Software Engineering for Machine Learning in Health Informatics

Mohammed Moreb (✉ moreb.mohammad@gmail.com)

Altınbaş Üniversitesi <https://orcid.org/0000-0001-8146-6918>

Oguz Ata

Altınbaş Üniversitesi

Research article

Keywords: Health Dataset Analysis, Machine Learning, Healthcare, Methodology, Software Development, Software Engineering

Posted Date: November 25th, 2019

DOI: <https://doi.org/10.21203/rs.2.17747/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published on January 29th, 2020. See the published version at <https://doi.org/10.1109/ACCESS.2020.2970178>.

Abstract

Background We propose a novel framework for health Informatics: framework and methodology of Software Engineering for machine learning in Health Informatics (SEMLHI). This framework shed light on its features, that allow users to study and analyze the requirements, determine the function of objects related to the system and determine the machine learning algorithms that will be used for the dataset.

Methods Based on original data that collected from the hospital in Palestine government in the past three years, first the data validated and all outlier removed, analyzed using develop framework in order to compare ML provide patients with real-time. Our proposed module comparison with three Systems Engineering Methods Vee, agile and SEMLHI. The result used by implement prototype system, which require machine learning algorithm, after development phase, questionnaire deliver to developer to indicate the result using three methodology. SEMLHI framework, is composed into four components: software, machine learning model, machine learning algorithms, and health informatics data, Machine learning Algorithm component used five algorithms use to evaluate the accuracy for machine learning models on component.

Results we compare our approach with the previously published systems in terms of performance to evaluate the accuracy for machine learning models, the results of accuracy with different algorithms applied for 750 case, linear SVG have about 0.57 value compared with KNeighbors classifier, logistic regression, multinomial NB, random forest classifier. This research investigates the interaction between SE, and ML within the context of health informatics, our proposed framework define the methodology for developers to analyzing and developing software for the health informatic model, and create a space, in which software engineering, and ML experts could work on the ML model lifecycle, on the disease level and the subtype level.

Conclusions This article is an ongoing effort towards defining and translating an existing research pipeline into four integrated modules, as framework system using the dataset from healthcare to reduce cost estimation by using a new suggested methodology. The framework is available as open source software, licensed under GNU General Public License Version 3 to encourage others to contribute to the future development of the SEMLHI framework.

1. Background

The field of Health Informatics (HI) aims at providing a large-scale linkage of disparate. Normally, the healthcare dataset is found to be incomplete and noisy; as a result, reading data from dataset linkage traditionally failed within the discipline of software engineering. Machine learning (ML) is a rapidly mature branch of computer science, since it can store data in large scale. Many ML tools that can be used to analyze the data, and reach the knowledge that can improve the quality of work for both staff and doctors; however, as for developer, there is no methodology that can be used until now. Regarding

software engineering, there is a lack of approaches to evaluate which software engineers tasks are better performed by automation, and which require human involvement, or human in the loop approaches [1].

Recently, a set of frameworks that are used to develop the analysis of data, such as Win- CASE [2], and SAM [3], while the market has vast analytical tools of data [4], that can discover interesting patterns, and hidden relationships to support decision makers. BKMR [5] used R package as a statistical approach with health effects to estimate the multivariable exposure-response function.

Augmentor [6] include as python image library for augmentation, while on visualization of medical treatment plans and patient data, used CareVis [7] as it's designed for this task. Other applications require a visual interface used COQUITO [8], for health- care data analytics widely known 3P tools [9] used. Many simple application, such as WEKA [10] that provides a GUI to many machine learning algorithms, and apache Spark [11] are used for cluster computing framework. Consequently, table 1 summarizes the main tools that used for big data for analytics according to task.

Software Engineering for Machine Learning Applications (SEMLA) [12] discusses challenges, new insights, and practical ideas regarding the engineering of ML and artificial engineering (AI). Based on software development, ML algorithms in clinical genomics generally come in three main forms: supervised, unsupervised and semi supervised [13]. Interflow system requirement analysis (ISRA) [14] is used to determine system requirements.

