Predicting Adverse Events after COVID-19 Vaccination

Research Project

Group 4

YUAN Fangxu
20799126
fyuanad@connnect.ust.hk

LONG Yuepeng
20806228
ylongag@connect.ust.hk

GUO Yuchen
20793419
yguobq@connect.ust.hk

LIN Lirong
20812186
llinav@connect.ust.hk

LEI Lijun
20797037
lleiad@connect.ust.hk

DECLARATION

We hereby declare that all the work done in this Project is of our independent effort. We also certify that this project is done solely within the course but no other scopes (e.g., other courses and research projects), and we have never submitted the idea and product of this Project for academic or employment credits.

Group	Date
Group 4	*2021.11*

Content

1 Background

In the context of a still global epidemic of the COVID-19 virus, the most effective and cost-effective way to prevent and control the outbreak is vaccination. Currently, the following vaccines have been approved and licensed for the prevention of COVID-19 in the United States:

- mRNA vaccines requiring two doses: Pfizer/BioNTech, Moderna
- Viral vector vaccine requiring only one dose: J&J/Janssen

The effectiveness of vaccines and adverse reactions after vaccination has become a global concern. After widespread vaccination, many reports of post-vaccination adverse reactions have gradually been generated and submitted to the Vaccine Adverse Event Reporting System (data source for this project).

Potential side effects of vaccination include[1]:

- May occur in the arm that received the injection: pain, redness, swelling, etc.
- In the rest of the body may occur: fatigue, headache, muscle pain, fear of cold, fever, nausea, etc.

After vaccination, coma (fainting) and other anxiety-related reactions such as rapid breathing, low blood pressure, numbness, or tingling may occur. In addition, although extremely rare, adolescents and young adults may develop myocarditis and pericarditis after mRNA COVID-19 vaccination (Pfizer-Biotech or Moderna), and women under 50 years of age may develop rare blood clots accompanied by platelet deficiency after Johnson & Johnson vaccination.

This study aims to evaluate side effects following COVID-19 vaccination and to predict the need for hospitalization based on the severity of the side effects. Ongoing monitoring and evaluation of these post-vaccination adverse reactions may improve our understanding of safety issues and rationalize medical resources.

2 Introduction

To achieve the project objectives that we mentioned above, the methods used in our project are broadly as follows.

For hospitalization prediction, to solve the problem of high-dimensional sparse features[2][3], we used the common dimensionality reduction methods, such as Principal Component Analysis (PCA), and Naïve Bayes, to reduce the dimensionality of original data.

Specifically, in PCA, we handled imbalanced data with under-sampling and with class weight, respectively to handle the zero-probability problem. In addition, significant symptoms were selected based on posterior probabilities.

For onset time prediction, we used baseline patient characteristics such as gender, age, and medication history as input features and used different algorithms, such as linear regression models, random forest, and neural network, to predict onset time.

The contribution of this paper is:

1. Carrying out an algorithmic implementation that uses sparse PCA and Naïve Bayes to extract principal components of sparse symptom features, where PCA is followed by logistic regression with balanced class weights to calculate the risk of hospitalization;
2. Evaluating several methods for predicting the onset time of adverse events.

The rest of the paper is organized as follows. Section 3 provides the data and coding foundations for open source. Sections 4 and 5 focus on the main issues in the project and the corresponding prediction scenarios. Sections 6 and 7 are the evaluation and individual case studies. Finally, several recommendations were made for algorithm improvements, re-selection, and forecasting model upgrades.

3 Framework

3.1 Dataset and Features

VAERS is a nationwide reporting system for approved vaccinations that is meant to discover early safety issues. The system accepts reports from healthcare practitioners, vaccine makers, and the general public. VAERS makes a yearly dataset available for download that has been independently confirmed for reliability and validity.

Only adverse events data connected with COVID-19 immunizations were selected from the VAERS 2021 datasets (up to April 20, 2021). Because a single individual might report several occurrences, each person was assigned a unique “VAERS ID”.

The MedDRA[4] Terms from the standard MedDRA codebook were used to encode the exact symptoms associated with each incident in VAERS. In total, there are 5487 different types of symptoms (MedDRA terms). However, on average, each patient had just 5 symptoms documented. The feature has a large size and is relatively sparse. We constructed a dictionary of these symptom names and assigned each individual an array of 0 (did not occur) and 1 (did occur) to indicate if a given symptom happened to them, and we used a list of indexes to record each patient's reported symptoms.

