International Journal of Disaster Recovery and Business Continuity
Vol.11, No. 1, (2020), pp. 2685–2694

An Effort Estimation Model for Agile Software Development Using Machine Learning In Healthcare

Shanu Verma[1], Dr. Rashmi Popli[2], Dr. Harish Kumar[3]
[1]*Research Scholar, J.C.Bose University of Science and Technology*
[2]*Assistant Professor, J.C.Bose University of Science and Technology*
[3]*Assistant Professor, J.C.Bose University of Science and Technology*
[1]*shanu.verma56@gmail.com,*[2]*rashmimukhija@gmail.com,* [3]*htanwar@gmail.com*

Abstract

Most of the software projects are not successful because of their wrong cost estimation. To palliate the wrong estimation, the accurate and effective estimation in all the software development process is required. Estimation has significant role in software development as it estimates its cost and the effort required thus resulting in overall success of software development. So far many effort estimation models have been developed for software projects, most among them produced accurate result, but did not provide flexibility during decision making of software development process. Choose the right method and right practices and applying them adequately in effort estimation of the software predicts the success of software development. The main objective of this paper is to estimate effort in healthcare while using the scrum in agile machine learning. The rising cost of Health care is one of the concerns for poor people around the world. Agile system in health care can provide an efficient framework for streamlining and improving governance. To Estimate the cost of healthcare the model is developed using machine learning algorithms. Story point as a unit of measure is used to assess the efforts involved in an issue. The generated model was tested using Precision, AUC, Recall, and Accuracy. The results of the estimate were then compared to story point estimation. The purpose of this paper is to estimate the size and effort in agile development from story points, which are user stories, are calculated using user learning techniques to improve the efficiency and accuracy of healthcare in various ways.

Keywords- *Software Effort Estimation, Story Point, Agile Methodology, Scrum, Machine learning algorithm, Naïve Bayes, Logistic Regression, Random Forests, Decision tree, thyroid, chronic disease*

1. Introduction

In recent years, healthcare analysis is in great demand in the most promising research areas. In this globalization era, the healthcare organization is having difficulty in responding [22]. For the dynamic, uncertain demand and constantly changing environment of the customer and this difficulty should be reduced because the need for health services. According to the Medicines report, 80 percent of health care projects fail without estimation, a third is never established and most projects are usually out of the budget. [5] Handling this raw data manually is very difficult. For the analysis of data, machine learning has emerged as an important tool and calculates effort estimation. [12] Poor management is the main reason for failure of software projects. Inaccurate estimation, constant change of requirements, incomplete requirements, Lack of communication between developers, unable to identify risk early stages of development, adoption of relevant process models, etc. Accurate estimation is critical to software

ISSN: 2005-4289 IJDRBC

success of development projects. Agile methodology allows the software products to be delivered in a very short time. [17] So that software development efforts like program, budget and others can be estimated at an early stage software development life cycle and this can be possible from agile methodology.

2. Literature Survey

- Velayutham et al (2020) paper describes thyroid dysfunction on south Indian population among young females. It study on seven colleges and conduct screening test based on TSH value to diagnose thyroid in female students. The paper concluded that one out of eight female students has abnormal TSH.

- Gultekin et al (2020) objective of this paper is calculating effort on each issue, and then the total effort is calculated using Scrum methodology. It demonstrate that error rate of story point is better than other estimation model. It works on five different datasets and compares the error rate among these datasets. The conclusion of this paper is other story point model can be used to provide high result.

- Singh KC et al (2020) paper discusses past research, present research in healthcare and describes future opportunities. It discuss medical areas where tools and operation to improve healthcare delivery. It describes the area of operation of health services is still New-born, with the immense challenges of aging population and rising health costs.

- Zhang et al (2020) survey paper describes testing of machine learning with fairness, correctness and robustness. This survey paper presents the testing properties of machine learning and summarizes with datasets and open source testing tools or framework work. This paper helps software engineers and machine learning researchers.

- Jackson, Yaqub and Li et al (2019) proposed that the continuous delivery of applied machine learning models in healthcare is frequently vulnerable to the deployments of isolated product with poorly developed architectures. For instance in healthcare trained the actuarial model which are used to implement the models in real world and distinguish from client facing software. Some systems prove difficult to maintain, to standardize on population, and to include latest design features and capabilities. Using an agile methodology System is ready for estimation the cost of healthcare and describes the role of product team in an existing research. In healthcare models of machine learning are universally implemented for the challenges and give accurate recommendations on actively address these issues [21].

