Journal of Health Informatics in Developing Countries
http://www.jhidc.org/
Vol. 15 No. 2, 2021

Submitted: Jan 21st, 2021 Accepted: Dec 11th, 2021

Applying Machine Learning to predict Hand-Foot-Mouth disease outbreaks in Vietnam

Dr Thanh Ngoc Nguyen*, Dr Dinh Ngoc Minh
School of Science, Engineering & Technology, RMIT University, Vietnam

Abstract

Applying Machine Learning to find out the patterns of hand foot mouth diseases is critical to understand the impacts of social-natural conditions to the outbreak of the diseases. This paper uses data from Vietnam to find out what factors contribute the most to the increase of cases and what models can help to predict this increase. We identify temperature as the important factors and the Random Forest Regressor is the model that produces best results.

Keywords: Machine Learning; HFMD; Health Informatics.

1. Introduction

Machine learning is a field of computer application that has very fast development. More and more data coming from the presence of ubiquitous computing devices such as smartphones, web applications, internet of things etc. have created huge opportunities for data mining and data analytics and mathematics modeling .

Machine learning has seen its applications in various fields including sales prediction, business decision making. The raw materials for machine learning are training data which are an accumulation of historical business transaction data [1, 2]. Many giant systems have been exploiting their data for machine learning to deliver a better experience to their users and customers. For example, Facebook and Google keep track of user preference and demographic data for better advertisement targets. Gmail uses machine learning for email spam filtering. There has been an increasing number of companies using machine learning to improve their business.

*Thanh Nguyen Ngoc-School of Science, Engineering and Technology, RMIT University Vietnam; Email: thanh.nguyenngoc@rmit.edu.vn.

Hand Foot Mouth Disease (HFMD) is a serious disease that often happens to children under 5-year-olds [3]. There are increasing infected cases in Vietnam, which currently is about 50 and 100K every year including several deaths. Predicting the pattern of the HFMD is important to control it. There have been many studies that apply Machine Learning to understand the patterns of HFMD in different contexts [2, 4, 5]. However, there has not been any studies on the impact of social and natural conditions on the outbreak of HFMD in Vietnam. Therefore, the aim of this study is to find out the link between social and natural conditions and the outbreak of HFMD in Vietnam.

The remainder of this paper is divided as follows. The related work will be discussed in section 2 followed by a research approach in section 3. Section 4 analyzes the data and section 5 presents the findings. Discussion and conclusion are provided in section 6 and section 7 respectively.

2. Related Works

There have been many studies that applied machine learning algorithms to understand the patterns of HFMD in different parts of the world.

Liu, Xu [6] conducted a study to predict cases of high risk of severe HFMD and compare effectiveness with existing models (Pediatric Critical Illness Score - problem is not designed for HFMD). The data features used for the study include (Severe/Mild), Gender (Male/Female), Vomiting (Yes/No), Age (month), Respiratory rate (/min), Peak temperature (°C), Fever duration (day), Blood glucose (mmol/L), Platelet (109 /L), Percentage of lymphocytes (%), Lactate dehydrogenase (IU/L), Alkaline phosphatase (IU/L), Creatine kinase (IU/L), Creatine kinase-MB (IU/L), Creatinine (μmol/L), Uric acid (μmol/L), Blood chlorine (mmol/L), Alanine aminotransferase (IU/L).

By using Random Forest Classifier, the system provides result up to an AUC of 0.916 score using 16 variables with a higher performance compared to traditional models. Lactate dehydrogenase, Creatine kinase - MB, Blood Glucose, Creatine kinase and Vomiting are top 5 indicators for severe HFMD.

In a similar vein, Liu, Liao [7] attempted to Identify clinical and MRI-related predictors for the occurrence of severe HMFD in children and assess the interaction effects between the indicators by using Machine Learning. The data used include:

- Demographic characteristics: age and sex
- Clinical symptoms and signs: oral ulcers, rash or herpes, duration of fever, vomiting, tachycardia, convulsion
- Altered consciousness: irritability, lethargy, drowsiness, and/ or coma, neck stiffness or positive Kerning's sign, muscle weakness, breathlessness, hypertension and elevated body temperature
- EV-A71 test results

• WBC count

• Chest radiograph

• MRI reports

As a result, they have identified top predictors which are WBC count, spinal cord involvement, spinal nerve roots involvement, hyperglycemia, brain/spinal meninges involvement, EV/AH infection. The performance is without balance approach: 92.3% accuracy, 0.985 AUC, 0.85 Sensitivity, 0.96 Specificity, without balance approach: 89.2% accuracy, 0.948 AUC, 0.8 Sensitivity, 0.93 Specificity.

