

The prediction of receiving flu vaccine based on machine learning

COMP4030 Data Science with Machine Learning

Xiting Wang
20489134
psxxw11@nottingham.ac.uk

Yifeng Peng
20493001
psxyp12@nottingham.ac.uk

Abstract—This study investigates the impact of merging behavioral, attitudinal, and doctor recommendation features on prediction performance and also explores how data analysis can be used to predict the probability of individuals receiving H1N1 and seasonal flu vaccinations. The dataset contains 35 features with varying numbers of missing values. Literature review and methodologies involving exploratory data analysis, preprocessing, and classification are discussed.

Index Terms—vaccination, data merging, exploratory data analysis, classification

I. INTRODUCTION TO THE DATASET AND RESEARCH QUESTIONS

The dataset is divided into three parts:

- training_set_labels.csv
- training_set_features.csv
- test_set_features.csv

The **training_set_features.csv** consists of 36 columns, with 35 columns representing features and the remaining column being 'respondent_id' used as the row index. Except for seven features, which do not contain missing values, the remaining columns have varying numbers of missing values, ranging from tens to tens of thousands.

The **training_set_labels.csv** contains two label columns that indicate whether an individual is vaccinated against H1N1 or seasonal influenza. The **test_set_features.csv** is the corresponding test set to **training_set_labels.csv** and is used for submitting prediction models and evaluating the predictive performance in the competition.

This study aims to investigate the potential impact of merging certain features, including individuals' behavioral features, attitudinal features, and doctors' recommendations regarding vaccines, on prediction performance. And in terms of behavioral and attitudinal aspects, whether there exists a positive association, where individuals with generally positive attitudes are more inclined to engage in self-protection measures.

II. LITERATURE REVIEW

A. The research process for similar datasets

Exploratory Data Analysis (EDA) refers to the process of visualizing and performing descriptive statistical analysis on data to understand its distribution, correlations, and outliers. It is the first step that data scientists undertake after acquiring

the data. During this stage, there are modules available, such as pandas-profiling, that can efficiently generate relevant statistical information about the data. For example, in Gordon et al.'s [1] study, it was demonstrated that Pandas Profiling is a versatile approach to data analysis. Researchers have also explored alternative EDA tools, such as DataPrep.EDA proposed by Peng et al. [2], which showed improved speed and user experience compared to pandas-profiling.

Data preprocessing and cleaning are typically the second step in data analysis. It involves addressing issues such as missing values, outliers, and duplicates using techniques like interpolation, deletion, or replacement. Data preprocessing and cleaning have always been challenging tasks, as mentioned by Rahm et al. [3], emphasizing the need to consider the inherent constraints of the data when designing suitable methods.

The third step in data analysis involves feature engineering. Common techniques in feature engineering include feature selection (Li et al. [4]), which evaluates the correlation and importance of features with the target variable to select meaningful features for modeling. Feature scaling (Juszczak et al. [5]) is used to normalize features and eliminate biases or imbalances caused by different scales or units. Feature encoding (Mougan et al. [6]) involves encoding non-numeric data to make it suitable for machine learning models, among other techniques.

The modeling stage is the later phase of data analysis, where researchers explore various machine learning algorithms. Common machine learning algorithms include decision trees, random forests, support vector machines, logistic regression, neural networks, and k-nearest neighbors, as mentioned in Mahesh et al.'s [7] study.

B. Handling and prediction on the same dataset

In the same prediction problem, Adalseno et al. [8] conducted the following steps. In the EDA phase, they utilized pandas profiling, dtale, and SweetWiz for comprehensive data analysis. For high-dimensional features, they employed feature selection techniques using the mlxtend library. In terms of prediction, they utilized the CatBoost algorithm with hyperparameter tuning using Optuna. The final prediction accuracy achieved was 0.8608. Furthermore, through further

improvements, the updated prediction accuracy as of 16/02/21 was reported as 0.8638, ranking seventh overall.