Age, sex, vaccine manufacture, current disease, medication usage, allergy history, and pre-existing illnesses were all considered baseline factors and were encoded as continuous variables (standardized) or Boolean variables (0 or 1), respectively.

Because pre-existing illnesses were recorded in VAERS as narrative text, we scanned each text and found the 17 most prevalent conditions, which we then converted into 17 distinct features (Boolean variables) for each person.

3.2 Data Interpretation

The incidence (rate) and distribution of adverse events after receiving COVID-19 immunizations in the United States were used to characterize the data. The number of adverse events was estimated by multiplying the number of persons vaccinated on the same day in VAERS by the total vaccination on that day to obtain the rate statistics. The rate and distribution of the occurrences were described using a bar plot and a line chart.

4 Hospital Prediction

We use machine learning approaches to classify the patients with needs for hospitalization for treatment after COVID-19 (a.k.a. SARS-Cov-2) vaccination and find the significant symptoms.

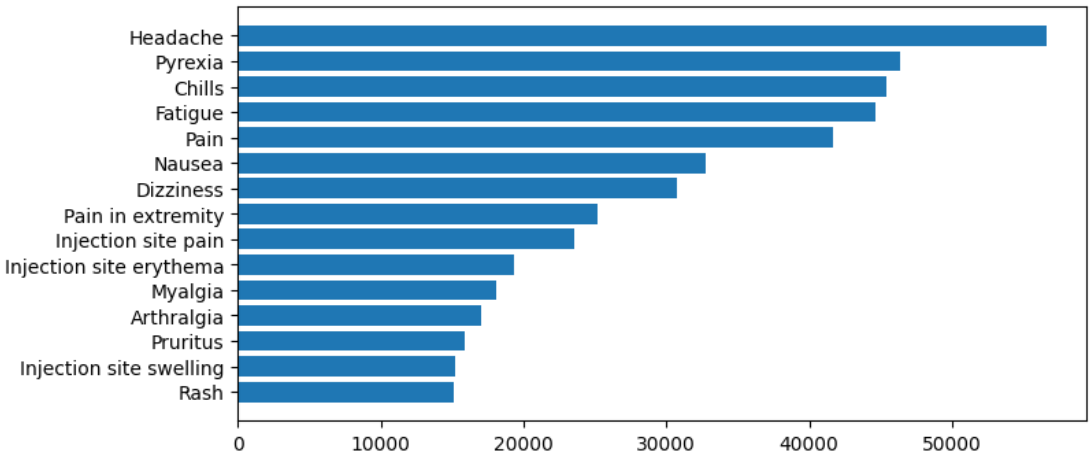

Figure 1 top 15 ranking symptoms of the adverse event. The 5 most common symptoms are headache, pyrexia (aka fever), chills, fatigue, and pain respectively

In this task, two difficulties need to be solved. Firstly, the input features are sparse encoded symptoms reported by patients. There were 5487 symptoms reported and encoded by MedDRA terms in total[4]. And in our dataset, there were 6725 symptoms. However, on average each patient only self-reported 5 symptoms resulting in a high dimensional and highly sparse input feature matrix. Secondly, since most patients with self-reported adverse effects do not need hospitalization, the output labels are highly imbalanced. Even if we predicted all the outcomes as no need for hospitalization, the accuracy can still be up to 95%. In this paper, we used

sensitivity and specificity to value our method, which are common evaluation criteria in a clinical setting.[5]