There are also efforts from a public health perspective trying to use meteorology to predict HFMD. For example, Liu, Bao [8] forecast the incidence of HFMD using Back Propagation (BP) Neural Network (NN) in JiangSu, China. Data used in the study includes Monthly case numbers of HFMD in JiangSu province, Monthly meteorological data from JiangSu Meteorological Service Center: rainfall (RF), sunshine duration (SD), relative humidity (RH), minimum temperature (MIN_T), maximum temperature (MAX_T), atmospheric pressure (AP), wind velocity (WV). As a result, they conclude that:

• Univariate spearman correlation analysis indicates that all meteorological factors were significantly associated with the incidence of HFMD except: SD, RH, WV

• Strong correlation includes (with correlation coefficient above 0.9): mean temperature, MAX_T, MIN_T, AP

• To avoid multicollinearity, only mean temperature and RF were used

• Further cross-correlation analysis indicated that both mean temperature and RF have significant relation with the incidence of HFMD at lag 1, with correlation coefficient of 0.235 (p=0.0216) and 0.251 (p=0.0146)

Although these studies try to understand the link between social and natural conditions to the outbreaks of HFMD, they mostly focus in China. There have not been similar studies that are conducted in Vietnam. The differences in terms of climate, weather, and social conditions between Vietnam and China are quite significant. Therefore, by this study we try to find out new patterns that fit to the Vietnam context. Our study is also motivated by the fact that Vietnam is one of the countries seriously affected by HFMD. While the healthcare sector in Vietnam is still underdeveloped, applying new technologies like Machine Learning into prediction and forecasting of HFMD in Vietnam could help the government better plan and respond to the epidemic.

3. Research approach and data collection

The social and meteorological data is easily obtained from public sources. The medical data which

involves the number of cases of HFMD in Vietnam is collected from an Infectious Tracking System governed by Administration of Medical Services, Ministry of Health, Vietnam.

The social data is collected from the General Statistics Office of Vietnam (link) (2016-2017) and the meteorological data is collected from Power Data Access Viewer (link) (2016-2018). The number of HFMD cases per month per province is provided by the Ministry of Health (2016-2018). Some provinces are missing HFMD data for some months. However, we couldn't confirm whether either data is missing or there are 0 cases in that province in that month. Hence, such cases will be ignored as they do not affect the overall outcome too much (missing cases are remote provinces and tend to have low TOTAL_CASES in months they have reports)

Please see the Appendix for more detailed collected data used in this study. Table 1 shows all features of the dataset:

Table (1) All features of the dataset

Feature name	Type	Unit	Description	Available year
POPULATION_DENSITY	Social	Population density	people/*km*2	2016, 2017
T2M_RANGE	Meteorological	Temperature Range at 2 Meters	°C	2016, 2017, 2018
PRECTOT	Meteorological	Precipitation	mm day-1	2016, 2017, 2018
T2MWET	Meteorological	Wet Bulb Temperature at 2 Meters	°C	2016, 2017, 2018
T2M	Meteorological	Temperature at 2 Meters	°C	2016, 2017, 2018
WS50M_MIN: Minimum Wind Speed at 50 Meters (m/s)	Meteorological			2016, 2017, 2018
PS	Meteorological	Surface Pressure	kPa	2016, 2017, 2018
T2M_MAX	Meteorological	Maximum Temperature at 2 Meters	°C	2016, 2017, 2018
TS	Meteorological	Earth Skin Temperature	°C	2016, 2017, 2018
WS10M_RANGE	Meteorological	Wind Speed Range at 10 Meters	m/s	2016, 2017, 2018
RH2M	Meteorological	Relative Humidity at 2 Meters	%	2016, 2017, 2018
WS10M_MIN	Meteorological	Minimum Wind Speed	m/s	2016, 2017, 2018

		at 10 Meters		
WS10M	Meteorological	Wind Speed at 10 Meters	m/s	2016, 2017, 2018
WS50M	Meteorological	Wind Speed at 50 Meters	m/s	2016, 2017, 2018
WS50M_MAX	Meteorological	Maximum Wind Speed at 50 Meters	m/s	2016, 2017, 2018
T2M_MIN	Meteorological	Minimum Temperature at 2 Meters	°C	2016, 2017, 2018
WS50M_RANGE	Meteorological	Wind Speed Range at 50 Meters	m/s	2016, 2017, 2018
WS10M_MAX	Meteorological	Maximum Wind Speed at 10 Meters	m/s	2016, 2017, 2018
QV2M	Meteorological	Specific Humidity at 2 Meters	kg kg-1	2016, 2017, 2018
TOTAL_CASES	HFMD	Total number of HFMD cases of a province in a month	cases	2016, 2017, 2018

4. Data Analysis

4.1 Feature engineering

We calculated the correlation between all features. POPULATION_DENSITY has the highest correlation score with TOTAL_CASES, this might due to the closer we are to each other, the higher the possibility of infection of HFMD. T2M, T2MWET, T2M_MAX, T2M_MIN have high correlation with each other, which are all about feature about temperature, representing the average, max and min of temperature (same as wind speed features: WS10M, WS10M_MAX, WS10M_MIN, WS10M_RANGE, WS50M, WS50M_MAX, WS50M_MIN, WS50M_RANGE), this explains the high correlation. Figure 1 describes the correlation between features.