- The goal of Sindhwani et al (2019) exploratory paper is to focus on the present status of implementation of the agile system in health care. The focus on this paper is to explore the definition of agile in healthcare, discuss barriers, implementation and characteristic. The system of healthcare was implemented to review the research of literature and to develop a technique to implement it. In the agile system which is flexible in nature and enables quickly respond according to customer demand and market variations on optimizing costs and quality by accepting upgraded processes, trainings and tools. In healthcare the agile system is used to create an efficient framework to establish and improvement of process. It can only be accomplished if a method of system is compulsory with work in proper way. The steps of common implementation steps propose agile

ISSN: 2005-4289 IJDRBC

training, start pilot projects and improvement using heterogeneous teams. The barrier in knowledge is lack of advisors and instructors and they can supply by sharing their knowledge and experience with the help of an example of agile in healthcare from the example of real life applications.

3. Goal and Research Questions

This research questions were raised in healthcare to estimate the software using agile effort estimation using machine learning.

I. Why do Healthcare Projects generally fail?
II. What are the input features that can be used for cost estimation can be of software project?
III. For Agile estimation which machine learning technique can be used
IV. What are the overall estimation accuracy of ML algorithms?
V. What are the supervised learning methods that can be used for cost estimation?
VI. What are the performance measures and evaluation results for cost estimation?

4. Proposed Approach

The proposed approach is based on chronic disease data set and thyroid disease data set. The data set are taken from UCI machine learning repository and Kaggle. These data set are used to evaluate the percentage of disease on yearly basis. In this paper, Story Points are taken as inputs and effort calculation will be done using scrum and result will be compared using machine learning technique. [8]

4.1. Story Point

The story point method is used to estimate the effort estimate. This method is the most well-known effort estimation technique that can be applies to all agile methods. [15] Typically, story points follow Fibonacci as a sequence. [13]

0,.5,1,2,3,5,8,13,20,40,100,.....

Story point estimation ranges from 1 to 5 where 5 represent extremely large story points and 1 represents extremely small story points. In user stories, users break the big story into smaller parts. [14] When story points are predicted for each user story for all teams there will be some different weight and opinions for their user stories.[13] A user story, which is brief, provides a means to communicate very effectively between the client and the user and to convey it to the designated client.[14]. The general format of user stories is

As a <user or role>, I want <work action> so that <output will come>

This user story format includes who, what and why, who represents the user or role, what tasks represent the action and why the output is represented. [20]

Example of user story: As a researcher, I want to predict thyroid disease so that solutions come up

The point of the story also represents the size of the T-shirt, and the t-shirt size classification for user stories: Extra Small (XS), Small (S), Medium (M), Large (L) and Extra Large (XL).

ISSN: 2005-4289 IJDRBC

The work with extra-large representation is complex and requires more time than other stories and the extra small representation work is simple.

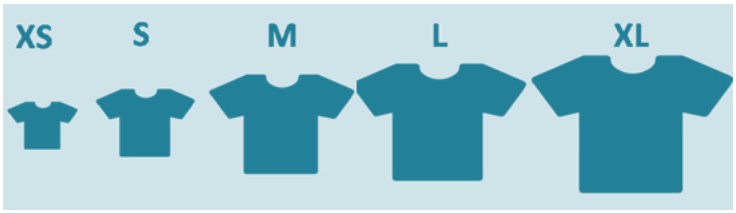

Figure 1 Estimate using T-shirt sizes

4.2. Scrum

In projects developed with Scrum, the software cost estimation differs from traditional software effort estimation model [21] [15]. Story point values are used to determine the cost of software developed with the scrum method. Some machine learning algorithms such as Linear Regression, Logistic Regression, Decision Tree, and K-Near Neighbors are used to find story point values in software developed with Scrum [3]. The following steps are used to calculate the effort estimation of a software project. [12][14]

