A Moblie App collects and displays Daily COVID-19 and Flu Cases

Student: Xiangyang Xu   
*Department of Electrical and Computer Engineering*  
*Stevens Institute of Technology*Hoboken, USA  
[xxu46@stevens.edu](mailto:xxu46@stevens.edu)

Advisor: Professor Shucheng Yu

*Department of Electrical and Computer Engineering*

*Stevens Institute of Technology*Hoboken, USA  
syu19@stevens.ed

*Abstract*— Machine learning is widely used to do prediction and give suggestion. When enough data is available, supervised learning is a good way to build accuracy models. Covid-19 is one of the hottest topics in this year and flu is also a problem confused people for many years. This project is to develop an android app to predict the trend of flu and covid-19, also can give people suggestion whether they need inject flu vaccine or not. All trend prediction and flu vaccine suggestion based on their assemble model. In this project, KNN, decision tree and Multi-layer perceptron are mainly used to train models.

Keywords— Covid-19, Flu, Android, Machine Learning, Deep learning

# Problem Introduction, Challenges and Related Work

## Problem Introduction

From Dec.2019, the covid-19 spread from Wuhan to the world. Millions of people suffered from this new virus. At the same time, the economy of many regions stopped growing.

Machine learning and deep learning tech are widely used to help people to fight against this kind of new virus. For example, L. Mertz and his team developed AI-driven tools to quantify Lung Images.[1] At the same time, in people's daily life, many apps are developed for keeping people away from the Covid-19 virus. J. Berglund and his team developed software to track the virus.[2]

The epidemic has not been controlled so far and winter has come. Therefore, anticipation and early warning of Covid-19 are necessary. Use an app to help people away from Covid-19 seems necessary to work. If we can predict the possible outbreak of the epidemic, then forecast it to the public.

Currently, people use machine learning methods to predict the trend in the price of goods. Based on the history price, it is feasible to predict future prices. A. Yousefi and his team have used this tech to predict long-term electricity prices.[3] Therefore, it is also feasible to use machine learning to predict the future trend of covid-19 as well.

This technology can be also used to forecast flu. At the same time, machine learning models can be used to predict whether a person needs a flu vaccine shot or not.

In part I, this section introduces the project background, challenges meet in this project and related work.

In part II, this section introduces the formal definition of the problem, including the framework of this project and the detailed parts of this problem.

In part III, this section introduces the algorithms which are used to make machine learning model, android app (client-side) development and server-side development.

In part IV, this section introduces the numerical solution of machine learning part of this project.

In part V, this section demonstrates the whole android app system including client-side and server-side.

In part VI, this section summaries this paper.

## Challenges

### Data Collection and Procession

In the past, research project assignments are with datasets given by professor. These datasets are all well designed for my projects, and no need to reprocess these datasets. This time one of the biggest problems is to get good datasets

### Machine learning

Unsupervised learning has many kinds of algorithms, such as linear regression, support vector machines. however, different algorithms have different performance on different task. Therefore, it’s necessary to pick some good algorithm for project.

### Android App Development

As brief description of my project, this app should be a user-friendly manner. When this app is being designed, ease of use must be considered. So, there are many restrictions and requirement when designing apps.

## Related Work

### COVID-19 Future Forecasting Using Supervised Machine Learning Models[4]

Furqan Rustam and his team use the dataset provided by the Center for Systems Science and Engineering, Johns Hopkins University. This dataset include features like latitude, longitude and the number of new inflect people.

They use 4 algorithms, Linear Regression, LASSO, Support Vector Machine and Exponential Smoothing to do future covid-19 forecasting. They set 85% of data as training data and 15% as test data.

ES and LASSO method perform good in forecast new cases with higher score and lower MSE. LR perform not bad and SVM cannot make a good prediction.

### Prediction of Influenza and the Associated Pneumonia in Taiwan Using Machine Learning[5]

Sing-Ling Jhuo and his team use a different kind of dataset and algorithm to study the trend prediction of Influenza. In their dataset, there are a lot more kinds of features such as temperature, relative humidity, PM2.5 and CO. Then their study dataset has more features.

Then they use a deep learning method, Multilayer Perceptron (MLP), to make model. Their study achieves 77.54% accuracy for the trend of influenza. The accuracy of elder population is especially good. The accuracy gets 81.16%.

# Formal Definition of the Problem

## Overall Framework

An android can collect data input by user, then machine learning model can calculate the predict of this data.

