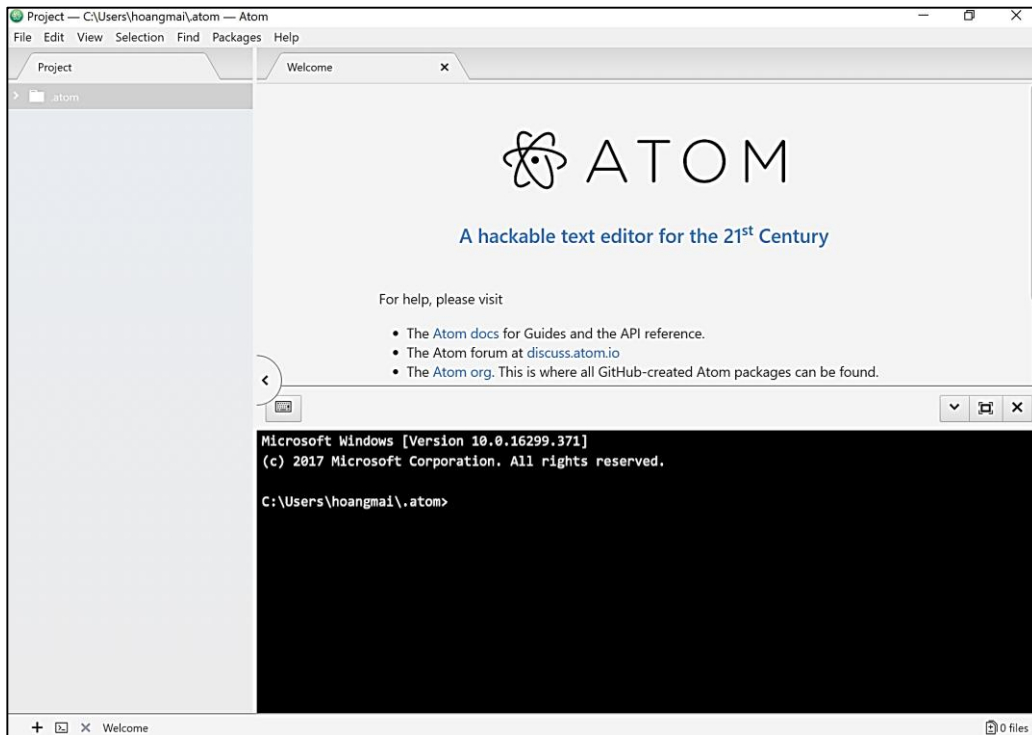
Both Python and R are currently top most significant technologies to implement both data analysis and predictive analytical projects, since they have provided rich tools and machine learning libraries that support data mining. In our project implementation, we have utilized both technologies where we have expected the best utilization from each. From our experiments, we have discovered that R is powerful for some data pre-processing phases, since they have supported very good libraries for manipulating large matrices and multiple dimensional arrays. However, we have meanwhile experienced a fairly slow execution performance of support vector machine and random forest models in R compared to those in Python. Therefore, we have decided to use Python as our main predictive analytics environment, while we still utilized some R packages for some of our data pre-processing and our feature selection.

#### **3.1. Installation of Anaconda, Spyder, R, and R Studio**

We go to <https://www.anaconda.com/download/> to download and install Anaconda, a Python distribution free and open source software<sup>3</sup>. This is a good platform to do data science work and perform machine learning algorithms, since it has provided more than 1,000 data science packages<sup>4</sup>. The software has also integrated with Spyder, an open source scientific programming IDE<sup>5</sup> which we intended to use to perform most of our analytical work. This platform has provided most import data preprocessing packages and machine learning libraries that we need to use for our predictive analytics, such as NumPy, Pandas, and sklearn. Next, we go to <http://cran.stat.sfu.ca/> to install the R language, then go to <https://www.rstudio.com/products/rstudio/download/> to download and install the R Studio IDE.