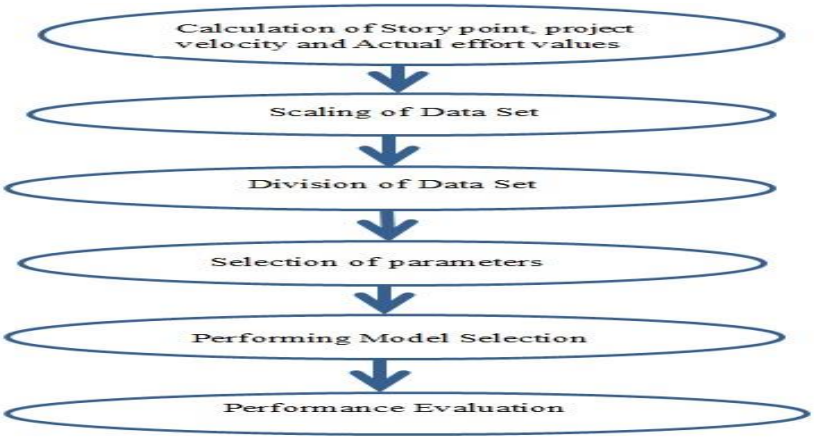

Figure 2 Calculate Effort Estimate steps

4. Implementation

4.1 Using Story point

Estimating effort in healthcare using machine learning algorithm and computation using story point. This table shows the five values, which are assigned to different types of user stories according to their size.

Table 1 Calculate story point

ISSN: 2005-4289 IJDRBC

Scale	Scaling Point Description
5	When story size has extreme large
4	When story size has large
3	When story size has moderately large
2	When story size has rough idea
1	When story size has very small

Complexity introduces uncertainty from estimation - greater complexity means uncertainty. Like the story size table, these rules are not fixed. These can only be adjusted by the team, however we have classified them to accommodate all the features of the agile software development method. [4]

Table 2 Complex Story point

Scale	Complexity Scaling Point Description
5	When story size has extreme large
4	When story size has large
3	When story size has moderately large
2	When story size has rough idea
1	When story size has tiny

The effort of a particular user story is determined using these two vectors, using the following simple formula:
Estimation Story=Complexity * size

4.2 Using Machine Learning Algorithm

Machine learning provides powerful tools to solve and find problems in data. Machine learning can be

ISSN: 2005-4289 IJDRBC

used to early the diagnose. Chronic diseases in patents using clinical trial data [19].Healthcare data are randomly divided into training data and test data to implement in machine learning. To train data using machine learning algorithms. And predict its software cost estimation we build three algorithms such as decision tree, Naive Bayes, K-NN to learn the model. The test data is used to test the model.

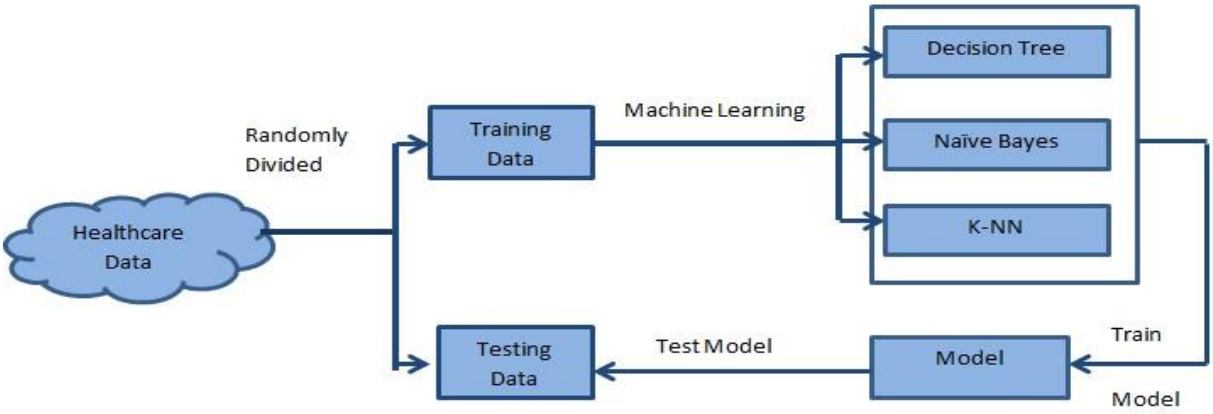

Figure 3 Train and test data by machine learning algorithms

Case Study

Study 1-Identify Chronic Diseases

This research paper has analyzed chronic diseases like diabetes, cancer, thyroid, lung disease, and cholesterol, high BP. Men are healthier than women according to India Fit Report. [11]