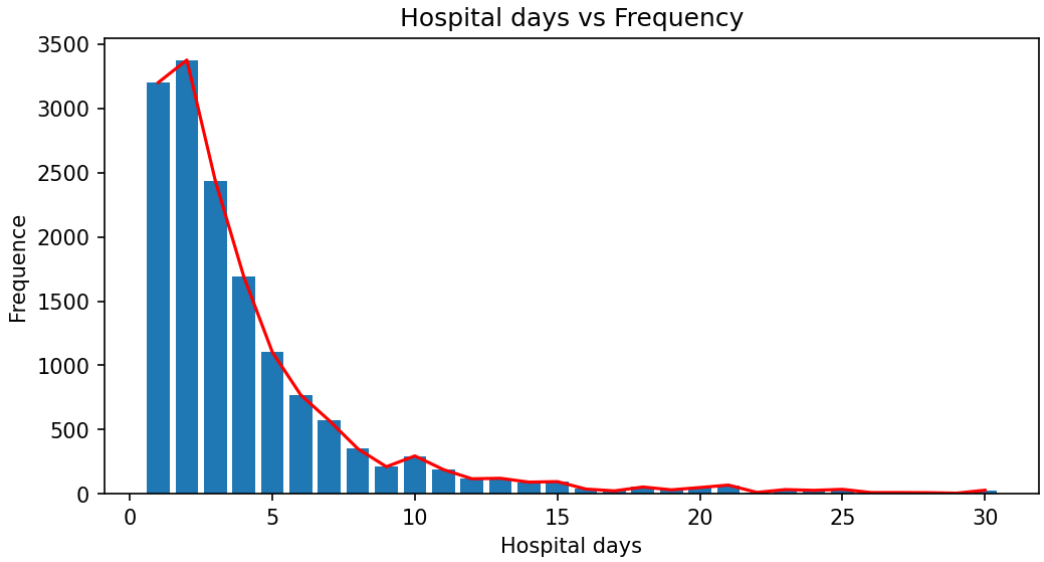

Figure 2 frequencies of hospital days (within 30 days). Most people who receive the vaccine are at risk of being hospitalized in the first week

These two ill-conditioned data structures were taken into consideration separately in our method as follow:

4.1 Dimensionality reduction with Sparse PCA

To solve the problem of high dimensional sparse features, we first used sparse PCA[6],[7] to reduce the dimensionality of features from 6725 to 5 or 10. On the other hand, as the label is highly imbalanced and our focus was the significant symptoms that would lead to patients' needs for hospitalization, we only used the symptoms reported by patients with the need for hospitalization to obtain the principal components. Later, we used the transformation to obtain the low dimensional features of all training set samples, and logistic regression[8] with balanced class weights to compute the risk of hospitalization.

4.2 Sparse Feature Selection with Naive Bayes

We used a Naïve Bayes classifier with sparse constraint [9] for the Bernoulli distribution model and Laplace smoothing to tackle the zero-probability problem. The risk of hospitalization is then computed. And the significant symptoms are selected according to the posterior probabilities.

5 Onset Time Prediction

In this task, we used the patient's gender, age, and medication history as input variables, and tried to predict the onset time (in days) from the time of vaccination to the onset of the adverse event. Also, we pointed out the key predictors in the task that have high significance.

Since the same patient in the dataset may submit multiple reports, we de-duplicated the dataset, leaving the earliest report submitted by each patient as the input data from that patient. Most of the input variables were binary variables, hence we standardized and centralized the dataset before prediction. There were 27 predictors (baseline variables) in total. The dataset was split into a training set and a test set in a ratio of 8:2.

We use the following 4 models for onset prediction and key predictor identification:

- Ordinary least squares (OLS). Predictor importance was assessed by the magnitude of p-values.
- Regularized regression (including Lasso, Ridge, and Elastic Net). Regularization parameters were chosen by 10-fold cross-validation. Predictor importance was assessed by the magnitude of coefficients.
- Random forest. The number of trees in the forest was set to be 500. The number of predictors sampled for splitting at each node was set to be 5 (square root of the number of predictors). Predictor importance was assessed by the mean decrease in accuracy.
- Artificial neural network. For simplicity, we used a simple multi-layer perceptron with 2 hidden layers. Each hidden layer had 5 neurons. ReLU activation function was applied. Predictor importance was assessed by the magnitude of the sum of weights of the first hidden layer.

All models were trained by the training set and were evaluated by the root mean square error (RMSE) on the test set.

6 Evaluation

6.1 Data Interpretation

From January 1, 2020, to April 20, 2021, a total of 245,908 persons reported adverse events in the VAERS system producing 336,486 adverse events. Of all events, 1477956 (43.9%) were reported on the first day of vaccination. The maximum duration between the event and vaccination date was 90 days.