### 3.2. Set up our web application environment

We use Atom as the main environment to build our web application. This open source text editor<sup>6</sup> acts as an excellent integrated environment for editing different file types, including Python source file, html files, css files, etc. To install this environment, we go



**Figure 1. The Atom text editor**

to <https://atom.io/> and follow all the default installation steps. To conveniently interact with the command shell in our computer, we have also installed a plug-in terminal into

<sup>3</sup> Wikipedia. Anaconda (Python distribution). [https://en.wikipedia.org/wiki/Anaconda\\_\(Python\\_distribution\)](https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)). Accessed 24<sup>th</sup> April, 2018.

<sup>4</sup> Anaconda. <https://www.anaconda.com/what-is-anaconda/>. Accessed 24<sup>th</sup> April, 2018.

<sup>5</sup> Wikipedia. Spyder (software). [https://en.wikipedia.org/wiki/Spyder\\_\(software\)](https://en.wikipedia.org/wiki/Spyder_(software)). Accessed 24<sup>th</sup> April, 2018.

<sup>6</sup> Wikipedia. Atom (text editor). [https://en.wikipedia.org/wiki/Atom\\_\(text\\_editor\)](https://en.wikipedia.org/wiki/Atom_(text_editor)). Accessed 28<sup>th</sup> April, 2018.

the Atom text editor. [Figure 1] has shown how an Atom editor interface which has integrated with the built-in terminal at the bottom.

We have also created a virtual environment to manage tools, packages and libraries that we have been using during our application development.

#### **4. Data preprocessing on NHANES 1999-2016**

We used R package “foreign” to read all XPT files and converted them into CSV files. Next, we concatenated all the data that belong to a single category from multiple data files. This step is necessary because each original data file contained data within only one year of survey. Our next step is to remove any category that has the total number of observations which is less than 20% of the total number of survey participants from 1999 to 2016. We have a total of 93,702 participants across the years. Therefore, we have removed any column that contains less than 18,740 rows. Finally, we have merged all data files that belong to one of the original component datasets, i.e. demographic, diet, examination, lab, and questionnaire, into a single data file. We also created a single dataset by merging these 5 component datasets. This final dataset includes 93,702 observations and 961 categories.

We performed missing data treatment by using module Imputer from the sklearn preprocessing package. We have been replacing our missing continuous values by the mean or median of the column, and the missing categorical values with the most frequent value that appears in the column. Our experiments have shown that the ‘mean’ strategy has outperformed the ‘median’ strategy in our data. For feature scaling and categorical data encoding, we employed modules StandardScaler and OneHotEncoder from the sklearn preprocessing library respectively.

#### **5. Performing predictive analytics on NHANES 1999-2016**

We first performed LR and MLR algorithms on our categorical and continuous targets variables respectively. These baseline models act as early filters that allow us to determine whether we should further apply support vector and random forest models to a certain prediction analytic that involved with a certain input and output data. This was because our experiments have shown that both LR and MLR have performed as fairly competitive models compared to SVMs and RFs, while their execution time is much shorter. Our strategy is then early discarding any prediction that performed too poor on our LR or MLR models.

Based on the original organization of NHANES data, we have organized the final dataset into the following components: demographics, diet, questionnaire, basic examination, oral examination, visual examination, audio examination, and laboratory test results. Since most of them mostly contain health-related factors, we have used demographical and basic examination factors as input variables to predict the other component datasets. The relations among health component datasets have been also explored.

[Table 1] has shown the performance on some best questionnaire target variables while we applied LR algorithm on demographics, diet and basic examination input data. The questionnaire group contains interview data that support all other datasets. Therefore, besides health-related information that this dataset mostly contains, there are also a few socioeconomic factors that support the demographic dataset, such as source of income, home condition, etc. That is why we have also obtained the prediction results on these

targets why applying our prediction models on the questionnaire target dataset. The target variable codes and variable labels at the first and second columns on [Table 1] are taken from NHANES codebooks, while variable descriptions describe the variable categories based on their domain. We have also achieved a few good models while performing LR on laboratory tests, including LBXHA (Hepatitis A antibody - LR AUC $\approx$ 0.716) and URXPREG (Pregnancy test result - LR AUC $\approx$ 0.696).