TABLE 1 Big Data Analytics Tools according to task.

Big Data Task	Tools
Data Languages	R, Python
Data Integration	Blockspring, Pentaho
Data Collection	Import.io
Data Cleaning	OpenRefine, DataCleaner
Data Analysis	Qubole, BigML, Statwing
Data Visualization	Tableau, Silk, Charito, Plot.ly
Data Mining	Rapid Miner, Weka, Oracle data mining, IBM Modeler.
Data Storage and Management	Hadoop, MongoDB, Talend.

Electronic healthcare (eHealth) frameworks [15] have replaced traditional medical frameworks. e-Health frameworks tend to improve mobile healthcare (mHealth) and enable patient-to-physician and patient-to-patient dealings for better healthcare and quality of life (QoL).

Machine learning is a field of software engineering that frequently utilizes factual procedures to enable PCs to "learn", with information from saved dataset. Unsupervised or information mining concentrates more on exploratory information investigation, and is known as learning supporting by data analytics. Patient lab test queue management and wait time prediction form a challenging and complicated job. Because each patient might require different phases operations (tasks) such as a check-up, various tests, e.g., a sugar level or blood test, X-rays or surgery. Each task can have different Medical tests from 0 to N for each patient according to his condition.

In this article based on grounded theory methodology [16], the researchers proposed a novel methodology, SEMLHI, to develop the framework by defining the research problem and methodology for the developers. SEMLHI framework includes a theoretical framework to support research and design activities that incorporate existing knowledge. SEMLHI framework is composed into four components that help developers observe health application flow from the main module, into sub modules, to run and validate specify tasks. This will enable multiple developers to work on different modules of the application simultaneously. The main contributions for SEMLHI framework are represented in the introduction of SEMLHI methodologies that enable researchers and developer to analyze health informatics software, from an engineering perspective. SEMLHI framework; supports the methodological approach to conduct research, on health informatics. It also supports the structure that presents a common set of ML terminology to use, compare, measure, and design software systems in the area of health. As a result, this will create a space where SE and ML experts could work on specific methodological approach for Health informatics software development teams to integrate the ML model lifecycle. Our methodology will apply for current systems, or in development of new system that use ML module for current system, that can be used in regular updates to add data to system, irregular updates, to add new feature such as new version of ICD diagnosis codes, minor model improvements for bug fixes, new functionalities required by client, New hardware or architectural constraints.

2. Methods

Dataset and preprocessing

The applications delivery of applied Machine Learning (ML) models in healthcare is often hampered by the existence of the isolated product deployments with poorly developed architectures and limited or non-existent maintenance plans, The "Translating Research into Agile Development" (TRIAD) method relies [17], and present five step method for designing a tailored EHR tool, while SEMLHI framework will help developer to use stander methodology for developing health systems, develop software supported by machine learning, at this state the dataset applied from Palestine hospitals.

In such a case, patient required to do more than one tests. In this article, we focus on helping patients and doctors to complete their treatment tasks by predictable test result based on ICD-10 [18], and helping hospitals save time and efforts for medical tests.

SEMLHI models and methodology will be presented by including new software systems connecting to real data set and present knowledge from data and using ML algorithms. The dataset case studies discussed in this article are set within the context of the Palestine Hospital and centers. Three hospitals

and nine medical centers were used for our data set. Furthermore, the data collection was conducted during the last two years, 458k Patients were identified with patient no. Overall, for pmc dataset included 141k patient with 1.63% missing, mean 1.08m, std dev 554k, min 10000, max 1.04M, For age label 141k patient with 1.63% missing, mean 32.24, std dev 26.25, min 0, max 88, with median equal to 29.