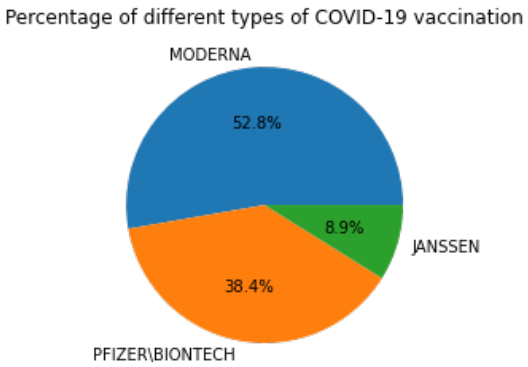

Figure 3 the proportion of people in the dataset who received the three mRNA types of vaccines

6.2 Evaluation of Onset Time Prediction

Table 1 shows the performance of different models in predicting the onset time. All models produced errors of around 5 days. Neural network (MLP) performed slightly better than other models, while random forest performed slightly worse than others. Since more than 2/3 of the input dataset has onset time within 3 days, this result indicated that more variables of patients needed to be considered in the task.

	Training RMSE	Test RMSE
OLS	5.37476826	5.34608356
Lasso	5.37477033	5.34600854
Ridge	5.37476781	5.34606870
Elastic Net	5.37477033	5.34600854
Random Forest	4.85620750	5.55952327
Neural Network	5.36143203	5.33521479

Table 1

Table 2 reveals the top 5 of the key predictors in the task that have high significance. Linear regressors (with and without L1/L2-norm penalization) all agreed with “Age”, “Other medication”, “Disability”, “Gender” and “Hypertension” as the best predictors when predicting shorter onset time (within 3 days), while they agreed “Age”, “Gender”, “Disability”, “Allergic history” and “Other medication” as the best predictors when predicting longer onset time (more than 3 days). Random forest and neural network highlighted the importance of some other predictors that were not included in linear regressors, such as “Current illness”, “Thyroid”, “Hyperlipidemia”, “Dementia” and “Migraine”.

	Key predictors (shorter onset)	Key predictors (longer onset)
OLS	Age Other medication Disability Gender Hypertension	Age Gender Disability Allergic history Other medication
Lasso	Other medication Age Disability Gender Hypertension	Age Gender Allergic history Other medication Disability
Ridge	Other medication Age Disability Gender Hypertension	Age Gender Allergic history Other medication Disability
Elastic Net	Other medication Age Disability Gender Hypertension	Age Gender Allergic history Other medication Disability
Random Forest	Age Current illness Allergic history Thyroid Hypertension	
Neural Network	Hyperlipidemia Allergic history Age Dementia Migraine	

Table 2

6.3 Evaluation of Hospitalization Prediction

The results of each method are shown in Table 3. The ROC curves of each method are shown in Figure 4 and Figure 5. Both the two methods have good performance in ruling in the patients with needs for hospitalization (high specificity). However, both have low sensitivity, which means a high rate of false negatives.

	Sparse PCA and logistic regression	Sparse naïve Bayes
Opitmal probability threshold	0.47	0.02
AUC	0.724323986	0.51508871
Training set sensitivity	0.59513435	0.108206245
Training set specificity	0.794572527	0.985125352
Validation set sensitivity	0.558379666	0.453534551
Validation set specificity	0.794441085	0.93525051