In the feature engineering phase, Shyam R [9] employed a data merging approach to combine all behavioral features into a new feature called "cleaness". Similarly, the opinion features were merged to create two new features representing people's opinions on seasonal and H1N1 vaccines. In the prediction stage, multiple models were used, including LinearSVC, GaussianNB, and 18 other models. The models were ranked based on their prediction accuracy, with the GradientBoostingClassifier performing the best. The average accuracy achieved by this model was reported as 0.835605.

III. METHODOLOGY

Overall, Yifeng Peng and Xiting Wang employed different approaches in each of the three stages. In Data Analysis Stage, Yifeng Peng focused on the visualization of numerical features and the understanding and merging of opinion-related features and doctor recommendation-related features. Xiting Wang, on the other hand, emphasized the visualization analysis of categorical features and the understanding and merging of behavioral-related features.

In Preprocessing Stage, Xiting conducted research on the correlation of categorized data based on the exploratory data analysis (EDA) stage. She employed a method to fill missing values in one column with the corresponding value from another column that has the highest correlation, which is considered the most likely value. For numerical data, the mode filling method was used to replace missing values with the most frequently occurring value in each column. Yifeng employed the mean filling method for numerical data and the mode filling method for categorical data.

In Classification Stage, Xiting and Yifeng utilized the data processed in their respective Stage 2 using different methods. Xiting applied the K-Nearest Neighbors (KNN) classifier and Random Forest classifier. Meanwhile, Yifeng experimented with the Decision Tree model and Support Vector Machine (SVM) model for classification purposes.

A. Data Analysis

In the data analysis stage, both team members in this study employed descriptive statistics to calculate basic statistical measures of the dataset, such as mean, median, standard deviation, maximum, minimum, etc., to gain insights into the distribution and summary information of the data. They utilized graphs, charts, and visualization tools to present the characteristics and patterns of the data, including techniques such as heatmaps and stacked bar charts. Furthermore, they conducted correlation analysis using Pearson's correlation coefficient to assess and explore the relationships between variables. Correlation matrices or heatmaps were created to visualize the correlations. Additionally, the **pandas_profiling**[CITATION] module was used as a supportive and convenient tool during the exploratory data analysis (EDA) phase of the study.

After conducting an initial exploration of the correlation between different features and the importance of their impact on

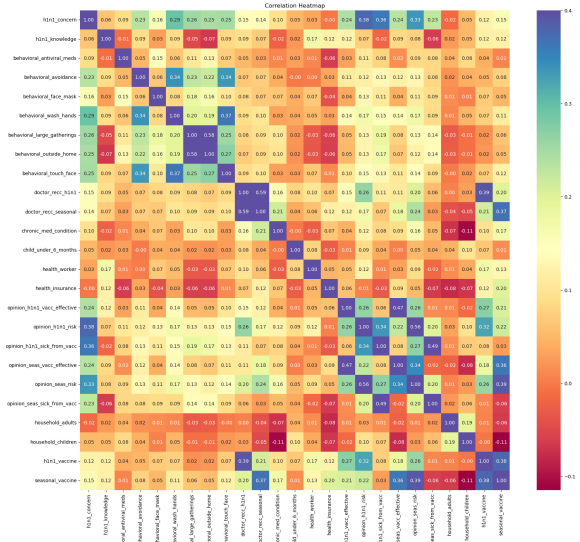


Fig. 1: Heatmap of the dataset

the prediction labels, it was concluded that certain unnecessary and redundant features could be eliminated:

- **No positive correlations:** household_children, household_adults, census_msa, hhs_geo_region
- **Imbalance, too much values are 0, cannot be used to predict:** behavioral_antiviral_meds, behavioral_face_mask, children_under_6_months
- **Too much missing values:** health_insurance, employment_industry, employment_occupation
- **High related features:** health_worker is high related with employment_industry, can be deleted.

