COVID-19 Data Analysis using Machine Learning

Upasana Singh and Jatin Batra***

ABSTRACT

The impact of the COVID-19 pandemic has led scientists to produce a vast quantity of research aimed at understanding, monitoring, and containing the disease; however, it remains unclear whether the research that has been produced to date sufficiently addresses existing knowledge gaps. We usemachine learning techniques to analyze this massive amount of information at scale. In machine learning and its subset (Deep Learning) methods are employed in various applications to solve multiple problems that occur due to uncertainty. Most of the machine learning and deep learning algorithms are trained to address the supervised learning problem, where the algorithms know the prediction requirement.

The results demonstrate 93% overall accuracy in predicting the mortality rate. We used several machine learning algorithms including linear regression, Random Forest, Decision Tree, and K-Nearest Neighbor (KNN) to predict the number of covid cases.

Keywords: *supervised learning, unsupervised learning, model selection, linear regression, Random Forest, Decision Tree, and K-Nearest Neighbor (KNN).*

1.0 Introduction

In late 2019, a novel form of Coronavirus, named SARSCoV-2 (stands for Severe Acute Respiratory Syndrome Coronavirus 2), started spreading in the province of Hubei in China, and claimed numerous human lives [1]-[3].. In January 2020, the WorldHealth Organization (WHO) declared the novel coronavirus outbreak a Public Health Emergency of International Concern (PHEIC) [4][5]. In February 2020, WHO selected an official name, COVID-19 (stands for Coronavirus Disease 2019), for the infectious disease caused by the novel coronavirus, and later in March 2020 declared a COVID-19 Pandemic [5][6].

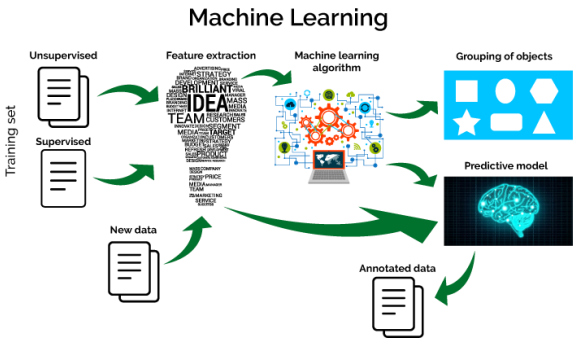
Coronavirus is a family of viruses that usually causes respiratory tract disease and infections that can be fatal in some cases such as in SARS, MERS, and COVID-19. Some kinds of coronavirus can affect animals, and sometimes, on rare occasions, coronavirus jumps from animal species into the human population. The novel coronavirus might have jumped from an animal species into the human population, and then begun spreading [7]. A recent study has shown that once the coronavirus outbreak starts, it will take less than four weeks to overwhelm the healthcare system. Once the hospital capacity gets overwhelmed, the death rate jumps [8].. The proposed system includes a set of algorithms for preprocessing the data to extract new features, handling missing values, eliminating redundant and useless data elements, and selecting the mostinformative features[9]. After preprocessing the data, we use machine learning algorithms to develop a predictive model to classify the data, predict the medical condition, and calculate the probability of number of cases in upcoming days[10].

2.0 Methods

**Corresponding author; Assistant Prof, CS & IT Department, Trinity Institute of Professional Studies, Delhi India. (Email: upasana.tips2018@gmail.com)*
Student, Department of IT, Trinity Institute of Professional Studies, Dwarka, New Delhi, Delhi, India, (Email: jathunnybatra@gmail.com).

2.1 Dataset

In this paper, we used a dataset of more than 117,000 laboratory-confirmed COVID-19 patients from 76 countries around the world including both male and female patients with an average age of 56.6.This dataset is a collection of the COVID-19 data maintained by Our World In Data. It and includes data on confirmed cases, gdp rate, deaths, hospitalizations, and testing, as well as other variables of potential interest. At the data cleaning stage, we removed useless and redundant data elements such as data source, admin id, and admin name. Then, Data imputation techniques were used to handle missing values. After analyzing the data, we found out that 74% of patients were recovered from COVID-19. To have an accurate and unbiased model, we made sure that our dataset is balanced. A balanced dataset with equal observations for both recovered and deceased patients was created to train and test our model. The data observations (patients) in the training dataset have been selected randomly and they are completely separate from the testing data. Figure 1 shows a machine



learningarchitectur
Fig1: Machine Learning Architecture

2.2 Feature Selection

The outcome label contained multiple values for the patient's health status. We considered patient that discharged from hospital or patients in stable situation with no more symptoms as recovered patients. A total of 80 features were extracted from symptoms and doctors' medical notes about the patient's health status. The next step is feature selection. The primary purpose of feature selection is to find the most informative features and eliminate redundant data to reduce the dimensionality and complexity of the model. We used univariate and multivariate filter method and wrapper method to rank the features and select the best feature subset.