Table 3

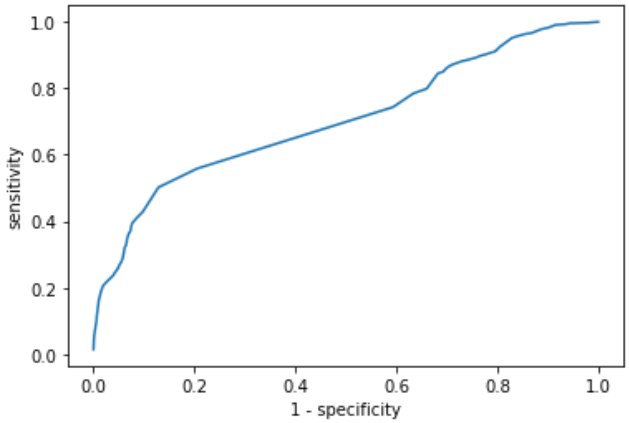

Figure 4

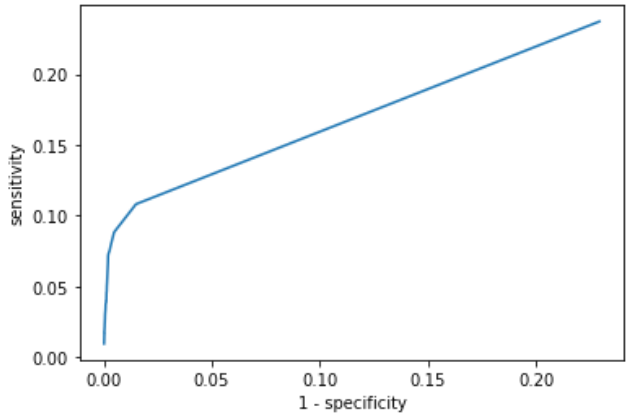

Figure 5

The top 20 significant symptoms leading to hospitalization obtained by sparse PCA and logistic regression method were 'Chills', 'Blood test', 'Pyrexia', 'Computerised tomogram', 'Unevaluable event', 'Dyspnoea', 'Pain', 'Nausea', 'Fatigue', 'Myalgia', 'Vomiting', 'Death', 'SARS-CoV-2 test negative', 'Hyperhidrosis', 'Malaise', 'Diarrhoea', 'Arthralgia', 'Pain in extremity', 'Abdominal pain', 'White blood cell count increased'.

And the top 20 significant symptoms leading to hospitalization obtained by sparse Naïve Bayes method were: 'Pyrexia', 'Pain', 'Fatigue', 'Chills', 'Headache', 'Nausea', 'Injection site pain', 'Injection site erythema', 'Pain in extremity', 'Dizziness',

'Myalgia', 'Arthralgia', 'Pruritus', 'Rash', 'Injection site swelling', 'Dyspnea', 'Injection site pruritus', 'Vomiting', 'Erythema', 'Asthenia'.

There are 4 overlapped symptoms of these two methods: 'Chills', 'Pyrexia', 'Pain', 'Fatigue'.

The sparse Naïve Bayes method successfully extracted the frequently occurring symptoms but failed to specify the symptoms that will lead to hospitalization. On the contrary, as the sparse PCA method obtained the principal components from the samples in need of hospitalization, it successfully extracted some of the severe adverse side effects that would lead to hospitalization.

7 Case Study

7.1 Onset Time Prediction

Interestingly, by observing the coefficients and p-values given by the regression models we used, there is evidence that Moderna's vaccine is more predictive of shorter onset time than the other two vaccines. This may be due to the fact that Moderna's vaccine has high amount of vaccine in a single dose, which causes patients' immune systems to respond more strongly, and therefore adverse events occur earlier. This could also be observed in the bar chart of the onset time and the corresponding number of patients (shown in Figure 6).

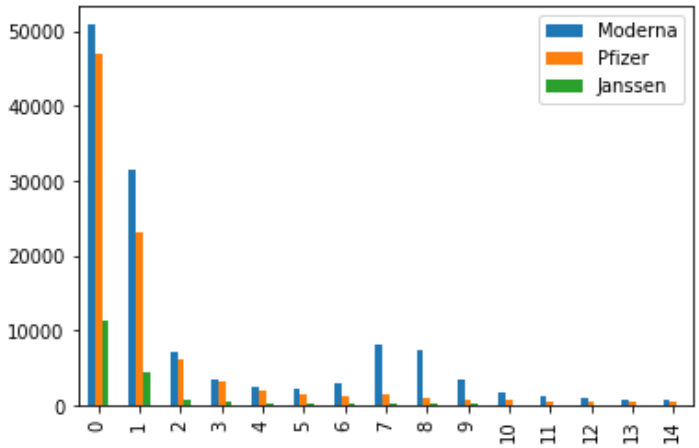

Figure 6

7.2 Hospitalization Prediction

	RF	LGBM	SVM	DT	XGB	GBM	Sparse PCA	Sparse Naïve Bayes
Sensitivity	0.69	0.67	0.42	0.66	0.69	0.72	0.56	0.45
AUC	0.72	0.72	0.6	0.71	0.73	0.71	0.72	0.52

Table 4

We compare our work with the work of Ahamad et al[10], which conducted the identification and classification of post-vaccination reactogenicity of COVID-19 vaccination, using the same data source. They used decision tree and random forest, support vector machine, and gradient boosting machine as classifiers to find the significant features leading to the hospitalization and death of patients. However, they pre-processed their data to solve the sparse symptom feature problem, by which only 86 most frequently appeared symptoms were selected and combined.