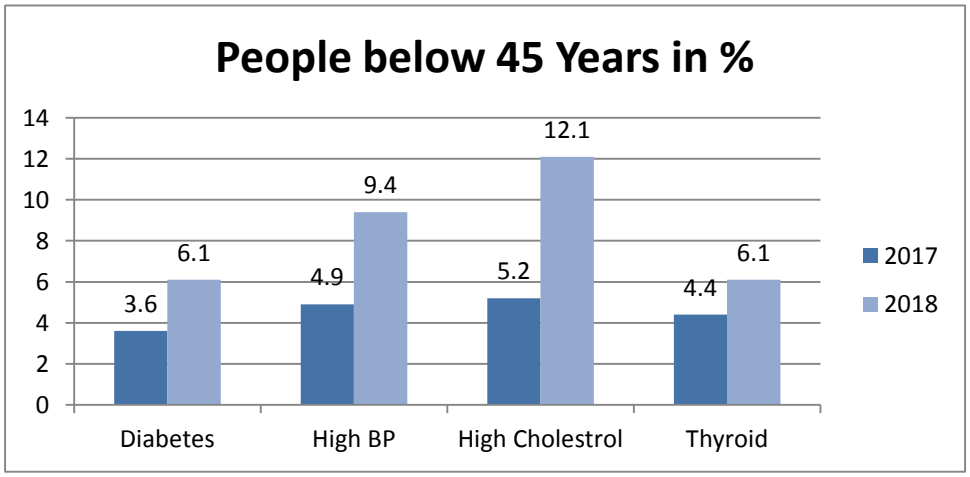

Figure 4 Chronic disease

According to the report, there has been an increase in chronic diseases among persons under 45 years of age. Let us take a look at chronic diseases affecting thyroid, cancer, diabetes, high blood pressure,

ISSN: 2005-4289 IJDRBC

cholesterol in India. Both diabetes and thyroid patient are increased in India. [11] Diabetes related cases have increased from 7.1% last year to 12% this year. During this period, thyroid cases have increased from 6.8 per cent to 10.7 per cent.

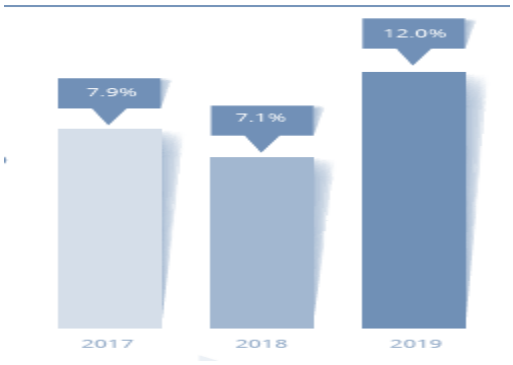

Figure 5 Raise of diabetes in past three years

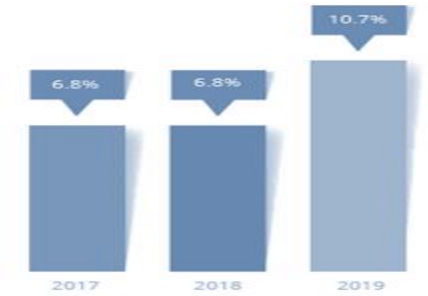

Figure 6 Raise of thyroid in past three years

Study 2- Thyroid Disease

Thyroid disease is detected based on TSH value. If the value of TSH is low, there is an indication of hypothyroidism and hyperthyroidism. [10] If the TSH value is normal, there is an indication of hypothyroxinemia. [18] If the value of TSH is high then there is an indication of hypothyroidism disease. The implementation of thyroid disease in the Python programming language uses a sample data set.

```
import pandas as pd
```

```
thd=pd.read_csv(r"C:\Users\Arshiya\Downloads\Thyroid_disease.csv")
```

```
thd.columns
```

```
Index(['age', 'sex', 'on thyroxine', 'query on thyroxine',
       'on antithyroid medication', 'sick', 'pregnant', 'thyroid surgery',
       'I131 treatment', 'query hypothyroid', 'query hyperthyroid', 'lithium',
       'goitre', 'tumor', 'hypopituitary', 'psych', 'TSH measured', 'TSH',
       'T3 measured', 'T3', 'TT4 measured', 'TT4', 'T4U measured', 'T4U',
       'FTI measured', 'FTI', 'TBG measured', 'TBG', 'referral source',
       'binaryClass'],
      dtype='object')
```

```
thd.head()
```