We use original data as selected dataset to develop a patient predicting test lab results predicting model, using SEMLIH framework, the realistic patient data are analysed carefully and rigorously based on important parameters such as patient treatment start time, end time, age, and detail treatment content for each different task. We identify the lab test require for patient based on their conditions and operations performed during treatment. The patient data has used only codified variables, including ICD 10 (International Classification of Diseases) codes, procedure codes, and medication orders often reduced to smaller subsets.

Available Features of Patients:

Patient dataset include 457914 cases and nine tables. Each table has different features. Some features connected with other table to build dataset describing the main attribute as mentioned in next section.

Monitoring of Patients, used collected electrocardiograph data to focused on five most common heartbeat cases in MIT-BIH arrhythmia database: (i) NORM (ii) LBBB (iii) RBBB (iv) VPC (v) APC

$E = \{e_1, e_2, e_3, \dots, e_n\}$, where E: a heartbeat class, n: number of heartbeat classes.

Laboratory test data, have Number of cases (columns) 200k cases; each case have basic attribute, such as, patient no, gender, age, department, diagnosis code, description, and date of lab test. The data represent as $L = \{l_1, l_2, l_3, l_4, \dots, l_n\}$, where l: the item of lab test reports, and n: number of items in the lab test reports.

Conceptual Framework and Design

3.1 SEMLHI methodology

SEMLHI methodology is used in software development in the Health area. For traditional application, development process includes many methodologies, such as Waterfall Methodology, Spiral Methodology, and Agile Methodology that can be used to define and develop the software. Table 2 illustrates the results of comparison between our methodology compared and the methodology of Vee [19] and agile [20]. SEMLHI's framework Methodology describes in details the process that is used when developing health software and the mechanism to integrate and use ML algorithms with the development software. SEMLHI methodology will give developer's new road map to design health application with system functions and implementing of the software. This framework includes 10 stages starting from defining the problem until reaching the stage of development and ending by the results that explain the following next section.

TABLE 2 The Comparison between Three Systems Engineering Methods Vee, agile and SEMLHI that proposed on this article.

	SEMLHI	Vee	Agile
Flexibility	very high	rigid	very high
Emphasis	risk	specification	customer
Logic	depth first	breadth first	depth first
Assumption	independent iteration	stable info	independent iteration
Scope	medium and large	large and complex	small
Iteration	very quick	slow	very quick
Delivery	One shot delivery	One shot delivery	incremental delivery

To develop HI system the first step is to design (encode data, define outlier and cleaning up the data), implement (Verification & Validation), maintain defined Workflows, structure Information, provide Security and privacy, test the performance, and then reuse software applications. Records in most data sets in HI are weakly structured and non-standardized. To apply ML for HI system a set, of patterns need to be used by algorithm to predict and visualize for ML algorithm and generate knowledge. The main patterns that are used in our framework are: Geographic location, Patient records, departments and hospitals, surgical history, obstetric history, family history, habits, immunization, assessment and plan, and test results.

Results And Discussion

Currently, the topic of health informatics is very vital for research, we propose a novel framework SEMLIH for health informatic system with machine learning. Next section describe the details for our propose framework.

3.2 SEMLHI framework

SEMLHI frameworks are specifically geared toward facilitating the development of software applications and include components that facilitate the analysis of the health data set. (Figure 1) summarizes the proposed framework as a conceptual framework in addition to the mechanism used to interact with operating system and hardware. Many users will work directly as developers or system analysts with approach framework or indirect by using result. As for users of Software Engineering, proposed frameworks interact with operating system components that are used by the framework and further how all software manage devise hardware with the main system devise that use by the framework.

Our framework is composed into four components or modules (software, machine learning model, machine learning algorithms, and health informatics data. (Figure 2) implements how each module interacts with the all modules to work as a framework.

3.2.1 ML Algorithms

Machine Learning algorithms, use to compute the parameters that might define a model [21], optimize its network topology and improve the system convergence without losing information. As supervised learning, k-Nearest Neighbors (KNN) [22] can be used for classification & predictive problems, KNN makes decisions based on the dominant categories of k objects, rather than a single object category. Figure 3 identify the most machine learning algorithms that used for health classification.