Based on the attributes and meanings of the feature data, the retained feature values can be broadly categorized into several groups:

- **The notability of h1n1:** h1n1_concern, h1n1_knowledge
- **The behavioral of protect themselves from the flu:** behavioral_avoidance, behavioral_wash_hands, behavioral_large_gathering, behavioral_outside_home, behavioral_touch_face
- **Doctor's Recommendation:** doctor_recc_h1n1, doctor_recc_seasonal
- **Personal Condition (Numerical):** chronic_med_condition
- **Opinion toward the vaccine:** opinion_h1n1_vacc_effective, opinion_h1n1_risk, opinion_h1n1_sick_from_vacc, opinion_seas_vacc_effective, opinion_seas_rick, opinion_seas_sick_from_vacc
- **Personal Condition (Categorized):** age_group, education, race, sex, income_poverty, marital_status, rent_or_own, employed_status

In relation to the research question of the impact of merging certain features on the prediction results, the behavioral feature values can be combined into two new feature values based on their correlation with H1N1 and seasonal vaccine, namely,

protection_h1n1 and **protection_seas**. The attitudinal feature values can be merged into two new feature values, namely, **opinion_h1n1** and **opinion_seas**. Furthermore, the doctor recommendations can be combined into a single new feature value, **doctor_recc**.

1) *Merging behavioral feature values (by Xiting):* weights were assigned to each feature based on their respective importance and subsequently summed:

$$\begin{aligned} \text{behavior} = & \text{behavioral feature}_1 * \text{weight}_1 + \\ & \dots + \text{behavioral feature}_n * \text{weight}_n \end{aligned} \quad (1)$$

Regarding the data exploration of grouping behavioral feature values, weight for each feature is determined by assigning importance based on its relevance to the H1N1 vaccine target. This process involves the following steps: (1) Obtain a set of importance weights for the existing five behavioral features based on their correlation coefficients with the H1N1 vaccine target. (2) Generate a set of importance coefficients using a random forest model and derive a second set of importance weights. (3) Combine the two sets of weights using a 0.5:0.5 weight ratio. (4) Repeat the same steps for the seasonal vaccine.

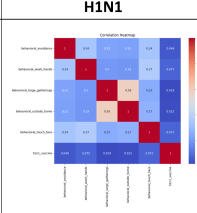
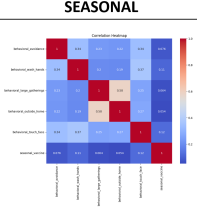
Vaccine	H1N1	SEASONAL
Heatmap		
Weight1 (By correlation coefficient)	0.2 0.32 0.08 0.09 0.31	0.18 0.26 0.15 0.13 0.28
Weight2 (By random forest)	0.16 0.28 0.11 0.14 0.31	0.14 0.32 0.11 0.08 0.35
Weight (Overall)	0.18 0.3 0.095 0.115 0.31	0.16 0.29 0.13 0.105 0.315

Fig. 2: The weight generated for h1n1 and seasonal vaccine

2) *Merging opinion and recommendation feature values (by Yifeng):*

$$\begin{aligned} \text{opinion_h1n1} = & \text{opinion_h1n1_vacc_effective} * \\ & \text{corr_ohve_hv_weight} + \text{opinion_h1n1_risk} * \\ & \text{corr_ohr_hv_weight} - \text{opinion_h1n1_sick_from_vacc} * \\ & \text{corr_ohsfv_hv_weight} \end{aligned} \quad (2)$$

(1) The weight of each feature value for H1N1 opinion was determined based on its correlation coefficient.

(2) The same method was applied for determining the weights of each feature value for seasonal opinion.

(3) For the two recommendation features, equal weights of 0.5 were assigned to each feature.

$$\begin{aligned} \text{doctor_recc} = & (\text{doctor_recc_h1n1} + \\ & \text{doctor_recc_seasonal}) / 2 \end{aligned} \quad (3)$$

B. Preprocessing

In the preprocessing stage, both team members initially removed the feature columns that needed to be discarded using

the drop() function. Then, based on the equations derived during the EDA stage, they merged the features to achieve data dimensionality reduction. Subsequently, Yifeng employed the mean filling method for numerical data and the mode filling method for categorical data. Xiting, on the other hand, used the mode filling method for numerical data and selectively performed interpolation based on the heatmap depicting the correlation between categorical features obtained during the EDA stage.