	age	sex	on thyroxine	query on thyroxine	on antithyroid medication	sick	pregnant	thyroid surgery	I131 treatment	query hypothyroid	...	TT4 measured	TT4	T4U measured	T4U	FTI measured	FTI	TBG measure
0	41	F	f	f	f	f	f	f	f	f	...	t	125	t	1.14	t	109	
1	23	F	f	f	f	f	f	f	f	f	...	t	102	f	?	f	?	
2	46	M	f	f	f	f	f	f	f	f	...	t	109	t	0.91	t	120	
3	70	F	t	f	f	f	f	f	f	f	...	t	175	f	?	f	?	

ISSN: 2005-4289 IJDRBC

Figure 7 Thyroid disease

Thyroid disease is represented in this chart by age and gender. Percentage of these disease are more in women but men members are increasing regularly.

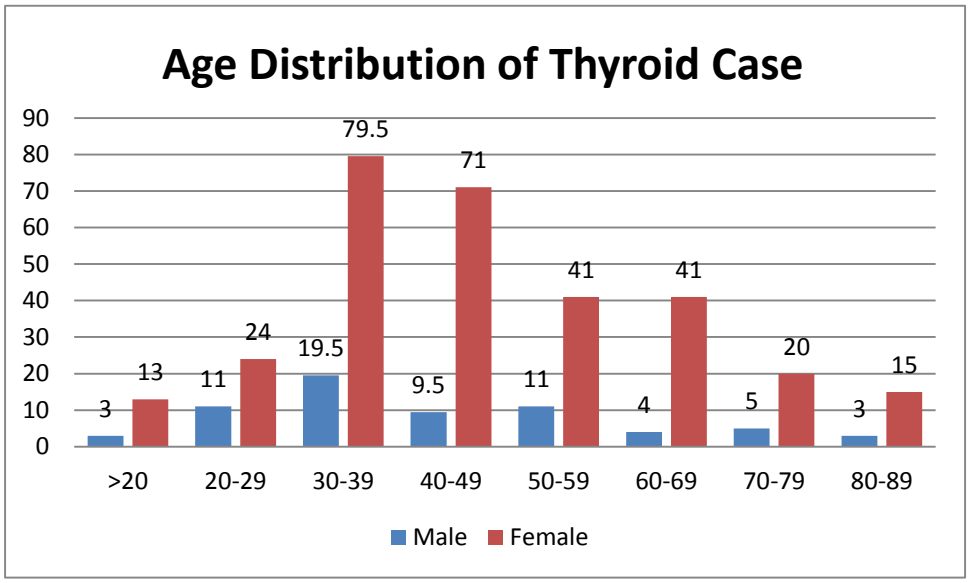

Figure 8 Thyroid case

5. Results

Prediction of type of health related disease using machine learning in Python. To develop software included in the sample dataset. Results are based on the training set and test set we used in the proposed approach. The data was then preprocessed, searched iteratively to complete it minimum requirements to meet defined objectives. Three machine learning techniques to work on the selected processed data: Random Forest, Logistic Regression, Naive Bayes. Selection was based on certainty of the thyroid sample data explored that matched machine learning techniques that can serve these purposes and consistent with the nature of the sample data. [16] Performance measurement using machine learning technology based on thyroid sample dataset Performance of each model using four selected performance measures: AUC, CA, Precision, Recall, error, F1-score and accuracy in table 3

Table 3 Final Result

Method	Accuracy	Recall	Precision	MSE	AUC	F1 Score
Logistic Regression	.92	.99	.93	.076	.84	.43

ISSN: 2005-4289 IJDRBC

Naïve Bayes	.71	.72	.95	.29	.57	.30
Decision Tree	.94	.95	.98	.061	.80	.71
Random Forest	.71	.97	.93	.090	.73	.39

Conclusion

All applied machine learning techniques were performed using evaluation Five-fold cross-validation. Machine learning techniques such as Naive Bayes, Random Forest, Decision Tree, and Logistic Regression are compared in Table 3. Decision Tree outperformed from Naïve Bayes, Random forest, logistic regression in all the other scores. It scored 94% in accuracy, 80% in AUC, 98% in Precision, 95% in Recall, and 71% in F1 Score. On the other hand, Logistic Regression outperformed the other three Recall, MSE and AUC measures. It scored 99%, 76%, 84% measures respectively. Despite the fact that Random Forests outperformed from other three Logistic Regression, Decision Tree, and Naïve Bayes in its Recall and MSE.