As all the data in our sample of dataset that prepared using SEMLHI framework, the output method will be supervised "label data" for this the KNN algorithm with multi label and evaluate our result (knn used for supervised learning, Kmean unsupervised). k-means can used for dataset that include million labeled, approximate nearest neighbours (ANN) which is usually 10x - 100x faster than KNN support-vector machines (SVM) works differently, and it is good and fast solution for many problems and will outperform k-NN almost, (Figure 4) show that Logistic Regression have high accuracy compared with expected and real predictions.

Based on original data that collected from the hospital in Palestine government in the past three years, five algorithms have been used to predict the lab test using Machine learning Algorithm (MAM) component on SEMLIH framework. ML approaches and algorithms [23] can achieve better performance than expert-selected. MLA uses two types of techniques, supervised learning (we have labeled input data to predict future outputs), and unsupervised learning (unlabeled data for input data). As for MLA module, first which techniques to use, then select the most suitable algorithms to use based on mathematical selection related to criteria, and then from list for algorithms, table 2 show the results of accuracy with different algorithms applied (KNeighbors Classifier, Linear SVC, Logistic Regression, Multinomial NB, and Random Forest Classifier), for 750 case, linear SVG have about 0.57 value compared with other algorithms, for clustering need to calculates the distance d between the two objects x and y by comparing the values of their n features by applying Minkowski metric.

Table 3: Evaluate the accuracy for machine learning models

Algorithm Name	Accuracy
KNeighbors Classifier	0.487694
Linear SVC	0.564566
Logistic Regression	0.560412
Multinomial NB	0.517013
Random Forest Classifier	0.488955

In supervised learning, the data set contains 'n' of rows (cases), each case needs to be evaluated using function $f: A \rightarrow B$ to compare with label A or label B according to 'f' function by evaluate E and compare to learn from training set of n. f have set of $n(d)$. Unsupervised learning, the data haven't labeled, apply in analysis, dimensional reduction, this type [24], use training set (t) which include n of objects $(t) = x_i \in A : 1 \leq i \leq 0$ that can be a category to n of class $C1, ck \in A$ by applying algorithm f in the evaluation phase to set which ck for input $x \in Cj : 1 \leq j \leq k$.

3.2.2 Machine Algorithm Model:

Machine learning helps us to extract useful features from dataset to solve, or predict health-related events [25]. MAM component include five sub-modules, read the data, prepare the data, train the model, test and evaluate the model, and predict new data. (Figure 5) describe the sequence of this stages.

The challenge for this component is to use the right type of algorithm, which optimally solves the dataset, while avoiding high bias or high variance. MAM main component to analysis the dataset by

set of conditions, as if the dataset > 50 samples, and have label then classification algorithms will select, if not cluster algorithms will apply if dataset need to predict quantity regression algorithms will be used, if not dimensional reduction will apply, next diagram show in details which algorithms to select for MAM Components.

3.2.3 Health Informatics Data

In ML, the data is essential; while the methods to apply to present and visualize knowledge are the most important step. Our dataset sample contains ten columns with 50k rows (cases) to use dataset on Health Informatics Data (HID) algorithms, require transformed into numerical features. Other data contain missing, duplicating or null value such as negative ages or extremely large integers could negatively affect the performance of our MLA, (Figure 4) describe main roles to detect the methodology to use classification, clustering, regression and reduction.

HID use data source and dictionary translate for label encoding to converting each value in a column to a number to reduce misinterpreted data used by Bayesian inference. Node Identifier work to analysis data as common process with determine patterns using patient-specific research identifier, dataset usually required multiple records from the same patient identified as being related in the de-identified database, for Outlier HID used a set of methods to analysis finding hidden groups to removing outliers, and in advance step need to calculate the outlier values for data, which appear to be erroneous and crop it from dataset. (Figures 8), shows that logistic regression has high accuracy compared with expected and real predictions.

