**Machine Learning**

**IT4060**

A picture containing text, clipart, vector graphics

Description automatically generated

ASSIGNMENT 2

|  |  |
| --- | --- |
| **STUDENT ID** | **NAME** |
| IT19125244 | Rogin Harshini N. V. |
| IT19074450 | Manilka G. S. |
| IT19364216 | P. J. Samuel |
| IT19038292 | Wickramasinghe A. Y. S. W. |

**Table** Of **Content**

### Introduction

1. Problem Statement
2. Methodology
   1. Dataset
   2. Pre-processing
      1. Selection Of Training Data and Test Data
      2. Data Analysis and Feature Selection
      3. Data Cleaning
      4. Data Normalization
3. Modelling Methods
4. Discussio**n**
5. Work Distribution
6. References
7. Appendix
8. **Introduction**

Machine Learning is an analysis of data and the use of software techniques to find patterns and regularities in sets of data. Machine Learning can deal with numerous tasks such as classification, regression, clustering, anomaly detection, etc. Machine Learning is being used in many more applications such as credit assessment, stock market prediction, fault diagnosis production system, medical discovery and fraud detection etc.

In recent times the sector is acquainted with the phrases pandemic, quarantine and waves etc. Due to the COVID-19 fashion global. A new virus referred to as influenza A (H1N1) advanced within the spring season of 2009. The virus first detected in the USA and immediately spread everywhere in the USA . And unfold to all over the world. This new H1N1 virus had a rare mixture of influenza genes that had in no way been visible before in animals or humans. Among 2009 and 2010, this virus lasted 19 months and killed over 3 lakh human beings. Man or woman immunization is provided through vaccines, and enough immunization in a community can in addition prevent disease spread via herd immunity. So, inside the occasion of an epidemic, vaccination is vital to saving the planet for the greater accurate.

# Problem Statement

A worldwide pandemic resulting from the H1N1 influenza virus has wiped out the world reportedly almost three lakhs deaths. In October 2009, a vaccine towards the H1N1 flu virus became widely available. The national 2009 H1N1 Flu Survey become conducted within the united states in past due 2009 and early 2010. This smartphone survey asked respondents in the event that they had received the H1N1 and seasonal flu vaccines, in addition to non-public questions. These extra questions targeted on their social, economic, and demographic backgrounds, as well as their views at the risks of contamination and vaccine effectiveness, in addition to their attitudes toward mitigating transmission. A higher understanding of how those characteristics are associated with personal vaccination styles can assist guide destiny public health efforts.

The goal is to forecast how probably human beings are to acquire H1N1 and seasonal flu vaccines. Particularly, Chances ought to be predicted: one for h1n1 vaccine and one for seasonal vaccine. The area beneath the receiver running characteristic curve (ROC AUC) for each of the 2 target variables can be used to assess performance. The general score may be the imply of those two ratings. A better price denotes better overall performance

# Methodology

## Dataset

The dataset used is from the National H1N1 Flu Survey.

* + - Training features
    - Training labels
    - Testing features

|  |  |
| --- | --- |
| **Feature’s name** | **Meanings of the features** |
| h1n1Concern | Level of concern about the H1N1 flu.   * 0 - Not at all * 1 - Not very * 2 - Somewhat * 3 - More. |
| h1n1Knowledge | Knowledge level about the H1N1 flu.  0 - no  1 - a little  2 - a lot |
| behavioralAntiviralMeds | Antiviral medications have been taken. |
| behavioralAvoidance | who are suffering from flu symptoms avoided to have the contacts with others or neglect. |
| behavioralFaceMask | Buy a mask to cover the face |
| behavioralWashHands | Wash and sanitize the hand |
| behavioralLargeGatherings | Avoided the large gatherings |
| behavioralOutsideHome | had reduced contact with people outside of their immediate family. |
| behavioralTouchFace | Had reduced contact with people outside of their immediate family. |
| doctorRecH1n1 | A doctor recommended the H1N1 flu vaccine. |
| doctorRecSeasonal | The doctor suggested a seasonal flu vaccine. |

