Application of Recommendation System on E-Learning Platform Using Content-Based Filtering with Jaccard Similarity and Cosine Similarity Algorithms

Yohanes Leonardus Sukestiyarno *, Hasballah Askar Sapolo , Hizir Sofyan

Posted Date: 23 June 2023

doi: 10.20944/preprints202306.1672.v1

Keywords: Recommendation System, Content-Based Filtering, Jaccard Similarity, Cosine Similarity, Mean Absolute Error

Article

Application of Recommendation System on E-Learning Platform Using Content-Based Filtering with Jaccard Similarity and Cosine Similarity Algorithms

Sukestiyarno [1,*], Hasballah Askar Sapolo [1] and Hizir Sofyan [2]

[1] Department of Mathematics, Universitas Negeri Semarang, Semarang, Indonesia; askarrsapolo@students.unnes.ac.id
[2] Department of Statistics, Universitas Syiah Kuala, Aceh, Indonesia; hizir@usk.ac.id
* Correspondence: sukestiyarno@mail.unnes.ac.id

Abstract: This study aims to apply a Recommendation System with Content Based Filtering method with Jaccard Similarity and Cosine Similarity algorithms on the E-Learning Platform. Recommendation systems deal with how to provide personalized recommendations to users efficiently. The Content Based Filtering method with Jaccard Similarity and Cosine Similarity algorithms can be used to calculate the similarity value between E-Courses on the E-Learning Platform. Implementation for Recommendation System using Google Colaboratory with Python programming language. In the application of the Recommendation System dataset, Coursera Free Dataset consists of 975 instances. The recommendation results use the Jaccard Similarity algorithm with an average similarity value of 0.3 while the value of Cosine Similarity with the average similarity value is 0.6 where the similarity value of Cosine Similarity is higher. Based on the results of the Mean Absolute Error in the low recommendation system, the average MAE value for all iterations of Jaccard Similarity algorithm is 0.013 and for the Cosine Similarity algorithm the average MAE value for all iterations is 0.014. This shows that the Recommendation System with Jaccard Similarity and Cosine Similarity algorithms can be used on the E-Learning Platform to provide efficiency solutions for personalized recommendations.

Keywords: Recommendation System; Content-Based Filtering; Jaccard Similarity; Cosine Similarity; Mean Absolute Error

1. Introduction

The E-Learning Platform has increased from year to year. In 2016, 21 million users were registered on the Coursera E-Learning Platform, in 2017 there were 28 million users, and in 2018 there were 35 million users. There is an increase of 7 million every year. In 2020 with the Covid-19 pandemic, the number of people participating in online learning increased dramatically. This can be seen from the online registration data on the Coursera Platform which has experienced a very significant increase, namely in 2019 there were 76 million registrants while in 2020 there were 143 million. The number of registrations has doubled, while from 2020 to 2021 there was an increase of 30%, namely 189 million registrants. Based on data from Coursera, Indonesia is included in the 10 countries with the most increase in learners on the Coursera Platform. Indonesia is in fifth place with an increase of 69% with 789,000 learners in 2021 (WEF, 2022).

After the WHO officially announced that Covid-19 became a symptom of an international pandemic. The Covid-19 pandemic has had a huge impact on all fields, especially in the field of education. Face-to-face learning cannot be carried out during Covid-19. This has made many countries change the traditional way of learning to online learning to stop transmission from the Covid-19 pandemic. This online learning process is adapted quickly in various countries so that the teaching and learning process can still be carried out even though it is not face-to-face.

This data shows that online learning platforms are starting to be trusted and can be an option to add expertise in various fields. In Indonesia itself, online learning platforms such as Coursera have also increased, and many startups have the same concept, namely online learning. Some of them are

Zenius, a company engaged in online tutoring in video format for elementary, junior high, high school, and college entrance preparation students. Zenius.net based learning can motivate and increase the concept of understanding from students by 60% on average ideal scores [1]. Ruangguru is also a company engaged in online learning for elementary, junior high, high school, and college entrance preparation students. The system provided by Ruang Guru creates new learning behaviors in Indonesia and future technology-based learning experiences through Ruang Guru products, students and teachers can learn critical thinking, creativity, collaboration, and communication [2]. In addition, there are also online learning platforms such as Dicoding which is used for education in the field of technology, Foodizz which is used for culinary business education, and many others.