To predict disease, we use icd-10 with multi label, as each patient have ICD code in health records can affect all regions of the retina, though, there is currently no classification system [23], to distinguish anterior (peripheral), and posterior (macular). We hypothesize that these classifications are characterized by D and refractive features, highlighting the disparity in types of disease.

Used collected electrocardiograph data to focused on D most common Diagnosis cases in Diagnosis Code database: $D = \{d_1, d_2, d_3, \dots, d_n\}$, d: a disease that applicable to diagnosis code, n: number of Disease classes using k mean algorithm with multi label, figure 7 describe pseudo code for k-nearest neighbor algorithm for Multilevel Learning.

Read the data

In Summary, to rad the data, from data source, such as CSV file, or other source available, the algorithm will automatic remove missing values, cleaning the data to remove noise, and other date in text description, add category column to encoding the test result name, as applied on our research, the result (720, 27), as we have 720 case categories, represented by 27 laboratory test, this step generate new dataset that include 18 columns, and 720 rows, figure 8, summarize the age with category feature.

Select ML algorithm and domain

To solve the medical data, as Multi-Class need to use supervised classifiers as all input data in training set we know, we can train supervised classifiers to train unseen patient description text to predict the

“Diagnose Code” Category by applying set of algorithms. This phase include select the automatic algorithm accuracy based on compare all algorithms to use.

3.2.4 Software

Software module includes sub class include reuse, performance, test, privacy, and security. For software testing main point to verify that the code is running correctly by exercising the code under known conditions and checking that the results are as expected [26]. This class used to test the resources use from memory, or CPU from the application, performance issues by first measure them, then profile the code, then optimize for that using the benchmark is the best choice to compare result to improve the performance for optimization. Code smells [27], found Genetic Algorithms used by 22.22% as most commonly machine learning techniques.

[Due to technical limitations, the formula could not be displayed here. Please see the supplementary files section to access the formula.] (1)

multi-labelled classification, a prediction containing a subset of the actual classes should be considered better than a prediction that contains none of them, i.e., predicting two of the three labels correctly this is better than predicting no labels at all. To measure a multi-class classifier a misclassification using micro-averaging and macro-averaging [16].

Conclusions

This research introduced a new methodology, that can develop health informatics application using machine learning. Our methodology used the grounded theory methodology to develop SEMLHI framework. Developers use SEMLHI methodology to analyzing and developing software for the HI model and create a space in which SE and ML experts could work on the ML model lifecycle.

Proposed framework includes a theoretical framework to support research and design activities that incorporates existing knowledge. Our work introduces a new approach form clustering and classification for ML in HI.

SEMLHI methodology includes seven-phase, designing (encode data and Define outlier and cleaning up the data), implementing (Verification & Validation), maintaining and defined Workflows, structured Information, security and privacy, testing and performance, and reusing software applications. SEMLHI framework includes four modules that organize the tasks for each module, and introduce a SEMLHI Methodological that enable researchers and developer to analyze health informatics software from an engineering perspective. The ultimate goal from a SEMLHI Methodological is to define a standardized methodology for software development in the Health area and include all stages from defining the problem until developing the application and get the result with the test stage.

Abbreviations

SEMLHI: Software Engineering for machine learning in Health Informatics; SE: software engineering; ML: machine learning; HID: Health Informatics Data; HI: Health Informatics; MAM: Machine Algorithm Model.

Declarations

Acknowledgment

We thank Health Minister of the State of Palestine, Dr. Jawad Awwad for the efforts has been made to allow us accessing Palestinian data set for patients and for all the teams that supported us during last two years for feedback that greatly improved the manuscript.

Authors' contributions

Moreb and Ata contributed equally to the work. Moreb Designed the study, prepare and analyzed the data, develop the framework using python, Ata test the framework, analysis the methodology steps, with evaluation framework. All authors have read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The data are available upon request to the corresponding author after signing appropriate documents in line with ethical application and the decision of the Ethics Committee.