|  |  |
| --- | --- |
| chronicMedCondition | having diseases one or more,   1. Asthma /lungs diseases. 2. Diabetes. 3. Heart related problems 4. Kidney problems. 5. sickle cell anemia/another anemia. 6. Neurological/neuromuscular problems. 7. Liver Problems . 8. Weak immune system caused by a chronic illness/medicines used. |
| childUnder\_6\_months | Having a child under 6 months |
| healthWorker | healthcare worker. |
| healthInsurance | Person with health insurance. |
| opinionH1n1VaccEffective | The effectiveness of the H1N1 vaccine, according to respondents.   * 1 - Not at all effective * 2 - Not very effective * 3 - Unsure * 4 - Somewhat effective * 5 - Extremely effective |

|  |  |
| --- | --- |
| opinionH1n1Risk | People’s views about the risk of contracting H1N1 flu in the absence of a vaccinum.   * 1 - Very Low * 2 - Slightly Low * 3 - Don't Know * 4 - Slightly High * 5 - Extremely high |
| opinionH1n1SickFromVacc | Respondents' fear of becoming ill after receiving the vaccination for the h1n1.   * 1 - Not at all concerned * 2 - Not very concerned * 3 - Don't know * 4 - Somewhat concerned * 5 - Very concerned |

|  |  |
| --- | --- |
| opinionSeasVaccEffective | People thoughts about the effectiveness of the seasonal flu vaccination.   * 1 – very low * 2 - Not much * 3 - Don't know * 4 – Somewhat high * 5 – very high |
| opinionSeasRisk | People thoughts about risk of getting sick with seasonal flu without a vaccine.   * 1 –Very Lowest risk * 2 – Lowest risk * 3 - Don't know * 4 – High risk * 5 –More Risk. |
| opinionSeasSickFromVacc | People scare about getting sick after the seasonal flu vaccination.   * 1 – Very low level of worry * 2 – Low level * 3 - Don't know * 4 – Highly worrying * 5 – High level of worrying |
| ageGroup | Age of the person |
| education | Educational qualification |
| race | Race |
| sex | gender |
| incomePoverty | Annual income |
| maritalStatus | Civil status of the employee |
| rentOrOwn | House category |
| employmentStatus | Status of the employment details |
| hhsGeoRegion | The people’s address. |

|  |  |
| --- | --- |
| censusMsa | Residence within metropolitan statistical areas (MSAs) as defined by the United States Census. |
| houseHoldAdults | Top code for adult amount in the home is 3. |
| houseHoldChildren | Top-coded to 3 for children amount in the home. |
| employmentIndustry | The industry in which the respondent works. Short random character strings are used to represent values. |
| employmentOccupation | Respondent's occupation type. Short random character strings are used to represent values. |

## Pre-processing

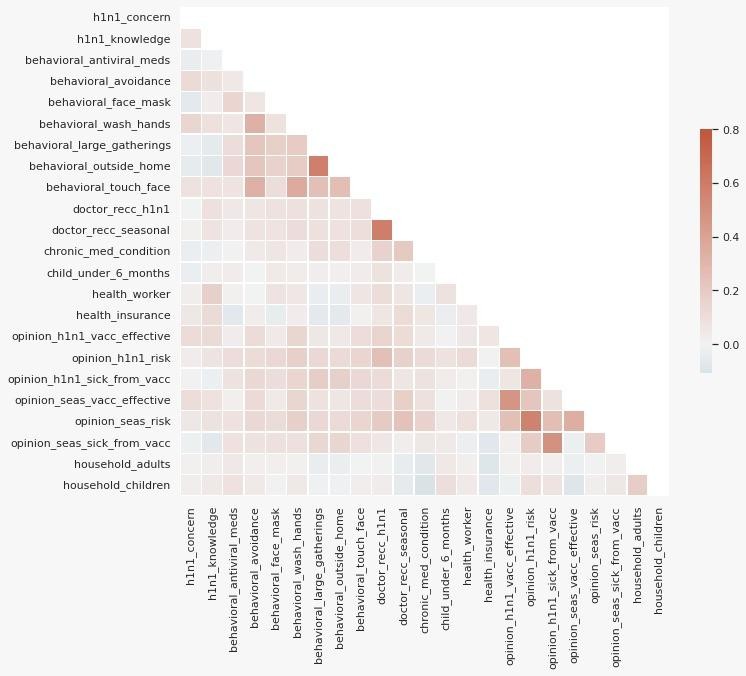
Data preprocessing is a data mining technique used to convert raw data into a usable and efficient format. Cleansing, transformation, and feature engineering are all part of data preprocessing.