Since only the AUC and sensitivity metrics but not the specificity metrics are shown in their work, we only compared the AUC and sensitivity. As our work is to consider all symptoms, so in general, the effect is certainly not as good as considering only 86 kinds of symptom baseline. Unexpectedly, our models which consider all symptoms, especially sparse PCA, get similar scores in Sensitivity and AUC compared to the baselines, even better than SVM in both metrics.

8 Related Work

There is a lot of work on the intelligent prediction of diseases. For example, V. Jackins et al.[11] took a patient's symptom dataset as input and applied data mining algorithms such as Random Forest and Gaussian Parsimonious Bayes to estimate whether the patient was affected by diseases such as heart disease and cancer and whether further medical treatment was needed.

Ma'mon M. Hatmal and others[12] conducted a statistical analysis of recorded data such as vaccine type, demographic data, and simplified recorded symptoms by surveying residents who received any COVID-19 vaccine, and performed statistical analysis on the recorded data, using the multi-layer perceptron (MLP), eXtreme gradient boosting (XGBoost), random forest (RF), and K-star to build a predictive model to predict the severity of side effects; the predicted severe cases may require more medical care or even hospitalization. They use the chi-square test to address the potential association of post-vaccination side effects with the number of received COVID-19 vaccine doses, the chi-square test (Yate's corrected) was also used to assess the potential association of post-vaccination side effects with the different types of COVID-19 vaccine.

9 Discussion

By accomplishing the onset time prediction and hospitalization prediction, we can apply them to real life and thus bring more safety to the people.

Nowadays, the epidemic is not getting better around the world, and even in such a serious situation, many people still choose not to get vaccinated. Experts say there are various reasons for this, including a lack of availability to the vaccine, an unwillingness to consider Covid-19 a threat, fear of side effects, a lack of faith in the vaccination or the agency that developed it, and belief in at least one of several conspiracy theories. Some of these factors overlap and combine; for example, if someone does not believe Covid-19 is a significant hazard, the vaccination may not be worth the risks.

We may not be able to increase vaccine production or change people's minds about the Covid-19, but we can help those who are afraid of the side effects that come with vaccination.

Later, by optimizing the neural network and expanding the database, we can better predict the different kinds of adverse reactions of different types of vaccines received by the prepared vaccine recipient through his information, and then recommend the most suitable vaccine to him. Even if symptoms develop after the vaccination, we can predict whether the vaccinated person needs to be hospitalized by describing the symptoms.

We can help those people who are worried about adverse effects following immunization to some level.

10 Conclusion

In the current study, a comprehensive case analysis was compared and our work compared all symptoms in terms of AUC and sensitivity scores for sparse PCA, ultimately finding that it was better than SVM on both metrics and also better than Naïve Bayes. At the same time, the neural network slightly outperformed other algorithms in predicting the onset time, which we speculated may be due to the hidden layer exploiting the complementarity between predictors to improve prediction performance.

11 Future Work

Machine learning for high-dimensional sparse features with highly unbalanced labels is always a challenging problem. Linear Discriminant Analysis (LDA), Locally linear embedding (LLE), Laplacian Eigenmaps, and other methods were not used in this work, but we will try to use them in the future work and will continue to explore machine learning methods that can effectively handle high-dimensional sparse features.

Acknowledgment

We would like to thank Prof. Raymond Wong for his constructive suggestions and devotion to teaching and the hard work of TA Weicheng WANG.