The huge potential in the field of online learning has increased significantly, especially during the Covid-19 pandemic, and this changed the way people learn significantly. However, providing personalized content for learners is a problem in online learning. Helping determine the right learning according to the needs and interests of students is an important problem to be solved by platform providers [3]. Recommendation System can help distance learning learners get personalized material that can increase efficiency in learning, satisfy the needs of learners, and offer learner-centered services [4].

In the era of the industrial revolution, 4.0 Machine Learning has become popular in various fields because of its ability to learn past and make smart decisions [5]. Recommendation System which is part of machine learning is one solution to provide personalized material in online learning. The Recommendation System has been widely developed in many fields such as music, film, news, and products in general. Recommendation System has been widely applied by large companies to meet the needs of customers. Linkedin, Amazon, and Netflix are just a few examples. Linkedin recommends relevant connections with people the user might know and fit their profile. Amazon suggests products that are relevant to what customers like. Netflix provides recommendations for movies or series that match what customers watch and like [6].

Recommendation systems have a big impact on business, large companies like Netflix that 75% of what customers watch comes from the Recommendation System, Youtube also reports that 60% of clicks on the homepage come from the Recommendation System, the CEO of Amazon in 2006 also stated that 35% of their sales came from Cross Selling, namely the Recommendation System. Researchers estimate the business value of recommendations and personalization at more than $1 billion per year. The amount of potential that can be obtained from the Recommendation System to create recommendations on personalized online learning can also have a big business impact on the company. This makes researchers choose a Recommendation System using Content-Based Filtering with Jaccard Similarity and Cosine Similarity algorithms as research [7].

Content-Based Filtering aims to group products with similar attributes, based on references from users. Then the Recommendation System with Content-Based Filtering will suggest different items with the same attributes [8]. Jaccard Similarity and Cosine Similarity algorithms were used in this study because these algorithms are widely used in building Recommendation Systems. Cosine Similarity algorithm uses vector similarity measured by the cosine angle of both vectors. The closer the cosine value is to 1, the smaller the angle between the two vectors in space. But the accuracy of the Recommendation System in the study was very low, the researchers said it was because the use of data sets was too small, which only used titles and descriptions from Netflix. And researchers suggest more complete attributes in the data set such as show duration, Netflix ratings, prominent cast, and others so that the Recommendation System can be more accurate [9].

The Jaccard Similarity algorithm is also widely used in building Recommendation Systems. Jaccard Similarity is a popular method for calculating similarities between users/items. In the calculation, only considers the number of general ratings between the two users. The benefits of using this method are maximum when the number of general ratings is greater [10]. In another study, used student data sets collected from undergraduate students with computer science backgrounds within 4 months. The student data set consists of 480 student data with 468 descriptions of learning objects and 8200 student ratings. To create a Recommendation System for Learning Management System (LMS). Of the 4 algorithms used, namely Pearson Correlation Similarity (PCC), Cosine Vector

 doi:10.20944/preprints202306.1672.v1

Similarity (CVS), Jaccard Similarity Correlation (JSC), and Euclidean Distance Similarity (EDS) with evaluation metrics, namely Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), the best algorithm in the study was Cosine Vector Similarity (CVS), the second Jaccard Similarity Correlation (JSC), the third Euclidean Distance Similarity (EDS), and finally Pearson Correlation Similarity (PCC) [11]. From previous studies, researchers chose Jaccard Similarity and Cosine Similarity algorithms to build a Recommendation System.

Data sets are also an important part of building a Recommendation System. The amount of data has increased dramatically with the number of internet users. But not all data available on the internet is useful or provides satisfactory results for users. Researchers created a Recommendation System to provide relevant information to users taking into account past data preferences. Data is personally filtered and customized as per user needs. With a large amount of data, the Recommendation System also becomes more accurate [6]. In this study, researchers used a qualitative data set obtained from Kaggle, namely the Dataset of the free courses available on Coursera which consists of 8 variables, namely URL, price, institution, title, skill you will gain, ratings, reviews, level type duration. The data consists of 975 instances, the variables used to create the Recommendation System are taken from two variables, namely title and skill you will gain. The skill variable you will gain will be used as an attribute in this study. These attributes will be vectorized so that the skill variable you will gain becomes a binary format using Python. Tools used to process data in the form of Python programming language through Google Colaboratory.

The Jaccard Similarity and Cosine Similarity algorithms will be used by importing the Scikit-Learn library which is a machine learning library for the Python programming language. The test results can be seen based on distance measurement values obtained from the SciPy library which is also a library of the Python programming language for SciPy providing algorithms for optimization, integration, interpolation, eigenvalue problems, algebraic equations, differential equations, statistics, and many other classes of problems.