**Table 1. Logistic regression AUCs on some questionnaire target variables**

Target variable	Variable label	Variable description	AUC
ALQ100	Had at least 12 alcohol drinks/1 yr	(yes, no)	0.627140475
AUQ130	General condition of hearing	(good: yes, not good:no)	0.685391902
BPQ020	Ever told you had high blood pressure	(yes, no)	0.70565387
HOQ065	Home owned, bought, rented, other	(home owned or being bought: yes, others:no)	0.728681183
IMQ020	Received Hepatitis B 3 dose series	(at least 3 doses: yes, less than 3 doses:no)	0.755735349
INQ020	Income from wages/salaries	(yes, no)	0.700995525
INQ030	Income from Social Security or Railroad Retirement	(yes, no)	0.785215822
RHQ030	Had regular periods in past 12 months	(yes, no)	0.850529948

**Table 2. LR, SVC, and RFC performance on some target variables**

Target variable	Variable label	Input variable groups	$\approx$ AUC (LR)	$\approx$ AUC (SVC)	$\approx$ AUC (RFC)
HOQ065	Home owned or being bought	demographics, diet, and basic examination data	0.73	0.73	0.77
IMQ020	Received Hepatitis B 3 doses series	demographics, diet, and basic examination data	0.76	0.75	0.77
INQ030	Income from Social Security or Railroad Retirement	demographics, diet, and basic examination data	0.79	0.76	0.8
RHQ030	Had regular periods in the past 12 months?	demographics, diet, and basic examination data	0.85	0.85	0.87

LBXHA	Had Hepatitis A antibody	demographics, diet, basic examination, and questionnaire data	0.72	0.74	0.76
-------	--------------------------	---	------	------	------

After sorting out good prediction models by applying LR and MLR, we found that there are good classification models as shown above. However, there has been no promising regression model that has been observed. Therefore, we only applied SVC and RFC algorithms on these top targets. Our results have been shown on [Table 2]. We also run Fisher Score algorithm to choose a few best inputs among hundreds of input factors, and obtained the results as shown on [Table 3]. We have modified the RFC models shown on [Table 2] in order to include only those input variables to observe the performance of those models on a limited set of inputs.[Table 3] also shows the results of this phase.

**Table 3. Random forest classification performance for top significant inputs**

Target variable	Top significant inputs	Top significant input labels	≈AUC (RFC)
HOQ065	RIDAGEYR RIDRETH1 DMDHHSIZ INDFMPIR DMDHRGND DMDHRAGE DMDHREDU DMDHRMAR	age, race, total number of people in the household (HH), ratio of family income to poverty threshold, household reference person's (HHR) gender, HHR age, HHR education level, HHR marital status	0.85
IMQ020	RIAGENDR RIDAGEYR RIDRETH1 INDFMPIR DMDHRGND DMDHRAGE DMDHREDU	gender, age, race, ratio of family income to poverty threshold, HHR gender, HHR age, HHR education level	0.75
INQ030	RIDAGEYR RIDRETH1 INDFMPIR DMDHRGND DMDHRAGE DMDHHSIZ	age, race, family income to poverty threshold, HHR gender, HHR age, total number of people in the HH	0.83
RHQ030	RIDAGEYR DMDHRAGE INDFMPIR DMDHHSIZ	age, HHR age, family income to poverty threshold, total number of people in the HH	0.84



LBXHA	RIDAGEYR RIDRETH1 DMDHHSIZ INDFMPIR DMDHRAGE DMDHREDU	age, race, total number of people in the HH, family income to poverty threshold, HHR age, HHR education level	0.7
-------	--	---	-----

## 6. Discussion

We have plotted the performance of LR, SVC and RFC to compare with each other based on the five target variables from [Table 2]. From [Figure 2], we can see that the performance of RFC is steadily higher than both LR and SVC, while the two remaining algorithms share their positions.