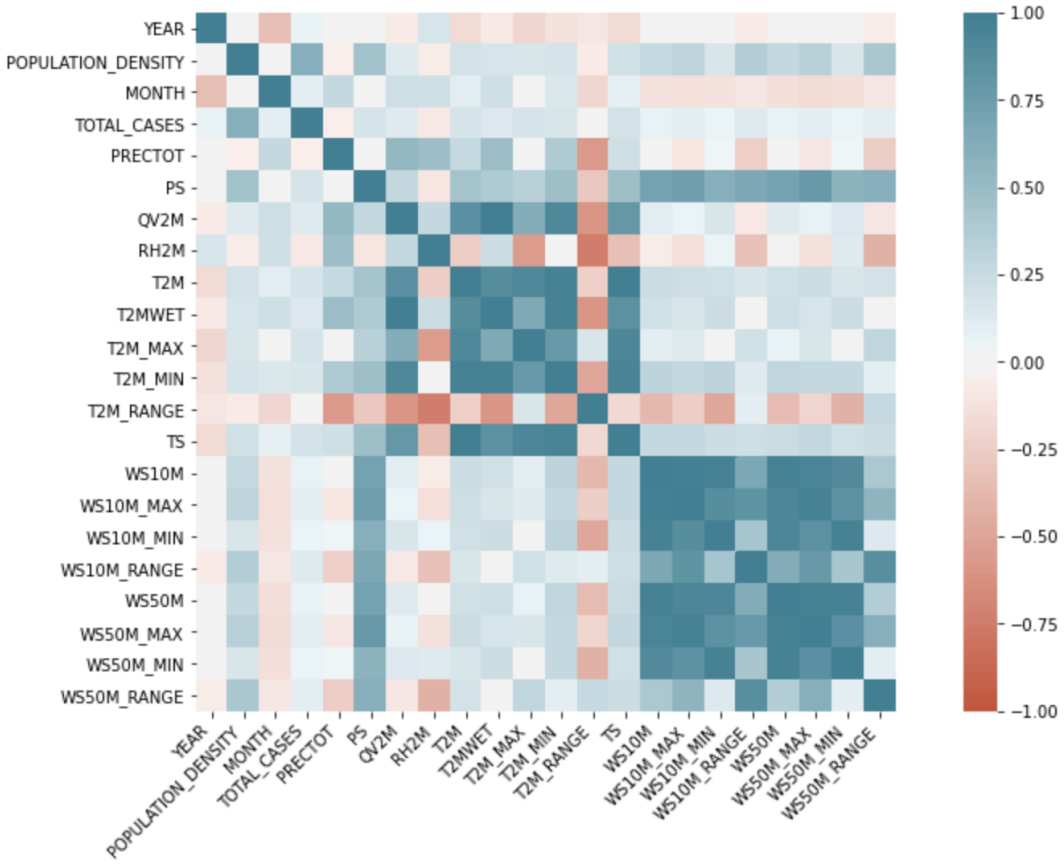


Figure (1) Correlations among data features.

4.2 Data training

The data is then split into training and testing sets with a ratio of 80% for training and the rest 20% for testing. We trained the data on 3 models: Linear Regression, Decision Tree and Random Forest.

Linear Regression is probably the most common and simple algorithm used in Machine Learning. Linear regression is a linear model that assumes the linear relationship between two variables (x, y). In simple words, linear regression allows calculation of variable y (output) from the variable x (input).

Decision Tree is a non-parametric learning method that can be used to predict an output value based on simple decision rules. Decision Tree is also easy to use and visualize.

Random Forest is a supervised learning method used for prediction. It is a robust algorithm that uses groups of decision trees to learn from input data when a Decision Tree regression is not enough for modeling. It will need to involve a number of decision trees.

The training process will involve training using all features including POPULATION_DENSITY (on dataset of 2016 and 2017) and without POPULATION_DENSITY (on dataset of 2016, 2017 and 2018 as we don’t have available POPULATION_DENSITY in 2018).