Research using Content-Based Filtering is most widely used for recommended films. Movie Lens Dataset which contained 9126 films classified by genre. There are a total of 11 genres. And movie ratings have been collected from 671 users. The algorithm used in this study is Euclidean distances to calculate distances against other users obtained and which have recommended minimum values. The results of the study show the output of various movies that have been recommended to users based on their previous behavior patterns [6]. This other study aims to offer general recommendations for each user, based on the popularity and/or genre of the film. The algorithm used is Cosine Similarity, Cosine Similarity is beneficial because it helps in finding the similarity of objects. After the Cosine Similarity result matrix was found later using the KNN function, researchers found the nearest neighbor of the film to be recommended to the user. Researchers state that the KNN algorithm is implemented in the model along with Cosine Similarity because it provides more accuracy than other distance metrics and its complexity is relatively low as well [12].

2. Literature Review

2.1. Content-Based Filtering

Content-based filtering uses attributes to make recommendations. In the Content-Based Filtering rating, the buying behavior of the user is combined with the content information available in the item. For example, suppose a customer gives a high rating on the Terminator movie, but we don't have access to ratings from other users. Then Content-Based Filtering can be used by searching for the same genre as an attribute to provide recommendations to customers. The selected genre will be used to predict whether the customer likes the item given based on the recommendation. Content-Based Filtering has several advantages in making new recommendations for an item when sufficient data is not available for that item. Because by using the right attributes, Content-Based Filtering is still used. In the case of films, it can use genres and recommend the same genre [13].

Content-Based Filtering is defined as the use of metadata held on items in the available catalog. In the case of Netflix, Netflix uses descriptions of movies, for example, based on specific algorithms

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1672.v1

the system can calculate recommendations based on liked items and find similar content. By comparing items and user profiles, or, if there are no users involved, Content-Based Filtering can still provide recommendations based on similar items [14].

2.2. Vectorizing Attributes

Vectorizing attributes are used to speed up Python code without using loops. Using such functions helps to efficiently minimize Python code time. By converting data that is still in the form of lists in Python into matrices, it allows data to be processed in a recommendation system. Vectorizing Attributes is important because the core of data processing, for example, data in the form of Comma Separated Value (CSV), still needs transformation into a matrix so that it can be processed in Machine Learning [15].

2.3. Jaccard Similarity

Jaccard Similarity is used to calculate the distance between two sets. To obtain Jaccard Similarity, it is to calculate the slice between two sets and then divide by the combined set.

$$sim(A, B) = \frac{A \cap B}{A \cup B} \tag{1}$$

Jaccard Similarity can then be defined according to the following formula.

$$d(A, B) = 1 - sim(A, B) = 1 - \frac{A \cap B}{A \cup B} \tag{2}$$

Getting closer to 1 value from Jaccard Similarity indicates similarity between two sets [16].

2.4. Cosine Similarity

Cosine Similarity is used to determine which users or items are nearby by providing recommendations. Cosine similarity is a measurement between two nonzero vectors of multiplication in space by measuring the cosine of the angle between them. Cosine of 0^0 is 1 and less than 1 for other angles.

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3}$$

Here, A_1 and B_1 is a component of the vector A_1 dan B_1 respectively. The value of Cosine Similarity indicates very high similarity between the two vectors, and the highest possible value is 1. When calculating the similarity between users or items, Cosine Similarity is applied to rank vectors and sort them based on the obtained Cosine Similarity value. Then we can see users or items that have high similarity by comparing between vectors [17].

2.5. Mean Absolute Error

Mean Absolute Error (MAE) is used to calculate the average difference between the calculated value and the actual value. MAE calculates the error between the actual value and the value predicted by the model. MAE is commonly used to calculate the accuracy of machine learning models. MAE can be defined as follows.

$$MAE = \frac{\sum_{\{u,i\}} |p_{ui} - r_{ui}|}{N} \tag{4}$$

where p_{ui} is the predicted value of the recommendation system, r_{ui} is the actual value of you and N is the sum of data [18].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1672.v1

3. Recommendation Process

In the process of making a Recommendation System using the Content-Based Filtering method with Jaccard similarity and cosine similarity algorithms then calculate the absolute Mean Absolute Error (See Figure 1). In addition, the libraries used in carrying out the analysis process are numpy, pandas, scipy and sklearn.