The following is an illustration of the correlation heatmap between features and the corresponding interpolation method used:

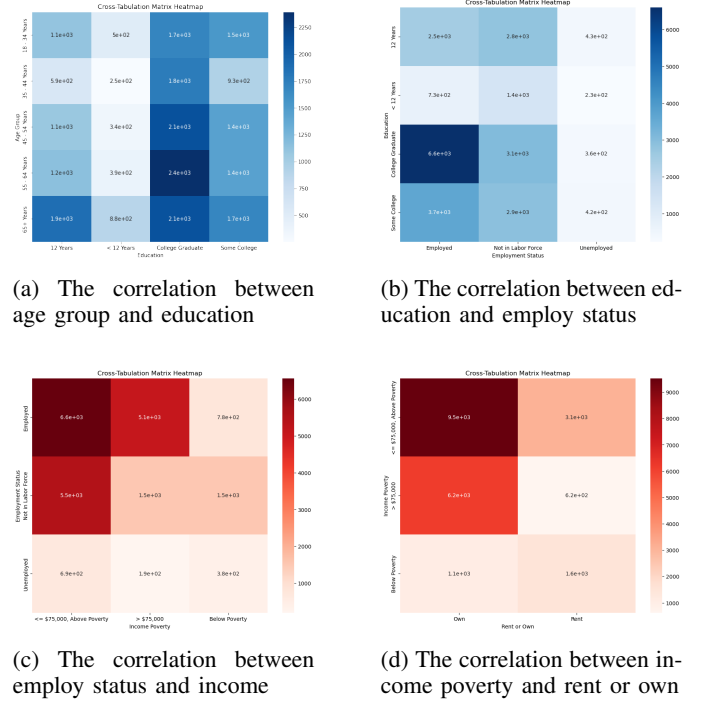


Fig. 3: Heatmaps for main categorical features

By the supportive theories from past literatures [10] [11] [12]

- Fill in the missing value of education based on age group
 - Generally, a relation trend for each age group and education level is College Graduate > Some College > 12 years > (<12 years)
 - It is clearly that in each age stage the College Graduate has the highest correlation. Consider fill any education level missing value based on the highest correlated education stage of each age group shown in the heatmap, which is "College Graduate"
- Fill in the missing value of employment_status based on education
 - 12 Years – Not in Labor Force
 - <12 Years – Not in Labor Force
 - College Graduate – Employed
 - Some College – Employed

c. Fill in the missing value of income_poverty based on employment_status

- For the missing values in income_poverty observing their employment_status value, all can be filled with \leq \$75,000, Above Poverty

d. Fill in the missing value of rent_or_own based on income_poverty

- Below Poverty – Rent
- ($>$ \$75,000) – Own
- \leq \$75,000, Above Poverty – Own

After completing the data filling process, both team members utilized a label encoder to convert categorical data into numerical form.

C. Classification

As this is a multi-label classification problem, in this stage, Yifeng opted to use Support Vector Machine (SVM) and Decision Tree (DT) models. SVM is an efficient linear classifier that can control overfitting through regularization parameters. The DT model is interpretable and can handle continuous, discrete, and mixed features due to its non-parametric nature.

Xiting, on the other hand, utilized the Random Forest and k-Nearest Neighbors (KNN) classifiers, as practiced in the experimental sessions. Random Forest offers the advantage of high accuracy as it combines predictions from multiple decision trees, resulting in more robust and reliable predictions compared to individual decision trees. It is particularly effective for datasets with a large number of input features or variables. When using the k-NN classifier, Xiting experimented with different values of k within a specified range to find the best performing value.

After conducting the initial pre-testing, it was observed that both the k-nearest neighbors (KNN) and random forest (RF) models exhibited significant advantages in terms of prediction results. Therefore, for further analysis, Yifeng decided to adopt the KNN model, while Xiting opted for the random forest model. They proceeded to develop and optimize their respective code implementations accordingly.