Then this app can display the prediction trend of covid-19 and flu or give app user suggestion whether he or she should take flu vaccine shot or not.

|  |
| --- |
|  |

1. Structure of Server Part and Client Side

## Data Procession Part

After collect dataset from internet, make them become a suitable dataset for supervised learning. Make sure that they have suitable labels and their features.

For example, the raw data of flu trend prediction records flu cases happened in Australia. The focus labels are the numbers of new cases. Therefore, we need to process data, like adding cases happens at same time and same place to one row.

## Machine Learning Part

Try different kinds of supervised machine learning algorithm to solve 3 problems, flu trend prediction, covid-19 trend prediction and flu shoot suggestion.

Then, compare their accuracy, MSE and other scores to choose suitable model. Pick about 2~3 models as final decision.

## Andriod APP Development Part

Client part is an android app which collects the data input by user and sends them to server; finally display the solution given by Server.

The main part of server part is assembling models based on ml model picked at machine learning part. Server part receive the data from client side and calculate them by machine learning models.

# Description of the Solutions

## Data Procession

### Covid-19 trend prediction data procession

The raw datasets include us\_counties\_covid19\_daily.csv (dataset 1 in following part), us\_county.csv (dataset 2 in following part), and Action.csv (dataset 3 in following part).

|  |
| --- |
|  |

1. Head of us\_counties\_covid19\_daily.csv (dataset 1)

|  |
| --- |
|  |

1. Head of us\_county.csv (dataset 2)

|  |
| --- |
|  |

1. Head of Action.csv (dataset 3)

I set dataset 1 as main dataset and I will merge necessary data from dataset 2 and dataset 3 to dataset 1. In dataset 1, there are two columns called fips and state, which will be used to merge dataset 2 and dataset 3. Fips is the code of county.

I use dictionary method (which is a special kind of hash table) in python to merge data. In merging dataset 1 and dataset 2, I set fips in dataset 2 as key and other local information as value. Then we can add local data to dataset 1. For example, the first row of dataset 1, fips is 1001, Then we can find the value of key 1001 in dictionary including its state, population and other elements. Then merge this to first row of dataset 1.

|  |
| --- |
|  |

1. Merge several datasets for Covid-19 dataset

Use dictionary method as before, I set state initial as key and other element as value for merging dataset 1 and dataset 3.

Then, we need to make policy change from date to yes or no. As we know the date and policy effect date. Therefore, it means policy is effective that date is after policy effect date, and I will record this as 1 and record policy no effective as 0.

Finally, drop the repeated data and unnecessary data to make final dataset for study. Randomly set 70% of dataset as train data and other 30% dataset as test data

### Flu Vaccine suggestion data procession

|  |
| --- |
|  |

1. Head of Flu Vaccine suggestion dataset (dataset 2)

Flu Vaccine dataset is a relatively mature dataset, no need to merge more data. However, there is some mysterious data in the dataset. For example, as in Fig.6. the data in hhs\_geo\_region column is ‘oxchjgsf’ and other data cannot be understood. Therefore, I need to drop these mysterious data for further study.

Finally, after drop the mysterious data to make final dataset for study. Randomly set 70% of dataset as train data and other 30% dataset as test data

### Flu trend prediction data procession

As is shown in Fig.7., the dataset record flu cases in Australia, but not how many new cases happen in a place. As I want to have dataset to train machine learning model to predict the trend of flu in a place, I need to change the dataset.

I set age 0-9 as group 1, 10-19 as group 2, 20-39 as group 3, 40~54 as group 4 and 55+ as group 5.

Then calculate the number of new cases in the same place, at the same time and in the same age group to form the new dataset.

|  |
| --- |
|  |

1. Head of Influenza (laboratory confirmed) Public datset 2008 to 2017 - Copy

Then as lack local information also no specific policy for flu. I search local population from internet. Add population to dataset.

Finally, drop some unnecessary data. Randomly set 70% of dataset as train data and other 30% dataset as test data.

## Machine Learning Part

### Instruction of Algorithmns used in this Project

#### Multiple Linear Regression & Lasso Regression

Linear regression is a linear approach model the relationship between value (yi) and one or more variables (xij). Their relation formulation i.e.

(1)

|  |
| --- |
|  |

1. Linear Regression Sample

MSE cost function of linear regression the average of the square distance between real value y and predict value i.e.

(2)

The difference between Lasso regression and linear regression is their cost function. The cost function of Lasso regression adds a penalty equal to the absolute value of coefficients i.e.