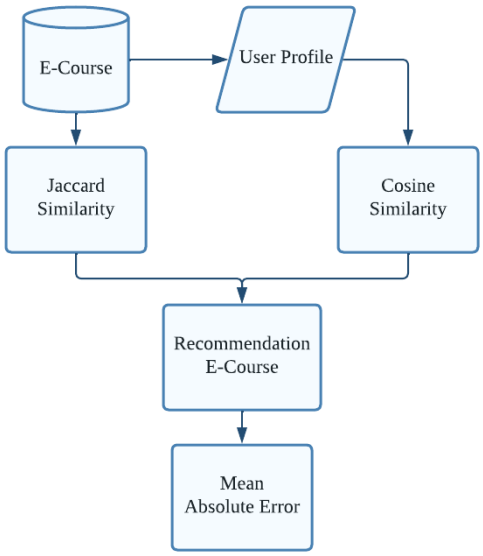

Figure 1. Recommendation Process.

3.1. Datasets

At this stage, it is carried out to obtain data related to or needed in this study. The source of data in this study is secondary data from Kaggle, namely Coursera Free Dataset.

3.2. Data Preparation

Data Preparation is the process of organizing data so that data can be processed to the stage of making a recommendation system. At this stage, Vectorizing Attributes is carried out by converting data that is still in the form of a list in Python into a matrix.

Table 1. Vectorizing Attributes.

Title	Skill 1	Skill 2	Skill 3	…	Skill n
E-Course 1	1	0	0	…	1
E-Course 2	0	1	0	…	0
…	…	…	…	…	…
E-Course n	0	0	0	…	1

3.3. Content-Based Filtering Using Jaccard Simmilarity Algorithm

The prepared data is tested with the Jaccard Similarity algorithm to manually calculate the similarity distance between items in Python as a test. Then calculate the similarity distance between items and all items with Jaccard Similarity and then sort items with the highest similarity as a recommendation.

3.4. Content Based Filtering Using Cosine Similarity Algorithm

The prepared data is tested with the Cosine Similarity algorithm to manually calculate the similarity distance between items in Python as a test. Make a user profile then calculate the similarity

 doi:10.20944/preprints202306.1672.v1

distance between items and all items with Cosine Similarity and then sort items with the highest similarity as a recommendation.

3.5. Accuracy Level Measurement Using Mean Absolute Error (MEA)

Measurement of the level of accuracy using both the Jaccard Similarity algorithm and the Cosine Similarity algorithm can be done by calculating the class between the actual value and the value predicted by the model using the Mean Absolute Error (MEA).

4. Results

4.1. Recommendation System with Jaccard Similarity

Manually from the calculation of Jaccard Similarity, the value of Jaccard Similarity between Financial Markets and Information Systems Auditing is 0.2083. For Financial Markets and The Global Financial Crisis Controls and Assurance, it is 0.2608. Information Systems Auditing and The Global Financial Crisis amounted to 0.3846. From these results, users who take the Information Systems Auditing course will be given recommendations for The Global Financial Crisis compared to Financial Markets due to the greater value of Jaccard Similarity.

The overall dataset of the recommendation system for the Information Systems Auditing, Controls, and Assurance e-course recommends the Global Financial Crisis e-course with a value of 0.384615, Risk in Modern Society with a value of 0.333333, Evaluación de peligros y riesgos por fenómenos naturales with a value of 0.3, Chemicals and Health with a value of 0.285714, Financing and Investing in Infrastructure with a value of 0.25.

4.2. Recommendation System with Cosine Similarity

Manually from the results of Cosine Similarity calculations, the value of Cosine Similarity between Indigenous Canada and In the studio: Postwar Abstract Painting is 0.7126. For Financial Markets and In the studio: Postwar Abstract Painting 0.1490. Financial Markets and the Indigenous Canada Crisis amounted to 0.1792. From these results, users who take the Indigenous Canada course will be given recommendations In the studio: Postwar Abstract Painting compared to Financial Markets because of the greater value of Cosine Similarity.

The entire recommendation system dataset for the Financial Markets e-course, Indigenous Canada, In the Studio: Postwar Abstract Painting recommends the Sexing the Canvas: Art and Gender e-course with a value of 0.688247, Formación docente basada en la práctica para desarrollar habilidades del siglo XXI with a value of 0.613139, American Deaf Culture with a value of 0.608330, Heritage under Threat with a value of 0.606977, Music and Social Action with a value of 0.587957.