Both team members agreed to select the two best-performing models and run them ten times to obtain an average accuracy. The training and testing ratio was set at 7:3. Since the predictions were randomly assigned for each of the ten runs, no random state was specified.

IV. RESULTS FROM EACH OF THE STAGES

A. Data Analysis

Based on the results obtained from the EDA stage, the formula for merging behavioral feature values can be expressed as follows:

$$\begin{aligned} protection_h1n1 = & behavioral_avoidance * 0.18 + \\ & behavioral_wash_hands * 0.3 + \\ & behavioral_large_gatherings * 0.095 + \\ & behavioral_outside_home * 0.115 + \\ & behavioral_touch_face * 0.31 \end{aligned} \quad (4)$$

$$\begin{aligned} protection_seas = & behavioral_avoidance * 0.16 + \\ & behavioral_wash_hands * 0.29 + \\ & behavioral_large_gatherings * 0.13 + \\ & behavioral_outside_home * 0.105 + \\ & behavioral_touch_face * 0.315 \end{aligned} \quad (5)$$

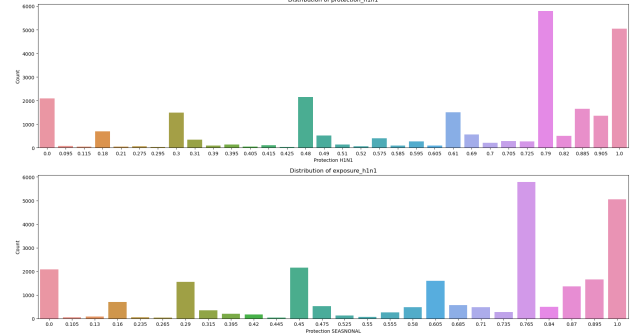


Fig. 4: Distribution for the new behavioral features

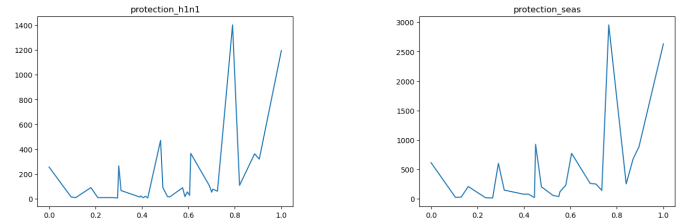


Fig. 5: Check the correlation for new behavioral features

For the new features related to opinions, the weights were calculated by setting new parameters and directly incorporating them.

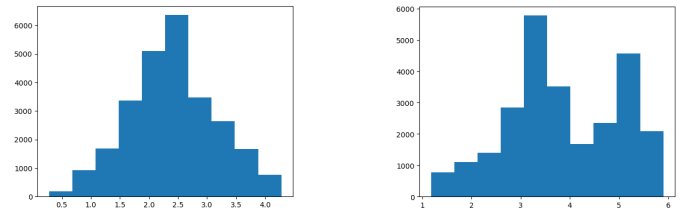


Fig. 6: Distribution for the new opinion features

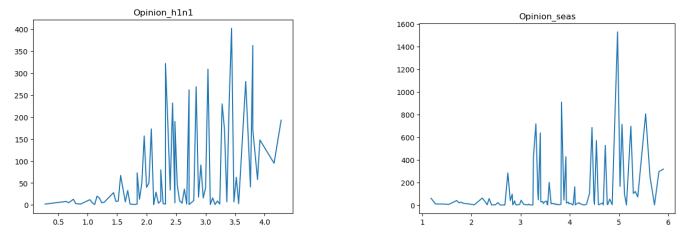


Fig. 7: Check the correlation for new opinion features

- After data fusion, the correlation coefficient has been improved, larger than the average value of each group of correlation coefficient before fusion. The new correlation coefficient are shown below:

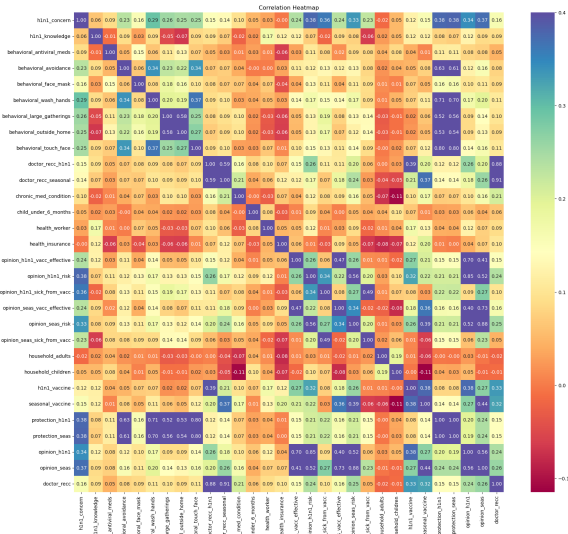


Fig. 8: The new heatmap after grouping features

B. Preprocessing

To explore the effectiveness of data dimensionality reduction operations, both team members performed feature engineering on the data obtained from different processing methods. Additionally, they retained a set of data without feature merging as a reference. In total, five CSV files were generated for the final prediction and evaluation, with the following contents:

- **DATASET 1: data_FE_AM.csv** - Mean and mode imputation, data dimensionality reduction completed.
- **DATASET 2: data_xFE_AM.csv** - Mean and mode imputation, data dimensionality reduction not performed.
- **DATASET 3: data_yesfusion**- Mode and correlation imputation, data dimensionality expansion, including both unmerged features and merged new features.
- **DATASET 4: data_nofusion** - Mode and correlation imputation, data dimensionality reduction not performed.
- **DATASET 5: data_pro** - Mode and correlation imputation, data dimensionality reduction completed.

C. Classification

The accuracy results obtained for the five datasets are presented in the following tables:

	DATASET1	DATASET2	DATASET3	DATASET4	DATASET5
KNN	0.797829	0.808187	0.814427	0.814177	0.80197
RF	0.818882	0.838225	0.827896	0.833833	0.825184

TABLE I: The accuracy rate for h1n1 prediction

	DATASET1	DATASET2	DATASET3	DATASET4	DATASET5
KNN	0.736179	0.730937	0.746911	0.734307	0.723075
RF	0.764395	0.77651	0.764695	0.760589	0.754561

TABLE II: The accuracy rate for seasonal prediction

V. DISCUSSION

After comparing the results, it can be observed that the random forest (RF) model performs slightly better than the k-nearest neighbors (KNN) model in terms of prediction accuracy. Based on the prediction accuracy, the accuracy of H1N1 is generally higher than that of seasonal vaccine. Additionally, it can be noted that the data obtained through mode and mean imputation without any feature merging achieves the best prediction performance when using the random forest model.

Upon evaluating the ROC curves and confusion matrices, it can be observed that both the KNN and RF models have ROC curves above the diagonal line, indicating that these two models outperform random guessing and are suitable prediction models. In terms of performance, the RF model exhibits a larger area under the ROC curve (AUC), indicating superior performance compared to KNN. This conclusion holds true across all datasets.

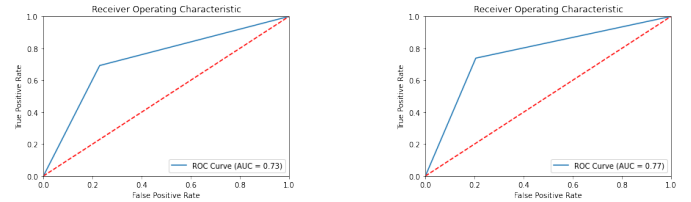


Fig. 9: ROC curve for KNN and Random forest model

In comparison to the results reported by Adalseno et al. [8], our prediction accuracy is lower by 0.3. However, when comparing our results to SHYAM R [citation], we observe that the accuracy is very similar. In fact, in Dataset 2, the performance of the Random Forest model for predicting H1N1 vaccination surpasses the accuracy of SHYAM R’s best-performing model, GradientBoostingClassifier [9].