In [Figure 3], we compare the performance of RFC on those targets between 2 different set of inputs: one is the full set and the other contains only the top input variables that have been shown in [Table 3]. From the graph, we can see that the performance of both kinds of input set are comparable to each other. This demonstrates that the Fisher Score algorithm has played an important role for feature selection on the NHANES dataset. Additionally, [Table 3] has shown that most of the most significant inputs belong to the demographic factors.

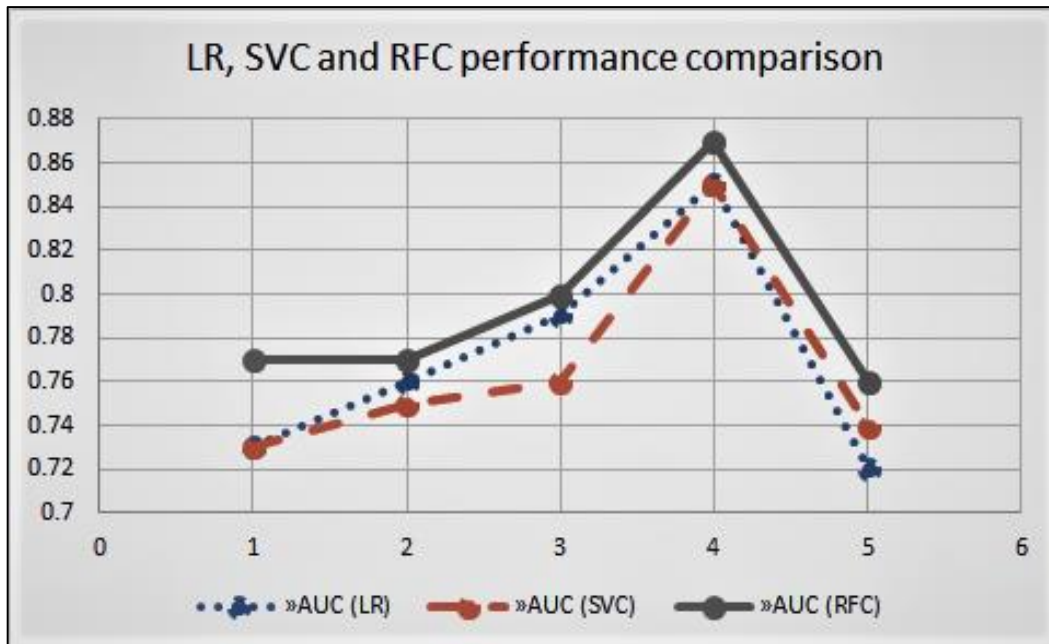


Figure 2. LR, SVC and RFC performance comparison

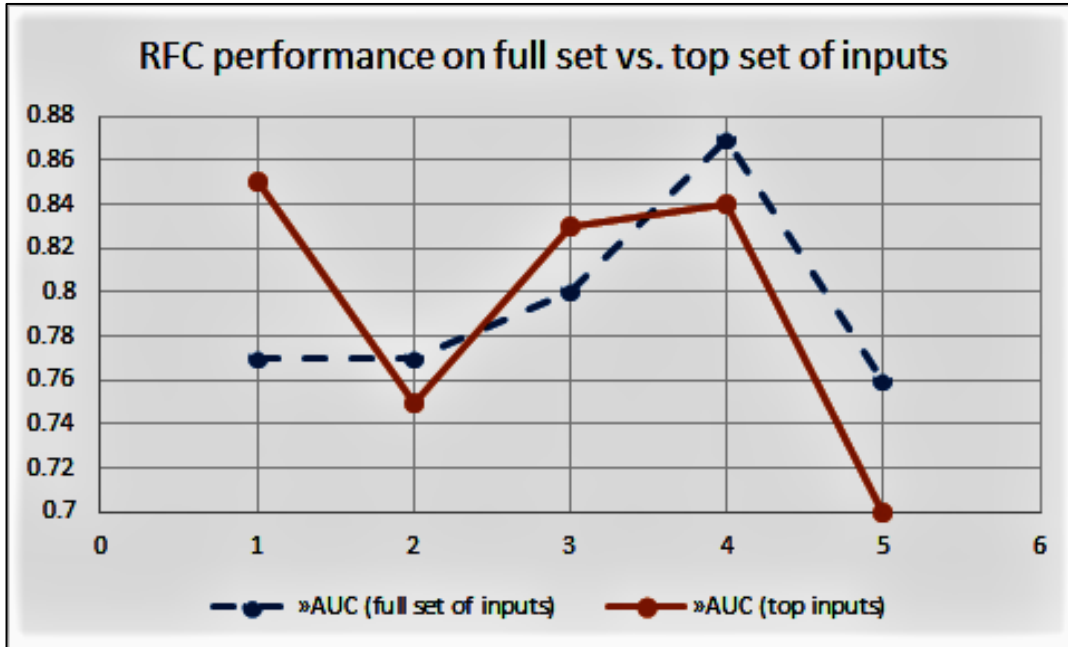


Figure 3. RFC performance based on full set vs. top set of input variables

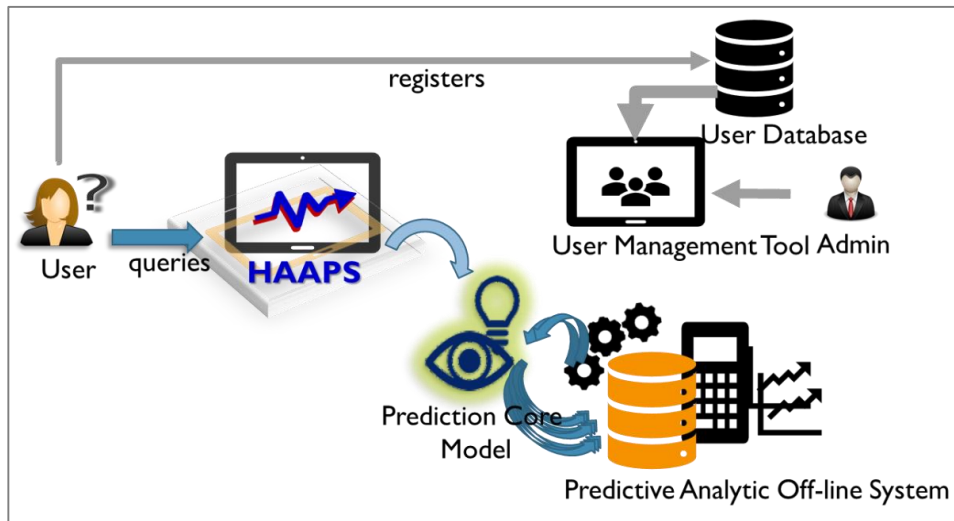
## 7. The Hepatitis A Antibody Prediction System

Based on our best results that have been presented in Section 5, we then applied the RFC model that predicted LBXHA shown in [Table 3] to build a Hepatitis A (HPA) Antibody Prediction System (HAAPS).

From one of our statistics on NHANES data, we have found that there were 22,751 participants from 2-year and older who are negative to their HPA antibody test. This accumulates  $\approx 47.68\%$  to the total of 47,715 participants from 2 years old at the time they performed their HPA antibody test. This statistical value has demonstrated that such Hepatitis A antibody prediction system is valuable to help predicting the status of HPA from a large amount of population. Since HPA virus can be transmitted through contaminated food, water intake or person-person direct contact<sup>7</sup>, it is recommended that all children “at age 1 year” should take HPA vaccination according to the Canadian Centers for Disease Control And Prevention.