### Selection of Training and Test Data

There are 26,707 rows in the training data. Three files were provided with the necessary information. 30% of the data was used as test data. After testing the accuracy, we train the model with the whole dataset without splitting.

### Data Analysis and Feature Selection

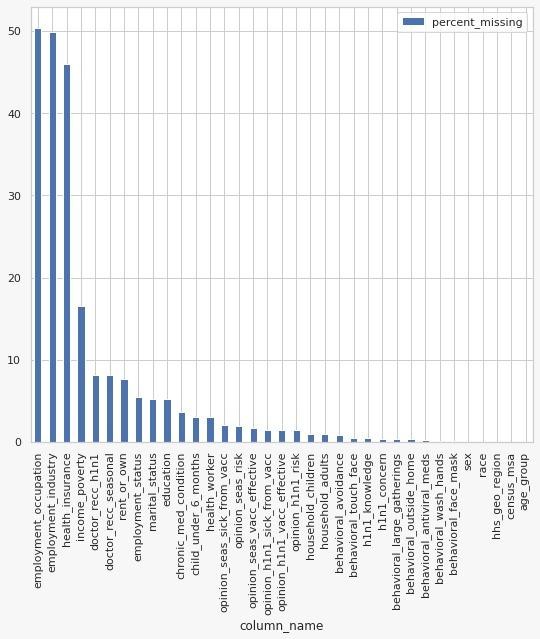
Although a dataset contains many features, it can include irrelevant information. The tracked metrics are collectively referred to as a feature analysis to identify features that are connected to the target. According to the feature analysis carried out, the correct features are selected for the considered label.



The correlation matrix above shows the correlation between characteristics and how they vary with each other. Behavioral\_large\_gatherings and behavioral\_outside\_home are related to each other, in the same way doctor\_recc\_h1n1 and Doctor\_recc\_seasonal are related. Opinion\_h1n1\_vacc\_effective, opinion\_h1n1\_risk, opinion\_h1n1\_sick\_from\_vacc are respectively correlated with opinion\_seas\_vacc\_effective, Opinion\_seas\_risk, Opinion\_seas\_sick\_from\_vacc. The other variables are not correlated with each other.

### Data Cleaning

Data cleaning could be a method of getting ready data for analysis by removing dangerous data, organizing the data, and filling in the null values is known as data cleaning. Finally, cleaning data prepares the data for the process of data mining. It extracts from the dataset most valuable information. Give dataset, the health\_insurance, employment\_industry and employment\_occupation data samples have a high percentage of missing values among the 26,707 instances and 35 characteristics. The following table shows missing values percentage for each of the given characteristics.