Ethics approval and consent to participate

The research meets all applicable standards with regard to the ethics of experimentation and research integrity, and the following is being certified/declared true. As an expert scientist and along with co-authors of concerned field, the paper has been submitted with full responsibility, following due ethical procedure, and there is no duplicate publication, fraud, plagiarism, or concerns about animal or human experimentation.

Competing interests

All authors report no conflicts of interest.

Consent for publication

Not applicable

Author details

Mohammed Moreb, is a PhD Researcher of Software Engineering at Altinbas University. The focus of his research is software engineering in Health informatics. He has over twelve years of experience in managing software development projects ranging from large government IT systems. Contact him by: moreb.mohammad@gmail.com

Oguz Ata, Head of Department at Software Engineering at Altinbas University, received the BSc degree Computer Engineering from Sakarya University in 2004, and the MSc degrees Computer Engineering from Beykent University in 2008, and PhD degrees in software engineering from Trakya Üniversitesi in 2012. He is a lecturer at Altinbas University. His research interests include software repository mining, software measurement and testing, process improvement, and requirements engineering. Contact him by: oguzata@gmail.com

References

1. Holzinger A, Plass M, Holzinger K, Crişan GC, Pintea CM, Palade V. Towards interactive machine learning (iML): Applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Cham; 2016. p. 81–95. doi:10.1007/978-3-319-45507-5_6.
2. Cakici B, Hebing K, Grünewald M, Saretok P, Hulth A. CASE: A framework for computer supported outbreak detection. BMC Med Inform Decis Mak. 2010.
3. Vickers AJ, Salz T, Basch E, Cooperberg MR, Carroll PR, Tighe F, et al. Electronic patient self-assessment and management (SAM): A novel framework for cancer survivorship. BMC Med Inform Decis Mak. 2010.
4. Hassan MK, Desouky AI El, Elghamrawy SM, Sarhan AM. Big Data Challenges and Opportunities in Healthcare Informatics and Smart Hospitals. Springer International Publishing; 2019. doi:10.1007/978-3-030-01560-2.
5. Bobb JF, Claus Henn B, Valeri L, Coull BA. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. Environ Heal. 2018;17:67. doi:10.1186/s12940-018-0413-y.
6. Aribisala B, Olabanjo O. Medical Image Processor and Repository—MIPAR. Informatics Med Unlocked. 2018;12 July:75–80. doi:10.1016/j.imu.2018.06.005.

7. Aigner W, Miksch S. CareVis: Integrated visualization of computerized protocols and temporal patient data. *Artif Intell Med*. 2006.
8. Krause J, Perer A, Stavropoulos H. Supporting Iterative Cohort Construction with Visual Temporal Queries. *IEEE Trans Vis Comput Graph*. 2016.
9. Pathinarupothi RK, Durga P, Rangan ES. Data to diagnosis in global health: A 3P approach. *BMC Med Inform Decis Mak*. 2018.
10. Witten IH, Frank E, Hall MA, Pal CJ. *Data Mining: Practical Machine Learning Tools and Techniques*. 2016.
11. To QC, Soto J, Markl V. A survey of state management in big data processing systems. *VLDB J*. 2018.
12. Khomh F, Adams B, Cheng J, Fokaefs M, Antoniol G. *Software Engineering for Machine-Learning Applications: The Road Ahead*. *IEEE Software*. 2018.
13. Diao JA, Kohane IS, Manrai AK. Biomedical informatics and machine learning for clinical genomics. *Hum Mol Genet*. 2018;27:R29–34.
14. Cheng PH, Chen YP, Lai JS. An interflow system requirement analysis in health informatics field. 2009 WRI World Congr Comput Sci Inf Eng CSIE 2009. 2009;7:712–6.
15. George C, Duquenoy P, Whitehouse D. *eHealth: Legal, ethical and governance challenges*. 2014.
16. Oliver C. Critical realist grounded theory: A new approach for social work research. *Br J Soc Work*. 2012.
17. Clark KD, Woodson TT, Holden RJ, Gunn R, Cohen DJ. Translating Research into Agile Development (TRIAD): Development of Electronic Health Record Tools for Primary Care Settings. *Methods Inf Med*. 2019;58:001–8. doi:10.1055/s-0039-1692464.
18. DiSantostefano J. *International Classification of Diseases 10th Revision (ICD–10)*. *J Nurse Pract*. 2009.
19. Weilkiens T, Lamm JG, Roth S, Walker M. B: The V-Model. In: *Model-Based System Architecture*. 2015.
20. Beck K, Beedle M, Van Bennekum A, Cockburn A, Cunningham W, Fowler M, et al. *Agile Manifesto*. *Software Development*. 2001.
21. Khomh F, Adams B, Cheng J, Fokaefs M, Antoniol G. *Software Engineering for Machine-Learning Applications: The Road Ahead*. *IEEE Softw*. 2018;35:81–4. doi:10.1109/MS.2018.3571224.
22. Salvador-Meneses J, Ruiz-Chavez Z, Garcia-Rodriguez J. Compressed kNN: K-nearest neighbors with data compression. *Entropy*. 2019.
23. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One*. 2018.
24. Clifton DA, Gibbons J, Davies J, Tarassenko L. *Machine Learning and Software Engineering in Health Informatics*. 2012;:37–41. <http://www.cs.ox.ac.uk/people/jeremy.gibbons/publications/ml-se-hi.pdf>.