<sup>7</sup> World Health Organization. Media Centre. Hepatitis A. <http://www.who.int/mediacentre/factsheets/fs328/en/>. Access 12 April 2018.

[Figure 4] has shown an overview of our system.



**Figure 2. The Hepatitis A Antibody Prediction System**

From the diagram, we can see that there are four main components in our system. They are:

**User Registration:** this component allows users to register to our HAAPS using their username, email and password. We used the Django Framework<sup>8</sup> in order to help fully protect our user information.

The screenshot shows the main page of the HAAPS. At the top, there are navigation links: HOME, Admin, Register, and Logout. The main heading is 'THE HEPATITIS A ANTIBODY PREDICTION SYSTEM'. Below the heading, there are several input fields and dropdown menus for user information: 'Age: 50', 'Race: Mexican American', 'Total number of people in household: 2', 'Family income to poverty threshold: 1.5', 'Household reference person age: 40', and 'Household reference person education level: College Graduate or above'. A 'Check>>' button is located below the input fields. At the bottom of the page, there is a message: 'You are likely positive with Hepatitis A antibody. This means that you were vaccinated, or you were once exposed to Hepatitis A Virus some time in the past.'

**Figure 5. The HAAPS main page**

**User Management:** is where our admin can manage user accounts.

**User Querying:** this main component allows user to query our system in order to obtain an HPA antibody status. The user first needs to enter some information such as their age,

<sup>8</sup> Django. <https://www.djangoproject.com/>. Accessed 24<sup>th</sup> April, 2018.

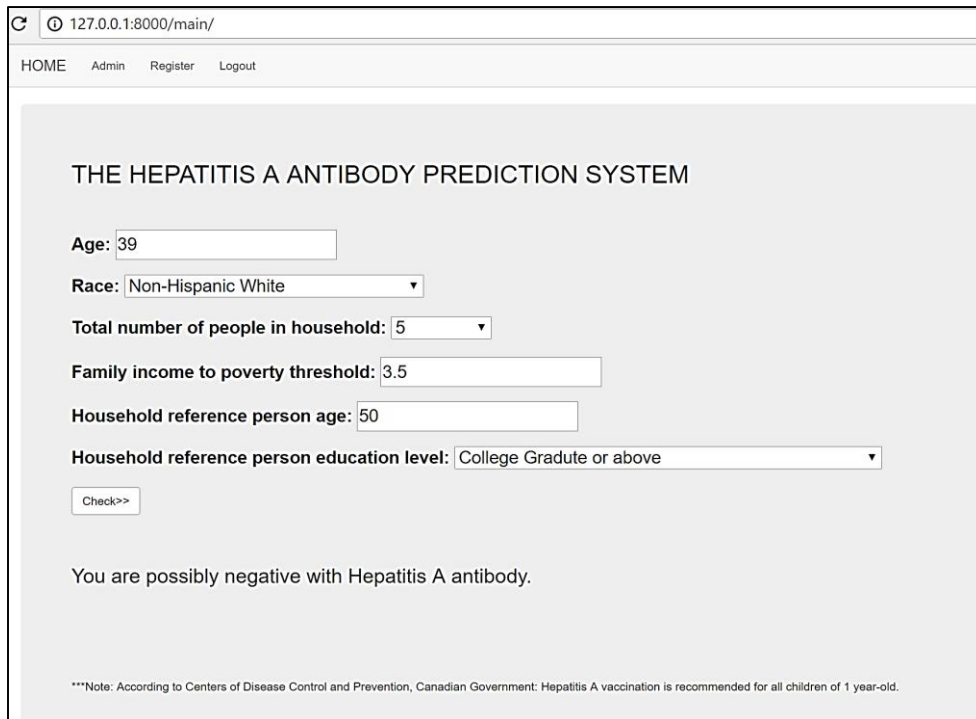
race, the total number of people in their household, etc according to the results that we have shown in [Table 3] then will get the prediction result.