(3)

#### K-Nearest Neighbors Algorithm

Different from Linear regression, KNN is a non-parameter method machine learning algorithm. KNN algorithm is that if most of the K nearest samples in the feature space of a sample belong to a certain category, the sample also belongs to this category and has the characteristics of the samples in this category.

#### Decision Tree Algorithm

|  |
| --- |
|  |

1. Decision Tree Model Sample

Decision tree algorithm is a kind of predictive model. The decision tree algorithm uses a tree structure and uses layered inference to achieve the final classification.

#### Multi-layer Perceptron

A multilayer perceptron (MLP) is a kind of neural network, a mlp at least consist by 3 layers of node, an input layer, an hidden layer and on output layer

|  |
| --- |
|  |

1. Decision Multi-Layer Perceptron Network Sample

As other Neural network as CNN, MLP also need activation function. In this project, classification model use relu and thah as activation function in hidden layers, and use sigmoid activation function in the last layer and regressive model use linear activation function in the last layer.

#### Score Method

In this project, F1 score and R2 score are mainly used for judging a model is good or not. F1 score is used for classification model and R2 model is used for regression score.

F1 score is a measure of a test's accuracy. It is calculated from the precision and recall of the test. It’s more reasonable than accuracy score for judging a classification model is good or not. For example, in a dataset that 10 people is suffering cancel and 90 people are healthy. A model classifies all people healthy has 0.9 accuracy score, it’s unreasonable.to think that this model is good.

TP – Number of True positives

FP – Number of False positives

FN – Number of False negatives

R2 score is coefficient of determination, which is used for how good reflects the fluctuation of the dependent variable y (labels) what percentage can be used as the independent variable x (features). The higher R2 score, the better the model explains the dataset.

R2 score is equal to 1 - sum of squares of residuals / total sum of squares. i.e.

(5)

RSS - sum of squares of residuals

TSS - total sum of squares

(6)

(7)

n - number of samples

f(xi) - predicted value of yi;

- ith value of the variable to be predicted;

- mean value of a sample

### Covid-19 Trend Prediction

In Covid-19 trend prediction, I try linear regression, Lasso regression, decision tree, KNN and multilayer perceptron.

At the first try of linear regression and Lasso regression, the R2 score is relatively low. So, I don’t use these two models to solve this problem. Decision tree, KNN and multilayer perceptron are promising models for solving this problem.

Based on their R2 score, I set the weighted average predict new cases of these three models as the final assemble model. i.e.

(8)

– New cases predicted by assemble model

- New cases predicted by decision tree model

- New cases predicted by KNN model

- New cases predicted by MLP model

– R2 score of decision tree model

- R2 score of KNN model

- R2 score of MLP model

### Flu Vaccine Suggestion

As linear regression and Lasso regression cannot be used to solve a classification problem, in Flu Vaccine suggestion models, I try KNN, Decision tree and MLP algorithm. As the first try of these three models make a good solution with high F1 score, I try to study and improve all these three models.

As their average F1 score and accuracy score are very close, I use all of them to make flu vaccine suggestion.

Final score is equal to the sum of prediction. 0 meant suggest this person not to take flu vaccine, 1 means suggest to this person take flu vaccine. Suggestion based on final score. The higher final score, the stronger suggestion will be given to user.

1. FFlu Vaccine Suggestion

| Final score | Suggestion |
| --- | --- |
| 3 | Very strongly suggest |
| 2 | Strongly suggest |
| 1 | Suggest |
| 0 | Not suggest |

### Flu Trend Prediction

In flu trend prediction, I try linear regression, Lasso regression, decision tree, KNN and multilayer perceptron.

At the first try of linear regression and Lasso regression, the R2 score is very low. I will not use them to assemble final model. I try MLP, however the score is still not enough high. Only KNN and Decision tree model is relatively high. These model is not so good as former two function, in my opinion mainly because data is very limited for machine learning.

Based on their R2 score, I set the weighted average predict new cases of these three models as the final assemble model. i.e.