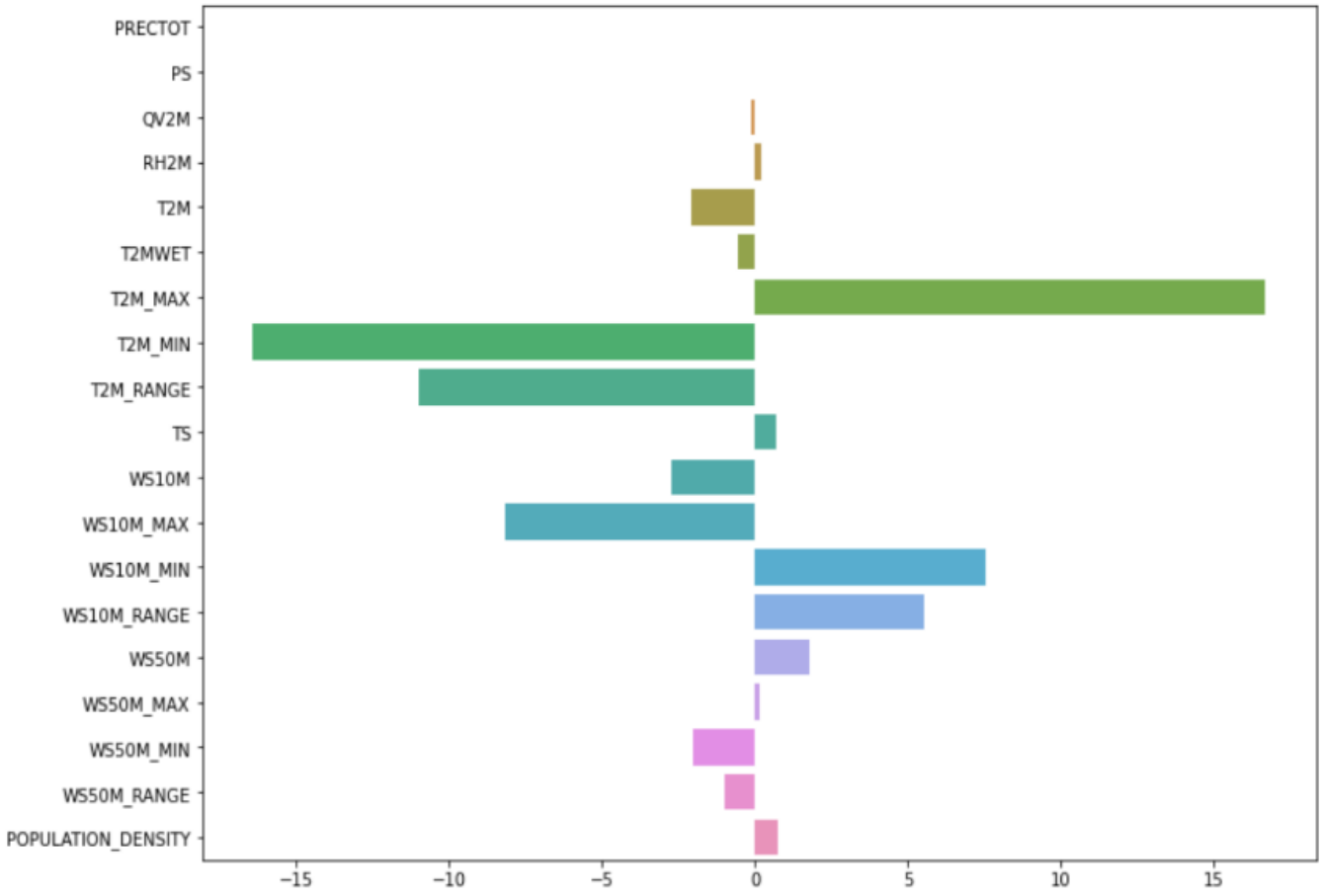
The exhaustive search, Grid Search, and Cross Validation with 5 folds will be used to find-tune the parameters for the Decision Tree Regressor and Random Forest Regressor model. Feature engineering was also implemented by recursively removing features based on feature importance of Decision Tree and Random Forest and the validation loss of the models in each recursion to find out the set of features that can best represent the outcome.

5 Findings

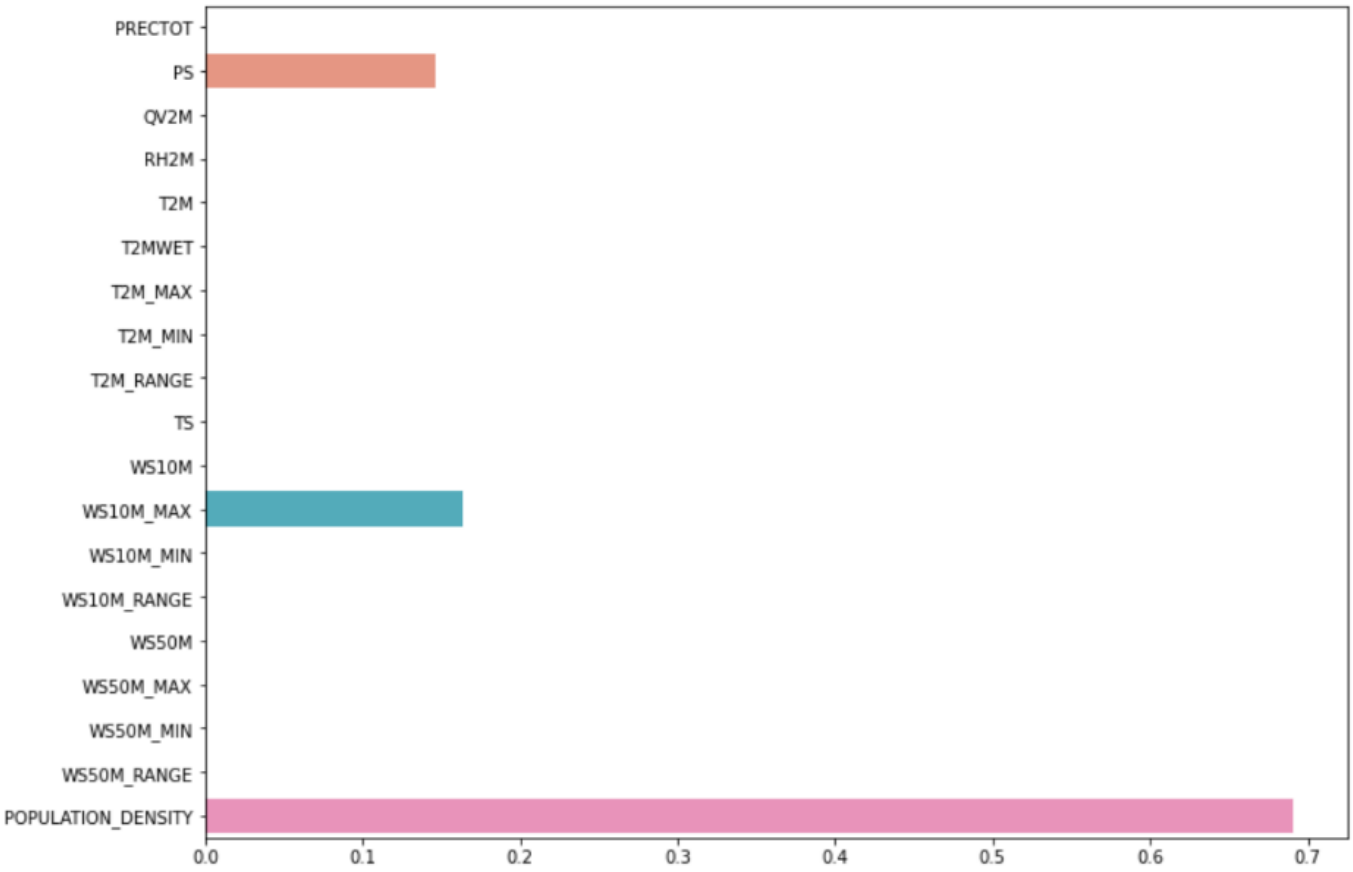
We have built and trained models for different cases to identify which data features have the biggest impact on the total number of HFMD cases. The correlation coefficients of 6 cases are presented as follows:

5.1. Without feature engineering

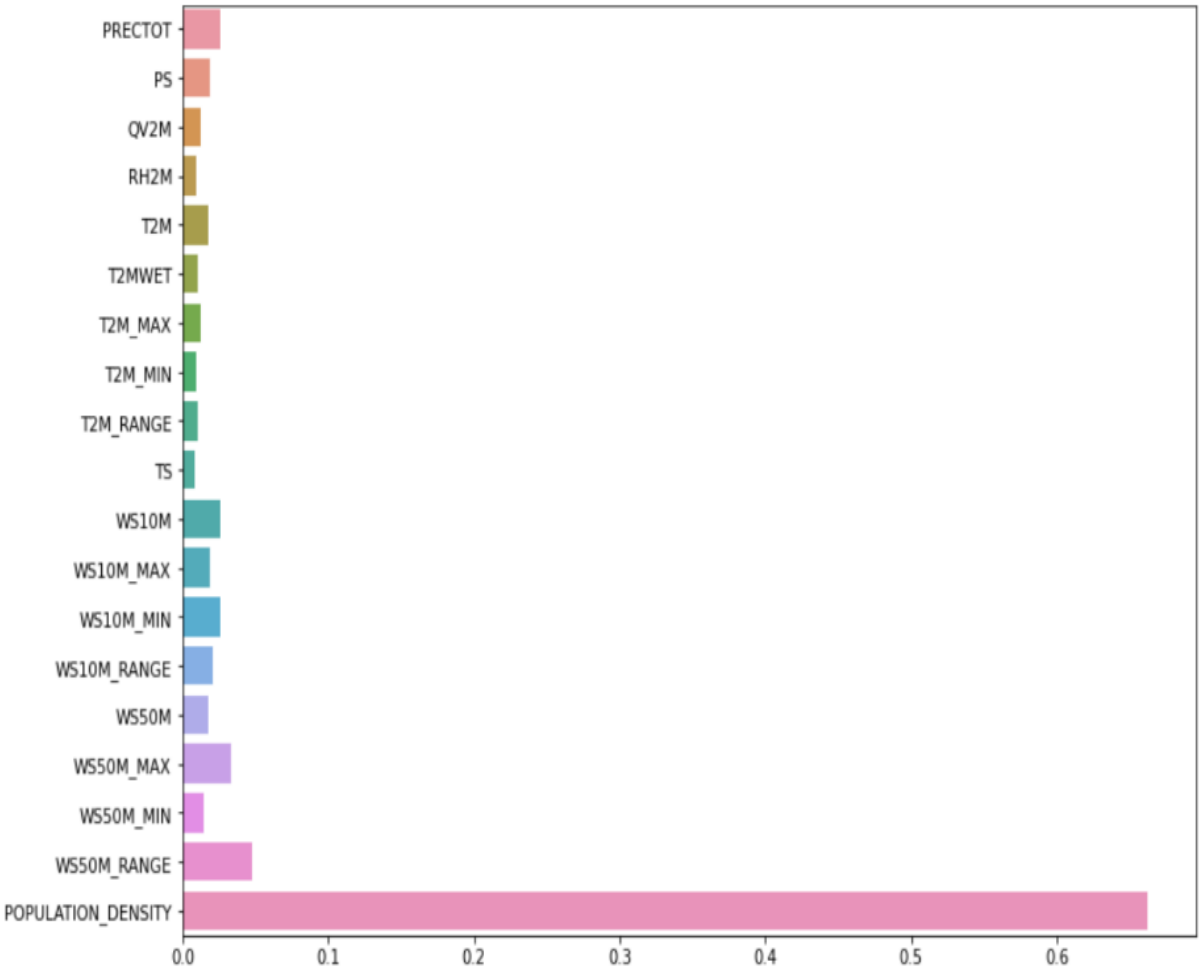
a. Linear Regressor



b. Decision Tree Regressor



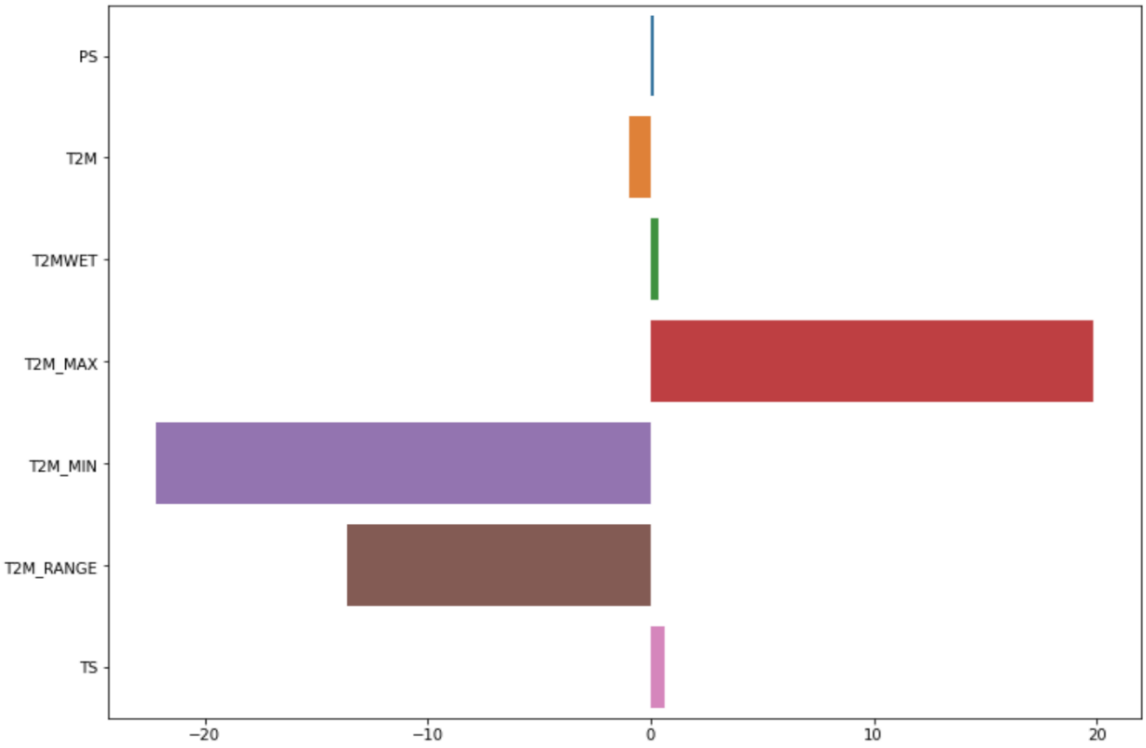
c. Random Forest Regressor



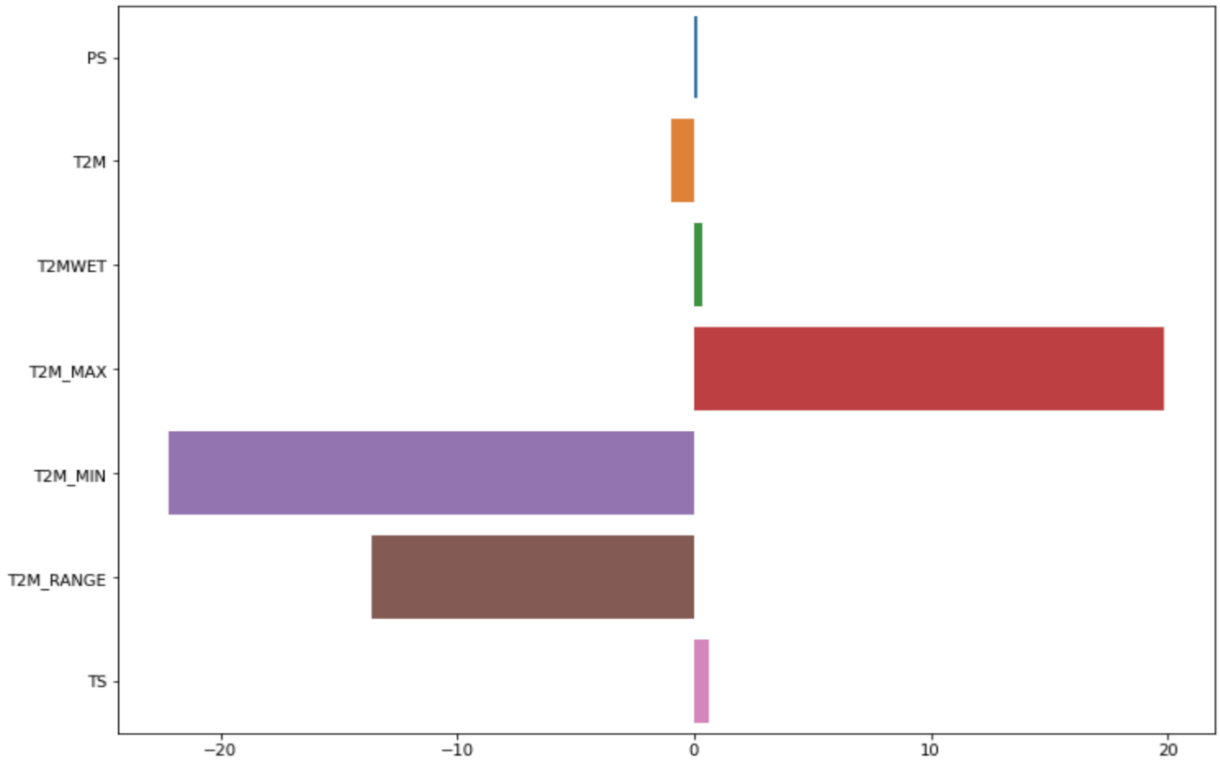
As we can see without feature engineering, population density is the feature that has the biggest impact on the total number of HFMD cases, regardless of the algorithms we used. The dependence of total cases on the population is quite obvious for any communicable diseases including HFMD.