**Prediction Core Model:** the prediction model is fetched from our off-line training system. The training system will then be responsible for analyzing, maintaining and improving our best model for HAAPS.

[Figure 5] has shown the HAAPS main page. In this case, the system has predicted that the user is positive to HPA antibody. An alternative case has shown in [Figure 6], where the prediction result is negative. Our system also pops up a note that shows a recommendation from the Canadian Centers of Disease Control and Prevention in this case.

## 8. Conclusion and future work

We have been employing LR, MLR, SVC, SVR, RFC, and RFR algorithms to build a lot of classification models and regression models on the NHANES dataset. The results have demonstrated a good contribution of NHANES dataset on the predictive analytical area. From our analytical work, we have achieved some good models for the prediction of some health and social factors, including home condition, source of income from Social Security or Railroad Retirement, Hepatitis B vaccination, Hepatitis A antibody, and menstrual period regularity. We have also found out that RFC prediction models have persistently outperformed LR's and SVC's. Additionally, Fisher Score algorithm has efficiently performed for our feature selections. This algorithm has helped our choices for top input factors usually comparable to full set of inputs. Another important finding



The screenshot shows a web browser window with the URL 127.0.0.1:8000/main/. The page title is "THE HEPATITIS A ANTIBODY PREDICTION SYSTEM". The navigation menu includes "HOME", "Admin", "Register", and "Logout". The main content area contains a form with the following fields and values:

- Age: 39
- Race: Non-Hispanic White
- Total number of people in household: 5
- Family income to poverty threshold: 3.5
- Household reference person age: 50
- Household reference person education level: College Graduate or above

A "Check>>" button is located below the form. Below the form, the prediction result is displayed: "You are possibly negative with Hepatitis A antibody." At the bottom of the page, a note reads: "\*\*\*Note: According to Centers of Disease Control and Prevention, Canadian Government: Hepatitis A vaccination is recommended for all children of 1 year-old."

**Figure 6. A negative prediction case**

is that demographical factors have played a dominant role in the prediction of some questionnaire and laboratory target outputs compared to other set of inputs.

Our regression models which are based on MLR, SVR and RFR, however, did not work well. We might need to include more similar data in the future to improve those results.

## Acknowledgements

This paper is part of the first author MSc Project.