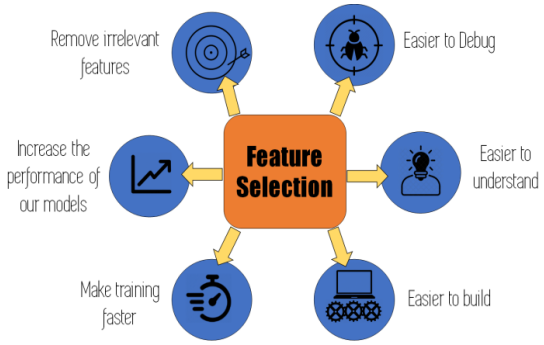


Fig 2:

It Select a subset of input features from the dataset.

- Unsupervised: Do not use the target variable (e.g. remove redundant variables).
- Correlation

- Supervised: Use the target variable (e.g. remove irrelevant variables).
- Wrapper: Search for well-performing subsets of features.
- RFE
- Filter: Select subsets of features based on their relationship with the target.
- Statistical Methods
- Feature Importance Methods
- Intrinsic: Algorithms that perform automatic feature selection during training.
- Decision Trees
- Dimensionality Reduction: Project input data into a lower-dimensional feature space.

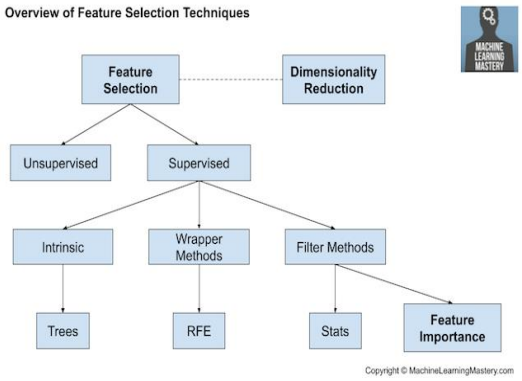


Fig3:

2.3 Predictive Algos

After selecting the best feature subset, we used various machine learning algorithms to build a predictive model. In this research, we used different algorithms including linear regression, Random Forest, Decision Tree, and K-Nearest Neighbor (KNN).

LinearRegression is supervised machinelearning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).

The Random Forest algorithm is an ensemble learning method combined of multiple decision tree predictors that are trained based on random data samples and feature subsets.

A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g.whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels.

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

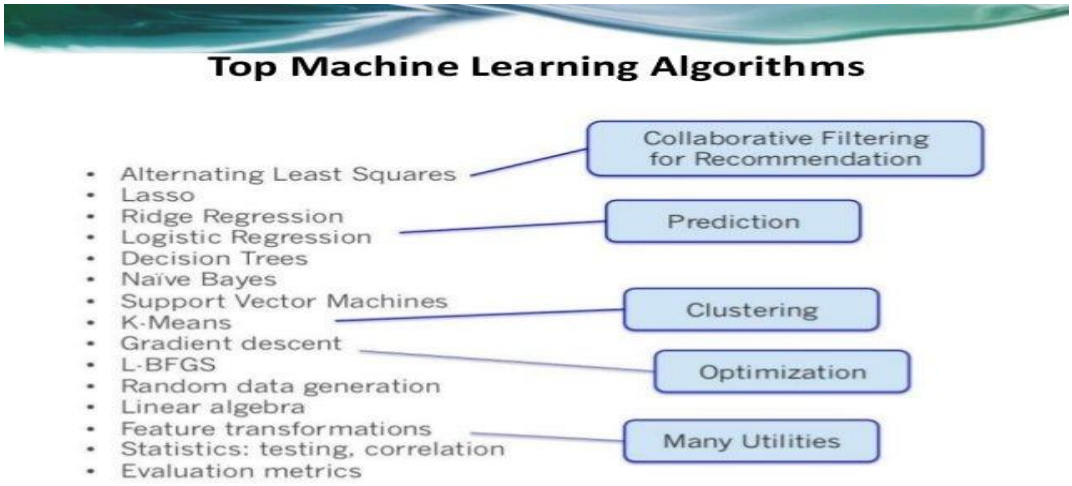


Fig 4:

3. ML Principle

Every learning process, deep or not, consists of two phases: the estimation of unknown dependencies in a system from a given data set (input) and the use of estimated dependencies to predict new outputs of the system. In this Subsection, we analyses the most common techniques used in both phases.

The input of a ML process is a set of instances. These instances are the things that can be classified, associated, or clustered. Each instance is an individual, i.e., independent example of the concept that must be learned.

Machine Learning (ML) platform provider is to use all the tools at our disposal to help our clients improve the quality and effectiveness of their marketing. At the same time, we have collated a set of principles that we will use as a code of ethics to abide by when developing our ML.

4.0 Result and Evaluation

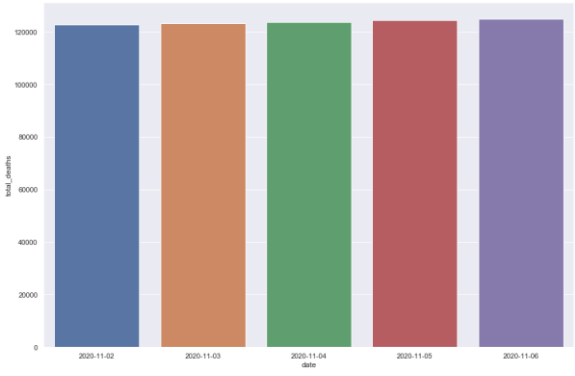


Fig 6: This graph represents total deaths w.r.t date in india.

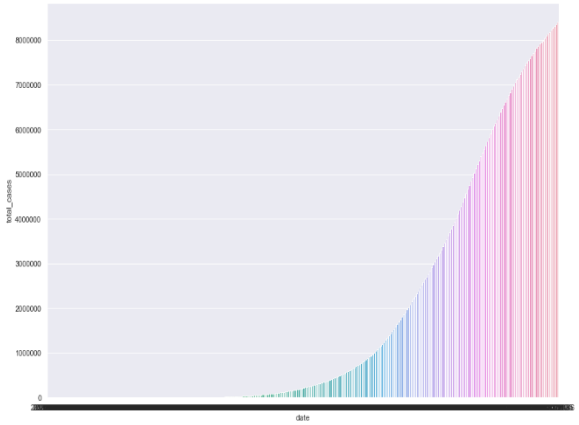


Fig 7:This graph represent total cases w.r.t date in india.

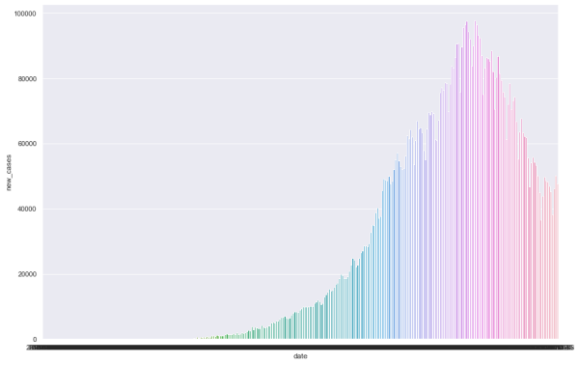


Fig 8: This graph represent new cases w.r.t date in india.