25. Boonchieng E, Duangchaemkarn K. Digital disease detection: Application of machine learning in community health informatics. 2016 13th Int Jt Conf Comput Sci Softw Eng JCSSE 2016. 2016.
26. Frochte J, Frochte J. Python, NumPy, SciPy und Matplotlib—in a nutshell. In: Maschinelles Lernen. 2019.
27. Azeem MI, Palomba F, Shi L, Wang Q. Machine learning techniques for code smell detection: A systematic literature review and meta-analysis. Information and Software Technology. 2019.

Figures

Figure 1

SEMLHI Frameworks as a conceptual framework.

Figure 2

SEMLHI Framework Components.

Linear Regression

- Find x,y using linear line up down, Using equation to describes a line that best fits the relationship between the input variables (x) and the output variables (y), by finding specific weightings for the input variables called coefficients (B).

K-Nearest Neighbors:

- its mainly work by determine the similarity between the data instances, require a lot of memory or space to store all of the data, if prediction needed training dataset to do a calculation, KNN is a bad choice, Random Forest (RF)

Logistic Regression:

- Find x,y using non-linear line, use logistic function to prediction the output to transformed using a non-linear function.

Linear Discriminant Analysis:

- classification algorithm for two class problem, and using to classify data and remove outlier.

Naive Bayes:

- Low cost of computation, probabilistic classifiers for supervised learning used for large range of complex problems, used on numeric or nominal data.

Figure 3

Machine Learning algorithms that used for health classification.

Figure 4

Logistic Regression have high accuracy compared with other algorithms.

Figure 5

The Mechanism of Machine Algorithm Model

Figure 6

Machine Algorithm Model Components Algorithms.

Algorithm 1: k- nearest neighbor algorithm for Multilevel Learning by using correlation, Diagnose code, and label weight with frequency

input: Heterogeneous data source, number of k, correlation n, training set ;
label of x

while While condition **do**

- 1 - all heterogeneous data set we will work on correlation multi label add to one label for i=1 to t, join table
- 2- ML to one label based on Freq, DG wight, Lowe's accuracy
- 3- create new role based on step 2
- 4- new role all to new schema create new role if micro=sen and test= normal then mml=3
- 5- classify ML based on DC category
6. predicate new disease based on role created;

end

Figure 7

Pseudo code k-nearest neighbor algorithm for Multilevel Learning.

Figure 8

Age and category Feature summarize with other's variables (test result, gender, icd 10).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Methodsformula.docx](#)