Contribution

GUO Yuchen:

1. **Conceptualization -** Participate in pre-discussions on the development of overall research goals and objectives.
2. **Methodology -** Development of methodology. Suggest using cluster in the project to create an unsupervised pre-classification (K-Means) according to the disease to get the hypothetical clusters with the mean values of the disease features as centers and set a threshold to calculate the MSE error of onset time to ensure a new cluster class center (Shown in the proposal but not adopted in the end because there was a better way).
3. **Data curation -** Scrub real data set disease top-15 and sorted research data for initial use and later reuse.
4. **Visualization -** Preparation, creation of the visualization data presentation with the real data set disease (top-15).
5. **Writing -** Writing the original draft of the proposal (including substantive translation); writing part of the final report with the introduction, conclusion, and future works.

LIN Lirong:

In this project, I came up with my own ideas for the conception of the experimental ideas for the whole project in the early stage (in the proposal): in the onset time prediction, I came up with the idea of using clustering to do the prediction of patients' diseases (although we do not use it at last). Then, I do the visualization of the Length of stay for the real data set part to show that of those who develop symptoms as a result of vaccination, how many of them are hospitalized. And of course the writing part of both proposal and the final report.

LEI Lijun:

In this project, I participated in the pre-and mid-project discussions, reviewed and studied papers related to the project (i.e., related to disease prediction), and studied how they conducted their experiments, what machine learning algorithms were used, what results were achieved, and how the results obtained by different methods were compared and wrote the relevant parts of the proposal based on this. In the data processing of the project, I was involved in some of the visualization work. During the development of the project, I reviewed background information and relevant papers for reference and wrote the background and related work sections of the final report.

YUAN Fangxu:

In this project, I was involved in the discussion of the project and the identification of the topic. In the middle of the project, I did the data cleaning work, completing the code for the hospital prediction part and conducting experiments. Finally, I was responsible for the completion of the Hospital Prediction part and Evaluation, case study during the writing of the report.

LONG Yuepeng:

In the early stage of the project, I was involved in the selection of the project topic and the discussion of the methodologies. In the middle stage of the project, I was involved in cleaning and filtering the raw data used in the project; meanwhile, I completed the code and experiments related to onset time prediction by studying and referring to the methods of other papers, then analyzed the results of the experiments. Finally, I was mainly responsible for the preparation of the onset time prediction part of the report.

References

[1] https://chinese.cdc.gov/coronavirus/2019-ncov/vaccines/expect/after.html

[2] Afnan M. Alhassan and Wan Mohd Nazmee Wan Zainon. Review of feature selection, dimensionality reduction and classification for chronic disease diagnosis. IEEE Access, 9:87310–87317,2021.

[3] Zhuoyuan Zheng, Yunpeng Cai, Yujie Yang, and Ye Li. Sparse weighted naive bayes classifier for efficient classification of categorical data. In 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), pages 691–696. IEEE, 2018

[4] Patricia Mozzicato. MedDRA. Pharmaceutical Medicine, 23(2):65-76,2009

[5] https://ebn.bmj.com/content/23/1/2

[6] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In Proceedings of the 26th annual international conference on machine learning pages 689-696,2009.

[7] Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Structured sparse principal component analysis. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pages 366-373.JMLR Workshop and Conference Proceedings,2010.

[8] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b Fortran subroutines for large-scale bound-constrained optimization. ACM Transactions on Mathematical Software (TOMS),23(4):550-560,1997.

[9] Armin Askari,Alexandre d'Aspremont, and Laurent El Ghaoui. Naive feature selection Sparsity in naive bayes. In International Conference on Artificial Intelligence and Statistics, pages 1813-1822.PMLR,2020

[10] Md Martuza Ahamad, Sakifa Aktar, Md Jamal Uddin, Md Rashed-A1-Mahfuz, AKM Azad, Shahadat Uddin,Salem A Alyami,Iqbal H Sarker, Pietro Lid,Julian MW Quinn,et al. Adverse effects of covid-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity. medRxiv,2021.

[11] Jackins V, Vimal S, Kaliappan M, et al. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes[J]. The Journal of Supercomputing, 2021, 77(5): 5198-5219.

[12] Hatmal M M, Al-Hatamleh M A I, Olaimat A N, et al. Side Effects and Perceptions Following COVID-19 Vaccination in Jordan: A Randomized, Cross-Sectional Study Implementing Machine Learning for Predicting Severity of Side Effects[J]. Vaccines, 2021, 9(6): 556.