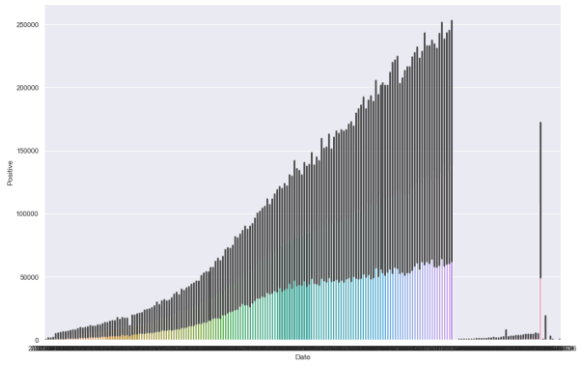


Fig 9: This graph represent total positives cases w.r.t date in india

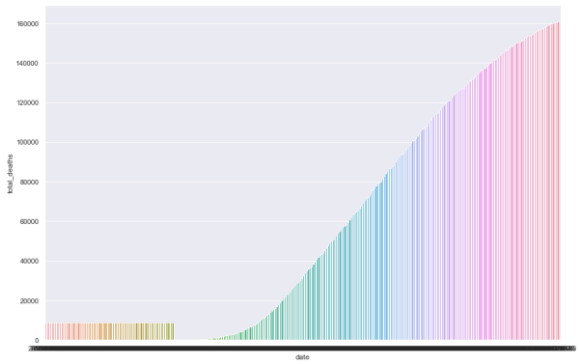


Fig 10: This graph represent total total deaths in brazil wrt date.

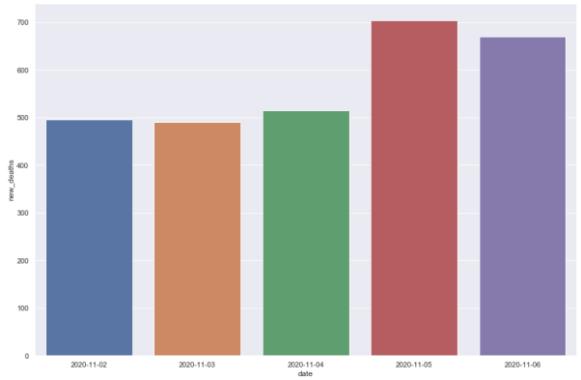


Fig 11: This graph represent new deaths in india w.r.t date.

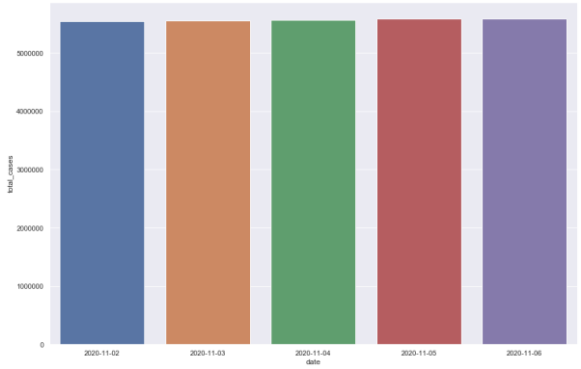


Fig 12: This graph represent total cases in brazil w.r.t date.

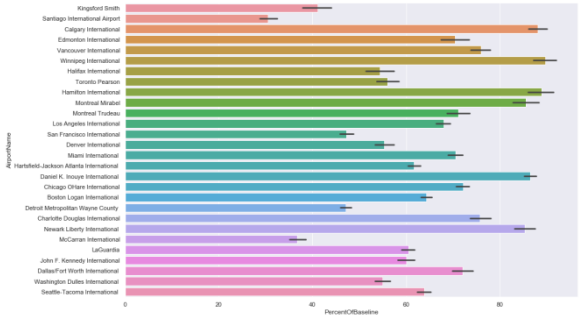


Fig 13: This graph shows the percent of world's airport baseline effected by covid.

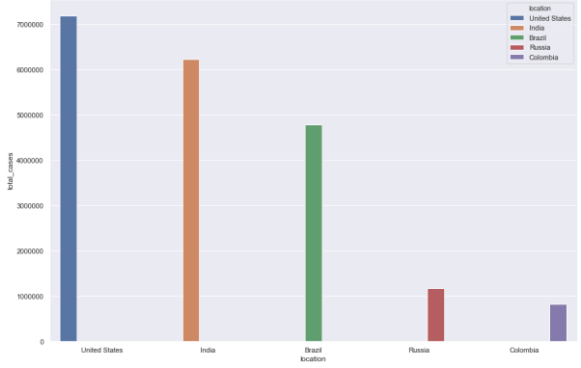


Fig 14: This graph represent total cases in top 5 countries .