(8)

– New cases predicted by assemble model

- New cases predicted by decision tree model

- New cases predicted by KNN model

– R2 score of decision tree model

- R2 score of KNN model

## Andriod APP Development Part

### Client Part

The main function of client part is let user input data such as his/her age and other easy-to-get information. Save this information to let app autofill if user use this app next time.

|  |
| --- |
|  |

1. Function diagram of client side

At the first page of this app, let user choose the function he or she want to use, including flu trend prediction, covid-19 trend prediction and flu vaccine suggestion. At next page, user will input his or her data. As the requirement of this app is let user be easy to use. I delete some data column which may let user hard to use and this leads the score of models decrease little. After user input all data, this client will send all data to server side with the first element of data is function name. Other elements are data which will be used to make prediction. Final part of app is to display the received suggestion from server side.

|  |
| --- |
|  |

1. Function diagram of server side

### Server Part

Before the server side runs, it loads all 8 models first. Then client will judge which models will be used to predict. After it gets the prediction and assemble the prediction of 2~3 models, final suggestion will be sent to client side.

### TCP/IP (Socket Part)

Server side and Client part connect by TCP/IP (socket), server side will be started first to listen and accept socket from client side. Next, server side receives the data from client side. After machine learning models make prediction, server side send prediction or suggestion to client side.

# Numerical Results and Analysis

### Covid-19 Trend Prediction Model

1. FCovid Trend Prediction Score

| Algorithm | Score | |
| --- | --- | --- |
| R2 Score | MSE Score |
| Linear Regression | 0.763 | 1034.628 |
| Lasso Regression | 0.763 | 1037.829 |
| Decision Tree | 0.782 | 953.66 |
| KNN | 0.832 | 734.83 |
| MLP | 0.789 | 922.74 |

As table 1 shown, Linear regression & Lasso regression have 0.76 R2 score, lower than decision tree 0.78, KNN 0.83 and MLP 0.78 R2 score

Also, Decision tree, KNN and MLP have better performance at MSE score.

Final model is assembled by 2~3 models, so I picked KNN, MLP & Decision tree to assemble the model on server-side.

### Flu Prediction Model

1. FFlu Trend Prediction Score

| Table Head | Score | |
| --- | --- | --- |
| R2 Score | MSE Score |
| Linear Regression | 0.046 | 21423.66 |
| Lasso Regression | 0.046 | 21423.57 |
| Decision Tree | 0.741 | 5810.42 |
| KNN | 0.706 | 7250.93 |
| MLP | 0.257 | 16689.06 |

As table II shown, Linear regression & Lasso regression have 0.04 R2 score, MLP 0.25 R2 score which are lower than decision tree 0.74 and KNN 0.70

Also, Decision tree and KNN have better performance at MSE score.

Final model is assembled by 2~3 models, so I picked KNN & Decision tree to assemble the model on server-side.

1. FFlu Suggestion Prediction Score

| Table Head | Score | |
| --- | --- | --- |
| F1 Score | Accuracy Score |
| Decision Tree | 0.767 | 0.769 |
| KNN | 0.785 | 0.771 |
| MLP | 0.796 | 0.796 |

As table III shown, Decision tree classification, KNN classification and MLP classification are all have similar F1 score and Accuracy Score.

Therefore, I pick all of them to assemble the final models.

Suggestion will be given based on the total score of the model solution.

# System Demonstration (1 Page)

## Client Part (Andriod APP)

|  |
| --- |
|  |

1. Function Choose page and one of data input pages

|  |
| --- |
|  |

1. Merged recevid suggestion from server side.

## Server Part

|  |
| --- |
|  |

1. Server received data from android side.

# Conclusion (1/4 Page)

##### References

1. L. Mertz, "AI-Driven COVID-19 Tools to Interpret, Quantify Lung Images," in IEEE Pulse, vol. 11, no. 4, pp. 2-7, July-Aug. 2020, doi: 10.1109/MPULS.2020.3008354.
2. [J. Berglund, "Tracking COVID-19: There’s an App for That," in IEEE Pulse, vol. 11, no. 4, pp. 14-17, July-Aug. 2020, doi: 10.1109/MPULS.2020.3008356.J. Clerk Maxwell, A Treatise on

Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

1. A. Yousefi, O. A. Sianaki and D. Sharafi, "Long-Term Electricity Price Forecast Using Machine Learning Techniques," 2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), Chengdu, China, 2019, pp. 2909-2913, doi: 10.1109/ISGT-Asia.2019.8881604.
2. F. Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," in IEEE Access, vol. 8, pp. 101489-101499, 2020, doi: 10.1109/ACCESS.2020.2997311.
3. S. Jhuo, M. -. Hsieh, T. Weng, M. Chen, C. Yang and C. Yeh, "Trend Prediction of Influenza and the Associated Pneumonia in Taiwan Using Machine Learning," 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Taipei, Taiwan, 2019, pp. 1-2, doi: 10.1109/ISPACS48206.2019.8986244.