References

1. Kumaravel Velayutham (2020) Prevalence of thyroid dysfunction among young females in a South Indian population
2. Muaz Gultekin(2020) "Story Point-Based Effort Estimation Model with Machine Learning Techniques " World Scientific Publishing Company DOI: 10.1142/S0218194020500035
3. Jie M. Zhang (2020) Machine Learning Testing: Survey, Landscapes and Horizons http://www.ieee.org/publications_standards/publications/rights/index.html
4. Diwas Singh KC (2020) Empirical Research in Healthcare Operations: Past Research, Present Understanding, and Future Opportunities https://doi.org/10.1287/msom.2019.0826
5. Teen-Hang Meen (2020) Selected Papers From 2019 IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability 2020, 12, 414; doi:10.3390/su12010414
6. Nachiket MOR (2020) Human Resources for Primary Healthcare in India Training & Certification.
7. Diwas Singh KC(2019) Empirical Research in Healthcare Operations: Past Research, Present Understanding, and Future Opportunities https://doi.org/10.1287/msom.2019.0826
8. Rahul Sindhwani(2019) Agile System in Health Care: Literature Review https://doi.org/10.1007/978-981-13-6412-9_61 Springer Nature Singapore Pte Ltd. 2019
9. Jackson, Yaqub and Li (2019). The Agile Deployment of Machine Learning Models in Healthcare, Computer 34(2): 118-119.
10. Shalini L(2019) A Hypothyroidism Prediction using Supervised Algorithm International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958
11. Lifestyle diseases: An epidemic among young Indians Anusha (2019) https://goqii.com/blog/lifestyle-diseases-an-epidemic-among-young-indians/
12. Hosahalli Mahalingappa Premalatha (2019) Effort Estimation in Agile Software Development using Evolutionary Cost- Sensitive Deep Belief Network

ISSN: 2005-4289 IJDRBC

13. Ahmed Bani Mustafa(2018) Predicting Software Effort Estimation Using Machine Learning Techniques
14. Ch. Prasada Rao(2018)," An Agile Effort Estimation Based on Story Points Using Machine Learning Techniques" Proceedings of the Second International Conference on Computational Intelligence and Informatics, Advances in Intelligent Systems and Computing 712, https://doi.org/10.1007/978-981-10-8228-3_20
15. Emanuel Dantas(2018) Effort Estimation in Agile Software Development: An Updated Review World Scientific Publishing Company DOI: 10.1142/S0218194018400302
16. Arwinder Dhillon (2018) Machine Learning in Healthcare Data Analysis: A Survey Journal of Biology and Today's World http://journals.lexispublisher.com/jbtw doi: 10.15412/J.JBTW.01070206
17. Dr. S. Rama Sree(2017) An Early Stage Software Effort Estimation in Agile Methodology Based On User Stories Using Machine Learning Techniques www.ijraset.com
18. S.Umadevi(2017) Applying Classification Algorithms to Predict Thyroid Disease International Journal of Engineering Science and Computing,
19. Niharika G. Maity (2017) Machine Learning for Improved Diagnosis and Prognosis in Healthcare 978-1-5090-1613-6/17/$31.00 ©2017 IEEE
20. Shashank Mouli Satapathy(2014) Story Point Approach based Agile Software Effort Estimation using Various SVR Kernel Methods
21. Ziauddin (2012) An Effort Estimation Model for Agile Software Development Vol. 2, No. 1, 2012, ISSN 2166-2924 World Science Publisher, United States
22. REBECCA KITZMILLER "Adopting Best Practices ''Agility'' Moves from Software Development to Healthcare Project Management"2006 Lippincott Williams & Wilkins, Inc.

ISSN: 2005-4289 IJDRBC