VI. CONCLUSIONS AND RECOMMENDATION FOR FUTURE RESEARCH

Based on the data exploration and final predictions, it can be concluded that in this dataset, dimensionality reduction does not significantly improve the accuracy of predictions. Additionally, the introduction of newly merged features through

data augmentation has minimal impact on the original prediction results. On the contrary, maintaining the original feature dimensions and using more basic methods like mode and mean imputation yield higher results.

Based on the observed heatmaps, there is indeed a positive correlation between protective behaviors against virus infection and individuals' attitudes towards vaccination. The merged data distribution and correlation graphs of behavior and opinion also reveal some associations among individuals in the new feature values. Although the distribution trends may not be entirely consistent due to different calculation methods, some similar reference points can still be identified. Furthermore, due to the high correlation between opinions on H1N1 and seasonal vaccines, the behavioral features show almost no difference in their correlation with the two types of vaccines.

Future research will continue to focus on data dimensionality and explore deeper insights. The use of random forest models to train the data will be employed, and based on the results, importance scores for all features will be obtained. Feature selection will be performed by setting thresholds based on the importance scores. The selected features will then be used for further predictions.

Additionally, considering the selection of more models for training is recommended. By exploring multiple models, researchers can identify the models that best fit the data. Cross-validation techniques can also be employed to evaluate the performance of different models and compare their effectiveness.

REFERENCES

- [1] Gordon, B., Fennessy, C., Varma, S., Barrett, J., McCondochie, E., Heritage, T., Duroe, O., Jeffery, R., Rajamani, V., Earlam, K. and Banda, V., 2022. Evaluation of freely available data profiling tools for health data research application: a functional evaluation review. *BMJ open*, 12(5), p.e054186.
- [2] Peng, J., Wu, W., Lockhart, B., Bian, S., Yan, J.N., Xu, L., Chi, Z., Rzeszutarski, J.M. and Wang, J., 2021, June. Dataprep. eda: task-centric exploratory data analysis for statistical modeling in python. In *Proceedings of the 2021 International Conference on Management of Data* (pp. 2271-2280).
- [3] Rahm, E. and Do, H.H., 2000. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4), pp.3-13.
- [4] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J. and Liu, H., 2017. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6), pp.1-45.
- [5] Juszczak, P., Tax, D. and Duin, R.P., 2002, May. Feature scaling in support vector data description. In *Proc. asci* (pp. 95-102). Citeseer.
- [6] Mogan, C., Alvarez, J.M., Patro, G.K., Ruggieri, S. and Staab, S., 2022. Fairness implications of encoding protected categorical attributes. *arXiv preprint arXiv:2201.11358*.
- [7] Mahesh, B., 2020. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*. [Internet], 9, pp.381-386.
- [8] Adalseno (2021). *Flu-Shot-Learning-Predict-H1N1-and-Seasonal-Flu-Vaccines*. GitHub Repository. <https://github.com/adalseno/Flu-Shot-Learning-Predict-H1N1-and-Seasonal-Flu-Vaccines>
- [9] SHYAM R (2020). *Flu Shot Prediction -Complete EDA and HPO*. Kaggle. <https://www.kaggle.com/code/darkknight98/flu-shot-prediction-complete-eda-and-hpo/notebook>
- [10] Lenehan, M. E., Summers, M. J., Saunders, N. L., Summers, J. J. & Vickers, J. C. (2015). Relationship between education and age-related cognitive decline: a review of recent research. *Psychogeriatrics*, 15, 154-162. doi:10.1111/psy.12083
- [11] Stronks, K., van de Mheen, H., van den Bos, J., & Mackenbach, J. P. (1997). The interrelationship between income, health and employment status. *International Journal of Epidemiology*, 26(3). doi: <https://doi.org/10.1093/ije/26.3.592>
- [12] Tunstall, R., Bevan, M., Bradshaw, J., Croucher, K., Duffy, S., Hunter, C., Jones, A., Rugg, J., Wallace, A. & Wilcox, S. (n.d.). The links between housing and poverty: An evidence review.