5.2.Without feature engineering

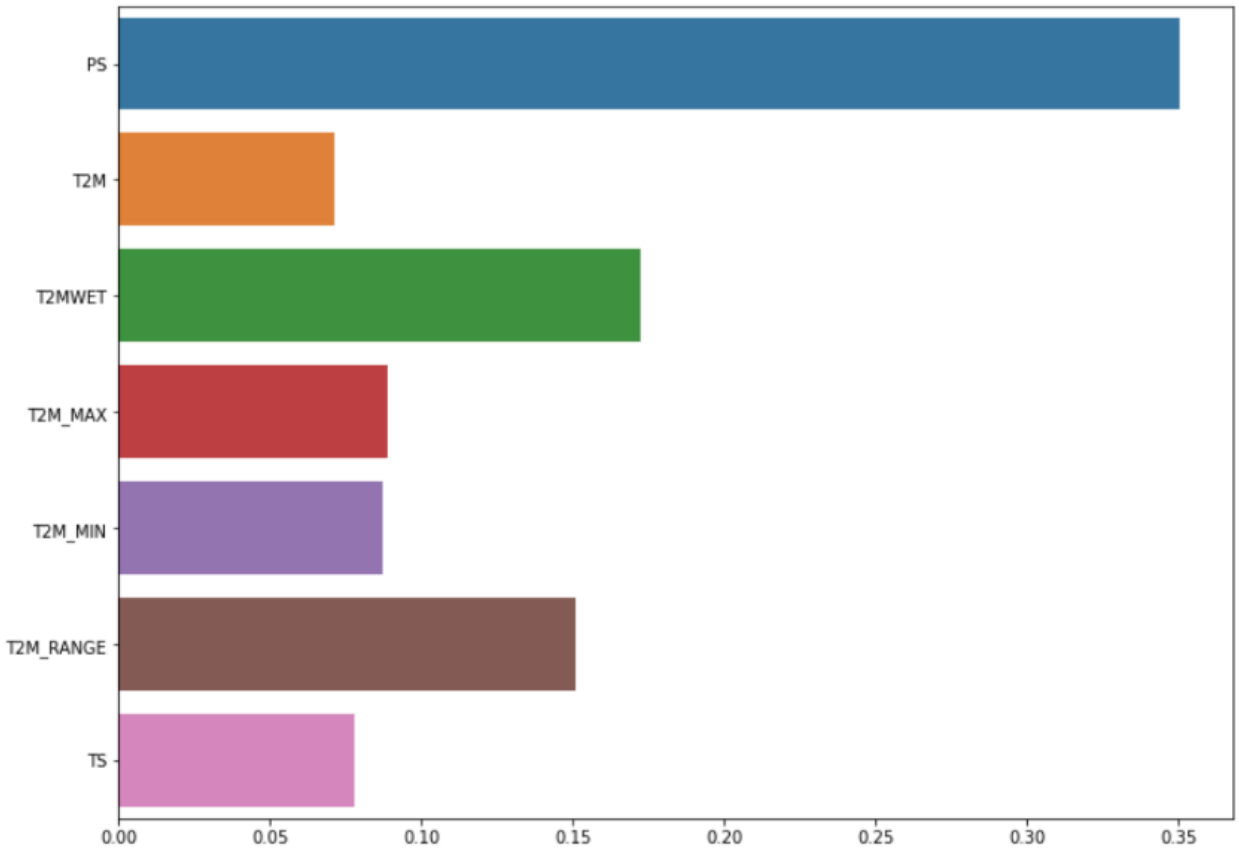
a. Linear Regressor



b. Decision Tree Regressor



c. Random Forest Regressor



6. Discussion

When training all features with the dataset of 2016-2017, the feature importance of Decision Tree and Random Forest show high importance of POPULATION_DENSITY. This is understandable as the more the population grows, the denser the population is, the higher chance of spreading the disease, which was mentioned above. But it doesn't tell us much aside from this fact. After feature engineering, Decision Tree Regressor algorithm gives highest feature importance score to PS, T2M_RANGE, TS, while Random Forest Regressor is the same with an addition of T2M, T2MWET, T2M_MAX, T2M_MIN, T2MWET, TS. These are temperature features, which indicates temperature has a high relation to the infectious rate. To be more precise, the higher the temperature, the higher the infectious rate is.