## References

- [1] Ma C1, Schupp CW, Armstrong EJ, Armstrong AW. Psoriasis and dyslipidemia: a population-based study analyzing the National Health and Nutrition Examination Survey (NHANES). *Journal of the European Academy of Dermatology and Venereology*. doi: 10.1111/jdv.12232 (2014)
- [2] Birch RJ, Bigler J, Rogers JW, Zhuang Y, Clickner RP. Trends in blood mercury concentrations and fish consumption among U.S. women of reproductive age, NHANES, 1999-2010. *Environmental Research*. doi: 10.1016/j.envres.2014.02.001 (2014)
- [3] O'Neil CE, Nicklas TA, Fulgoni VL. Consumption of apples is associated with a better diet quality and reduced risk of obesity in children: National Health and Nutrition Examination Survey (NHANES) 2003-2010. *Nutrition Journal*. doi:10.1186/s12937-015-0040-1 (2015)
- [4] Wattigney WA, Irvin-Barnwell E, Pavuk M, Ragin-Wilson A. Regional Variation in Human Exposure to Persistent Organic Pollutants in the United States, NHANES. *J Environ Public Health*. doi: 10.1155/2015/571839 (2015)
- [5] Mozumdar A, Liguori G. Persistent increase of prevalence of metabolic syndrome among U.S. adults: NHANES III to NHANES 1999-2006. *Diabetes Care*. doi: 10.2337/dc10-0879 (2011)
- [6] Befort CA, Nazir N, Perri MG. Prevalence of obesity among adults from rural and urban areas of the United States: findings from NHANES (2005-2008). *J Rural Health*. doi: 10.1111/j.1748-0361.2012.00411.x (2012)
- [7] Lankester J, Patel C, Cullen MR, Ley C, Parsonnet J. Urinary triclosan is associated with elevated body mass index in NHANES. *PLoS One*. doi: 10.1371/journal.pone.0080057 (2013)
- [8] Zhu Y, Pandya BJ, Choi HK. Prevalence of gout and hyperuricemia in the US general population: the National Health and Nutrition Examination Survey 2007-2008. *Arthritis Rheum*. doi: 10.1002/art.30520 (2011)
- [9] Breiman L. Random Forests. *Machine Learning* 45, no. 1: 5– 32. (2001)
- [10] Aizerman, Mark A., Braverman, Emmanuel M. & Rozonoer, Lev I. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*. 25: 821–837. (1964)
- [11] Lara J. Estimating the costs of achieving the WHO–UNICEF Global Immunization Vision and Strategy, 2006–2015. *Bulletin of the World Health Organization*, January 2008, 86 (1). doi: 10.2471/BLT.07.045096. (2008)
- [12] Bart Baesens. *Analytics in a big data world. The essential guide to data science and its applications*. Published by John Wiley & Sons, Inc., Hoboken, New Jersey (2014)
- [13] Schinazi, Rinaldo B. Multiple Linear Regression. *Handbook of Psychology*. John Wiley & Sons, Inc., 2012:364-368 (2012) [4]J. Kimura and H. Shibasaki, Editors. *Recent Advances in Clinical Neurophysiology. Proceedings of the 10th International Congress of EMG and Clinical Neurophysiology, (1995) October 15-19; Kyoto, Japan*
- [14] N. Khanna, H. A. Eicher-Miller, H. K. Verma, C. J. Boushey, S. B. Gelfand and E. J. Delp, "Modified dynamic time warping (MDTW) for estimating temporal dietary patterns," 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, (2017) pp. 948-952. doi: 10.1109/GlobalSIP.2017.8309100
- [15] A. Mohawish, R. Rathi, V. Abhishek, T. Lauritzen and R. Padman, "Predicting Coronary Heart Disease risk using health risk assessment data," 2015 17th International Conference on E-health Networking, Application & Services (HealthCom), Boston, MA, (2015) pp. 91-96. doi: 10.1109/HealthCom.2015.7454479

- [16] S. O. Torres, H. Eicher-Miller, C. Boushey, D. Ebert and R. Maciejewski, "Applied Visual Analytics for Exploring the National Health and Nutrition Examination Survey," 2012 45th Hawaii International Conference on System Sciences, Maui, HI, **(2012)** pp. 1855-1863. doi: 10.1109/HICSS.2012.116
- [17] N. Khanna, H. A. Eicher-Miller, C. J. Boushey, S. B. Gelfand and E. J. Delp, "Temporal Dietary Patterns Using Kernel k-Means Clustering," 2011 IEEE International Symposium on Multimedia, Dana Point CA, **(2011)** pp. 375-380. doi: 10.1109/ISM.2011.68
- [18] L. R. Long and G. R. Thoma, "Computer assisted retrieval of biomedical image features from spine X-rays: progress and prospects," Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001, Bethesda, MD, **(2001)** pp. 46-50. doi: 10.1109/CBMS.2001.941696
- [19] Xiaoqian Xu, D. J. Lee, S. Antani and L. R. Long, "Pre-Indexing for Fast Partial Shape Matching of Vertebrae Images," 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06), Salt Lake City, UT, **(2006)** pp. 105-110. doi: 10.1109/CBMS.2006.129