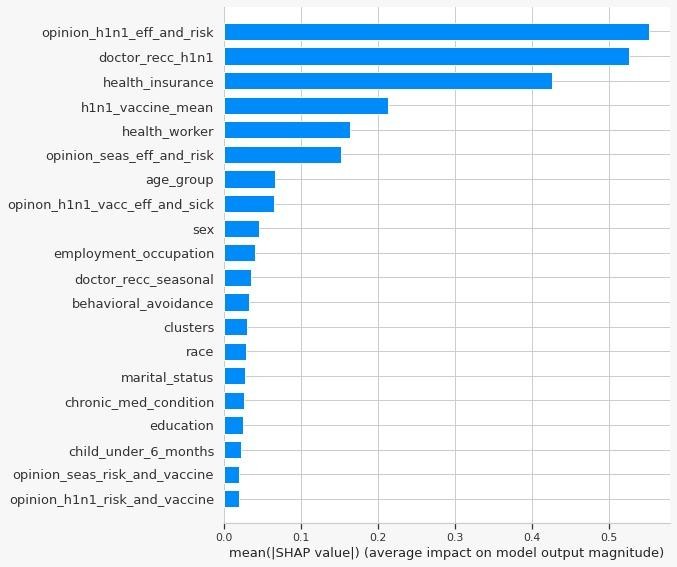
We have used different strategies to eliminate these factors from time to time. The strategies are,

* + - Based on average value
      * Here the missing values are filled with the average value of the entire column
    - Based on median value
      * This can work in some cases, here the missing values are filled with the median of the specific column.
    - Consider as a new category
      * Fill the missing categorical value to -999.
    - Based on cluster mean value
      * Here we group data into different clusters and take the mean of the cluster and fill that value.

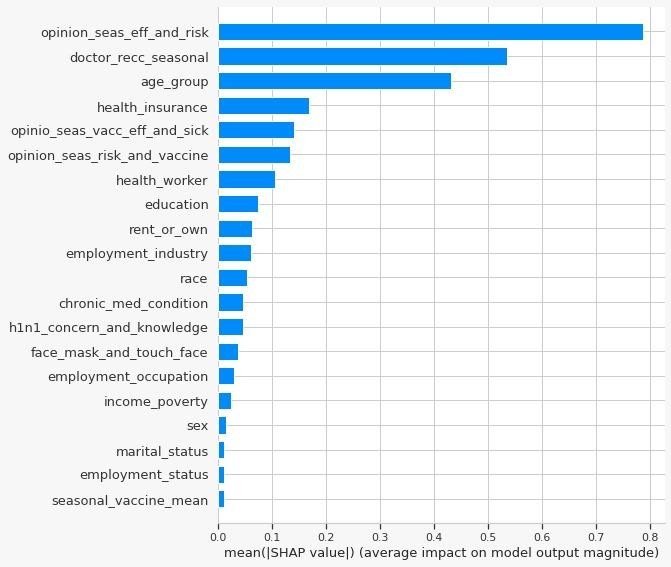
The best strategy we have found as a result is based on a combination of two approaches. There are other methods, such as drop columns which have more missing values. Because we miss the data patterns. But, we haven't tried this approach.

Some of the fields with numeric values have been filled in with cluster mean values. Empty fields with categorical values were filled with some negative values.

We also plotted the SHAP values of the Cat Boost classifier for the two labels to find the features important for the prediction.



The above plot for Seasonal vaccine indicates that the most important positive correlation comes from opinion\_h1n1\_eff\_and\_risk, doctor\_recc\_h1n1, health\_insurance, h1n1\_vaccine\_mean and health\_worker.



The above plot for the h1n1 vaccine indicates that the most important positive correlation comes from opinion\_H1n1-Eff-and-Risk, doctor\_recc\_seasonal, health-insurance, Age-group and health\_insurance.



It appears that about 50% of people have infected with the seasonal flu shot, about 20% of people have infected with the H1N1 flu shot. In terms of class balance, the seasonal vaccinum target is alleged to possess balanced categories, however the H1N1 vaccinum target has moderately unbalanced categories.