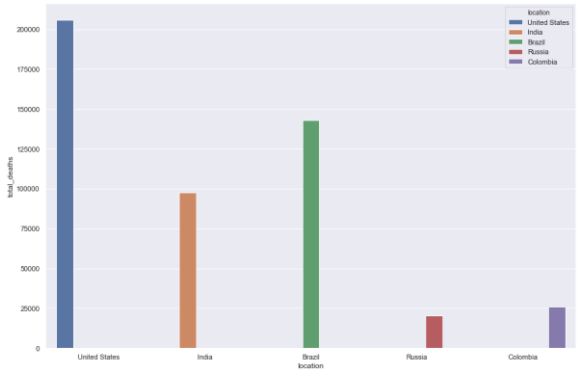


Fig 15: This shows total deaths in top 5 countries.

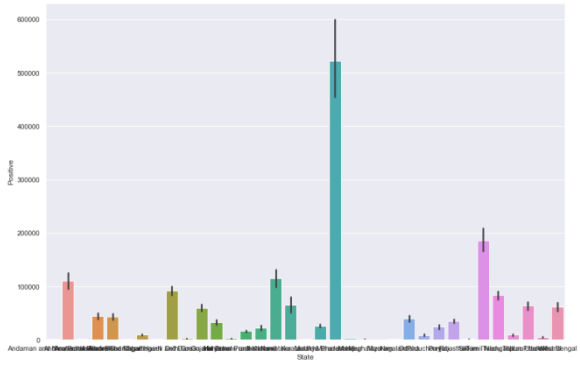


Fig 16: This represents statewise total positive cases.

References

Todd Ellerin, HumaFarid, Douglas Krakower, Howard E. LeWine, Claire McCarthy, Babar Memon, John Sharp, Robert H. Shmerling, Jacqueline Sperling, Harvard Health Publishing Coronavirus Resource Center Experts.

Li, Q. et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. N. Engl. J. Med. NEJMoa2001316, https://doi.org/10.1056/NEJMoa2001316 (2020).

Xu, B., Gutierrez, B., Mekaru, S. et al. Epidemiological data from the COVID-19 outbreak, real-time case information. Nature Sci Data7,106(2020).ttps://doi.org/10.1038/s41597-020-0448-0 In: Nature.

M. Pourhomayoun, E. Nemati, M. Sarrafzadeh, B. Mortazavi, ContextAware Data Analytics for Activity Recognition.

C. Cortes, V. Vapnik, in Machine Learning, pp. 273–297 (1995).

V. Vapnik, The Nature of Statistical Learning Theory [17]L. Breiman, "Random Forests". Machine Learning, 2001.

Yoon, Y.; Cho, J.H.; Yoon, G. Non-constrained blood pressure monitoring using ECG and PPG for

personal healthcare. J. Med Syst. 2009, 33, 261–266. [CrossRef]

Dick, R.S.; Steen, E.B.; Detmer, D.E. The Computer-Based Patient Record: An Essential Technology for Health Care; National Academies Press: Washington, DC, USA, 1997.

Zhuang, Z.Y.; Churilov, L.; Burstein, F.; Sikaris, K. Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. Eur. J. Oper. Res. 2009, 195, 662–675. [CrossRef]

Huang, M.J.; Chen, M.Y.; Lee, S.C. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Expert Syst. Appl. 2007, 32, 856–867. [CrossRef]

Murdoch, T.B.; Detsky, A.S. The inevitable application of big data to health care. JAMA 2013, 309, 1351–1352. [CrossRef]

Wu, X.; Zhu, X.; Wu, G.Q.; Ding, W. Data mining with big data. IEEE Trans. Knowl. Data Eng. 2014, 26, 97–107. [CrossRef]

Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. Data Mining: Practical Machine Learning Tools and Techniques; Elsevier: Amsterdam, The Netherlands, 2016.

Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. Science 2015, 349, 255–260. [CrossRef] [PubMed]

Kononenko, I. Machine learning for medical diagnosis: History, state of the art and perspective. Artif. Intell. Med. 2001, 23, 89–109. [CrossRef]

Sriram, T.V.S.; Rao, M.V.; Narayana, G.V.S.; Kaladhar, D.S.V.G.K. A Comparison and Prediction Analysis for the Diagnosis of Parkinson Disease Using Data Mining Techniques on Voice Datasets. Int. J. Appl. Eng. Res. 2016, 11, 6355–6360.

Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning; MIT Press: Cambridge, MA, USA, 2016.