When training the data of 2016-2018 (without POPULATION_DENSITY feature), Linear Regression doesn't have much changes after feature engineering. As for Decision Tree Regressor, the features with high importance are PS, T2MWET, T2M_MAX, T2M_MIN. Random Forest Regressor also, again, shares the same top feature importance with an addition of TS. These are temperature features, which, once again, indicates temperature has high relation to the infectious rate.

Feature engineering helps reduce validation loss significantly for Decision Tree Regressor and

Random Forest Regressor.

Table 2 represents the test loss of all models when trained on all features and after feature engineering.

Table (2) Test loss of all models

	All features 2016-2017 (with Population)	**Feature engineering 2016-2017**	**All features 2016-2018 (without Population)**	**Feature engineering 2016-2018**
Linear Regression	~154.81	~133.68	~183.37	~182.94
Decision Tree Regressor	~114.27	~91.68	~130.1	~122.38
Random Forest Regressor	~93.47	~114.01	~160.67	~124.44

7. Conclusion

We concluded that temperature has the highest impact on the infectious rate of HFMD, the higher the temperature, the higher the infectious rate. This conclusion is similar to the other studies[9-11]. Our contribution in this paper is twofold. First, we identify temperature as a key influential factor to the outbreak of HFMD. Second, we find out that the Random Forest Regressor is the most appropriate model for predicting the outbreaks based on temperature.

For future work, we will investigate individual provinces to remove the difference in population and population density which has huge effects on the infectious rate of HFMD. Some provinces and cities such as Ho Chi Minh city have very high infection totals which can be due to their large population and high population density. The number of such provinces are very low, which can cause high variation of the model. Instead of looking at each month as a separate instance, time-series models such as ARIMA or LSTM can be used to imply time as a feature to the dataset, which couldn't be done in this paper due to the lack of data.

8. Declarations

8.1.Conflict of Interest Statement

The author has no conflict of interests to declare.

8.2.Funding Disclosure

This research is supported by RMIT Internal Research Grant 2020.

9. References

1. Cao, C., et al. *Research on the environmental impact factors of hand-foot-mouth disease in Shenzhen, China using RS and GIS technologies*. in *2012 IEEE International Geoscience and Remote Sensing Symposium*. 2012. Ieee.
2. Takahashi, S., et al., *Hand, foot, and mouth disease in China: modeling epidemic dynamics of enterovirus serotypes and implications for vaccination.* PLoS medicine, 2016. **13**(2): p. e1001958.
3. Owatanapanich, S., et al., *Risk factors for severe hand, foot and mouth disease.* Southeast Asian J Trop Med Public Health, 2015. **46**(3): p. 449-59.
4. Xu, M., et al., *Non-linear association between exposure to ambient temperature and children's hand-foot-and-mouth disease in Beijing, China.* PloS one, 2015. **10**(5): p. e0126171.
5. Cheng, Q., et al., *Ambient temperature, humidity and hand, foot, and mouth disease: A systematic review and meta-analysis.* Science of the Total Environment, 2018. **625**: p. 828-836.
6. Liu, G., et al., *Developing a machine learning system for identification of severe hand, foot, and mouth disease from electronic medical record data.* Scientific reports, 2017. **7**(1): p. 1-9.
7. Liu, X., Y. Liao, and Z. Zhu, *Machine Learning Algorithms for Important Feature Evaluation and Prediction of Severe Hand-Foot-Mouth Disease in Hunan Province, China.* Machine Learning, 2019. **6**: p. 15-2019.
8. Liu, W., et al., *Forecasting incidence of hand, foot and mouth disease using BP neural networks in Jiangsu province, China.* BMC infectious diseases, 2019. **19**(1): p. 1-9.
9. Xu, Z., et al., *The effect of temperature on childhood hand, foot and mouth disease in Guangdong Province, China, 2010–2013: a multicity study.* BMC infectious diseases, 2019. **19**(1): p. 1-10.
10. Hao, J., et al., *Impact of Ambient Temperature and Relative Humidity on the Incidence of Hand-Foot-Mouth Disease in Wuhan, China.* International journal of environmental research and public health, 2020. **17**(2): p. 428.
11. Pearson, D., et al., *Temperature and hand, foot and mouth disease in California: An exploratory analysis of emergency department visits by season, 2005–2013.* Environmental Research, 2020: p. 109461.

Appendix

Kaggle Notebook for data analysis

https://www.kaggle.com/toanquach/hfmd-analysis-wo-population/notebook