### Data Normalization

It is a process of scaling attributes of data so that it falls within a narrower limit range, the limits within -1.0 to 1.0 Or 0.0 to 1.0. It is commonly used method for classifications. Normalization is needed in which coping with attributes on different scales; otherwise, the effectiveness of an equally critical characteristic (on a lower scale) may be diluted due to other attributes having values on a larger scale.

Virtually placed, when a couple of attributes exist but their values are on unique scales, this may bring about negative records models while appearing records mining operations. As a end result, they may be normalized to position all of the attributes on the identical scale. We tried to normalize the statistics the usage of special normalization strategies. Normalization of Z-rankings and min-max normalization

.

# Modelling Methods

Random Forest Classifier

* + - Model Parameters
      * nEstimators- 1000
    - Score: 0.8081

Logistic Regression

* + - Model Parameters
      * C - 1.0,
      * interceptscaling - 1
      * max\_iter - 100
      * penalty – ‘l2’
      * Total - 0.0001
    - Score : 0.8308

Catboost classifier

* + - Model Parameters
      * Iterations - 45000
      * reg\_lambda - 30
      * custom\_loss - 'logloss'
    - Score : 0.8632

decision tree classifier

* + - Model Parameters
      * Criterion - gini
      * MaxDepth - 6
      * MinSampleSplit - 2
    - Score : 0.8392

# Discussion

As a team we have tried different methodologies to predict the results regarding the prediction of H1N1 and Seasonal Flu Vaccines and we got a best score 0.8632 and rank 49 out of 2692 on 27th of May 2022.

By doing this assignment we were able to practice the theories covered in the ML module. Addionally we are lerning data cleaning, transformation, models and their fine tunings and the pros and cons of using those models after analysing the data set.

1. **Workload distribution**

|  |  |  |
| --- | --- | --- |
| **IT Number** | **Name** | **Work contribution** |
| IT19125244 | Rogin Harshini N. V. | Data preprocessing and building model |
| IT19038292 | Wickramasinghe A. Y. S. W. | Importing libraries and observing the data set |
| IT19364216 | P. J. Samuel | Testing the performance of model and improving the model |
| IT19074450 | Manilka G. S. | Analysis of the data |

# 7. References

1. Kaggle.com. 2022. *H1N1 and Seasonal Flu Vaccines*. [online] Available at: <https://www.kaggle.com/datasets/arashnic/flu-data?resource=download> [Accessed 29 May 2022].
2. 2022. [online] Available at: <https://www.cdc.gov/flu/pandemic-resources/2009-h1n1-pandemic.html> [Accessed 29 May 2022].
3. DrivenData. 2022. *Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines*. [online] Available at: <https://www.drivendata.org/competitions/66/flu-shot-learning/> [Accessed 29 May 2022].
4. Seaborn.pydata.org. 2022. *Plotting a diagonal correlation matrix — seaborn 0.11.2 documentation*. [online] Available at: <https://seaborn.pydata.org/examples/many\_pairwise\_correlations.html> [Accessed 29 May 2022]..
5. GeeksforGeeks. 2022. *Data Preprocessing in Data Mining - GeeksforGeeks*. [online] Available at: <https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/> [Accessed 29 May 2022].
6. Trifacta. 2022. *Data Cleaning in Data Mining: Evaluating Data | Trifacta*. [online] Available at: <https://www.trifacta.com/data-cleaning-in-data-mining/> [Accessed 29 May 2022].
7. GeeksforGeeks. 2022. *Data Normalization in Data Mining - GeeksforGeeks*. [online] Available at: <https://www.geeksforgeeks.org/data-normalization-in-data-mining/> [Accessed 29 May 2022].
8. Appendix

**Git Link**

https://github.com/YasiruWickramasinghe/ML\_Assignment2\_IT19125244\_IT19074450\_IT19364216\_IT19038292

**Video Link**

https://drive.google.com/drive/folders/1YSO2XzFWMVyCZVP6NehDSzSsRrftDNWU?usp=sharing