Dixon-Woods, M.; Bonas, S.; Booth, A.; Jones, D.R.; Miller, T.; Sutton, A.J.; Shaw, R.L.; Smith, J.A.; Young, B. How can systematic reviews incorporate qualitative research? A critical perspective. Qual. Res. 2006, 6, 27–44. [CrossRef]

Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 2015, 13, 8–17. [CrossRef] [PubMed]

Hartigan, J.A. Clustering Algorithms; Wiley: Hoboken, NJ, USA, 1975; Volume 209.

Birant, D.; Kut, A. ST-DBSCAN: An algorithm for clustering spatial-temporal data. Data Knowl. Eng. 2007, 60, 208–221. [CrossRef]

Kohonen, T. The self-organizing map. Neurocomputing 1998, 21, 1–6. [CrossRef] [23]. Dara, R.; Kremer, S.C.; Stacey, D.A. Clustering unlabeled data with SOMs improves classification of labeled

real-world data. In Proceedings of the IEEE International Joint Conference on Neural Networks IJCNN'02, Honolulu, HI, USA, 12–17 May 2002; Volume 3, pp. 2237–2242, doi:10.1109/IJCNN.2002.1007489.

Wang, B.; Mezlini, A.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods 2014, 11. [CrossRef]

Nguyen, T.; Tagett, R.; Diaz, D.; Draghici, S. A novel approach for data integration and disease subtyping. Genome Res. 2017, 27, 2025–2039. [CrossRef]

Ramazzotti, D.; Lal, A.; Wang, B.; Batzoglou, S.; Sidow, A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nat. Commun. 2018, 9, 4453. [CrossRef] [PubMed]

Nissim, N.; Boland, M.R.; Tatonetti, N.P.; Elovici, Y.; Hripcsak, G.; Shahar, Y.; Moskovitch, R. Improving condition severity classification with an efficient active learning based framework. J. Biomed. Informatics 2016, 61, 44–54. [CrossRef] [PubMed]

Nissim, N.; Shahar, Y.; Elovici, Y.; Hripcsak, G.; Moskovitch, R. Inter-labeler and intra-labeler variability of condition severity classification models using active and passive learning methods. Artif. Intell. Med. 2017, 81, 12–32. [CrossRef] [PubMed]

Cortes, C.; Vapnik, V. Support-vector networks. Mach. Learn. 1995, 20, 273–297. [CrossRef]

Quinlan, J.R. Induction of decision trees. Mach. Learn. 1986, 1, 81–106. [CrossRef]

Quinlan, J.R. C4. 5: Programs for Machine Learning; Springer: Berlin, Germany, 1993.

Fix, E.; Hodges, J.L. Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties; Technical Report; DTIC Document; Defense Technical Information Center: Fort Belvoir, VA, USA, 1951.

McCallum, A.; Nigam, K. A comparison of event models for naive bayes text classification. In Proceedings of the AAAI-98 Workshop on Learning for Text Categorization, Madison, WA, USA, 22–27 July 1998; Volume 752, pp. 41–48.

Heckerman, D.; Horvitz, E.; Nathwani, B.N. Toward Normative Expert Systems: Part I. The Pathfinder project. Methods Inf. Med. 1992, 31, 90–105. [CrossRef]

Heckerman, D.; Nathwani, B.N. Toward Normative Expert Systems: Part II. The Pathfinder project. Methods Inf. Med. 1992, 31, 106–116. [CrossRef]

Lawson, C.L.; Hanson, R.J. Solving Least Squares Problems; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1995.

Kleinbaum, D.G.; Klein, M. Analysis of matched data using logistic regression. In Logistic Regression; Springer: Berlin, Germany, 2010; pp. 389–428. [CrossRef]

Miao, D.Q.; Zhao, Y.; Yao, Y.Y.; Li, H.X.; Xu, F.F. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model. Inf. Sci. 2009, 179, 4140–4150. [CrossRef]

Rokach, L.; Maimon, O. Data Mining with Decision Trees: Theory and Applications; World Scientific Publishing: Singapore, 2014.

Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. Classification and Regression Trees; Chapman and Hall/CRC: Boca Raton, FL, USA, 1984.

Serrano, K.J.; Yu, M.; Coa, K.I.; Collins, L.M.; Atienza, A.A. Mining health app data to find more and less successful weight loss subgroups. J. Med Internet Res. 2016, 18. [CrossRef]