4.3. Accuracy of Recommendation System Using Mean Absolute Error (MEA)

Mean Absolute Error (MAE) is used to calculate the average difference between the calculated value and the actual value. MAE calculates the error between the actual value and the value predicted by the model. Testing the model against the Jaccard Similarity algorithm is done by calculating the MEA, sklearn.metrics library and then the mean_absolute_error() function is used to obtain the MEA value. The results of recommendations from the Jaccard Similarity and Cosine Similarity algorithms are carried out to formulate mean_absolute_error (ecourse_actual. ecourse_prediction) so that they can obtain MEA values.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2023 doi:10.20944/preprints202306.1672.v1

Table 2. Mean Absolute Error.

Iteration	Jaccard Similarity	Cosine Similarity
1	0.02733812949640280	0.01525179856115100
2	0.01237410071942440	0.02071942446043160
3	0.00690647482014388	0.01534772182254190
4	0.00402877697841726	0.01016786570743400
5	0.01438848920863300	0.00738609112709832

5. Discussion

5.1. Recommendation System Using Content-Based Filtering with Jaccard Similarity and Cosine Similarity

Application of Content-Based Filtering method with Jaccard Similarity and Cosine Similarity algorithms on Coursera Free Courses Dataset with skills you will gain as attributes. Exploratory Data Analysis is carried out on the data set for initial investigation, Data Preparation is carried out for organizing data by doing Vectorizing Attributes, namely converting list-shaped data into a matrix with the title as rows and skills you will gain as columns. Then the dataset is applied Jaccard Similarity and Cosine Similarity so that it can provide e-course recommendations based on the highest similarity value. The application of the recommendation system with this method shows excellent results, as evidenced by the MAE results generated in each iteration showing a low error rate.

5.2. Efficiency Recommendation System Using Content-Based Filtering with Jaccard Similarity and Cosine Similarity

The application of the Content-Based Filtering method with the Jaccard Similarity and Cosine Similarity algorithms can be used as an efficient solution in providing e-course recommendations. In the application of Content-Based Filtering with the Jaccard Similarity and Cosine Similarity algorithms, manually the similarity value was obtained based on the selected e-course. For the Coursera Free Courses Dataset consisting of 975 instances, manual calculations take a long time for large data sets because they have to select e-courses one by one. Application of the Content-Based Filtering method with Jaccard Similarity and Cosine Similarity algorithms to the entire data set by creating a similarity table first, the recommendation system can provide e-course recommendations within a few seconds. The Application of this method shows that it can provide e-course recommendation efficiently

5.3. Accuracy of Recommendation System Using Mean Absolute Error (MEA)

Accuracy of Recommendation System using Mean Absolute Error (MAE) Based on the MAE value resulting from the iteration process obtained in the recommendation system with the Content-Based Filtering method with the Jaccard Similarity and Cosine Similarity algorithms, it is quite low. This shows that the error rate in the recommendation system is low, in the Jaccard Similarity algorithm the average MEA value for all iterations is 0.013 and for the Cosine Similarity algorithm, the average MEA value for all iterations is 0.01377.

6. Conclusion

This study concludes that the application of the recommendation system with this method shows very good results, as evidenced by the results of MAE produced in each iteration showing a low error rate, which can provide efficient solutions. The suggestions from this study can be used on the same data characteristics, with Content-Based Filtering methods with Jaccard Similarity and Cosine Similarity algorithms to create a recommendation system. This research can be developed by adding historical user data or can apply the Collaborative Filtering method, a Hybrid Filtering method to create a better recommendation system.

References

1. B. R. A. Safitri and L. Herayanti, "Pengaruh Video Pembelajaran Berbasis Zenius. Net Dalam Meningkatkan Motivasi Dan Pemahaman Konsep Siswa," JISIP (Jurnal Ilmu Sosial dan Pendidikan), 2020.
2. M. Chinmi and R. F. Marta, "RuangGuru as an Ideation of Interaction and Education Revolution during COVID-19 Pandemic in Indonesia," Revista Romaneasca pentru Educatie Multidimensionala, vol. 12, no. 2Sup1, pp. 118–129, 2020, doi: 10.18662/rrem/12.2Sup1/297.
3. G. Xu, G. Jia, L. Shi, and Z. Zhang, "Personalized Course Recommendation System Fusing with Knowledge Graph and Collaborative Filtering," Comput Intell Neurosci, vol. 2021, 2021, doi: 10.1155/2021/9590502.
4. J. Xiao, M. Wang, B. Jiang, and J. Li, "A personalized recommendation system with combinational algorithm for online learning," J Ambient Intell Humaniz Comput, vol. 9, no. 3, pp. 667–677, Jun. 2018, doi: 10.1007/s12652-017-0466-8.
5. I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Computer Science, vol. 2, no. 3. Springer, May 01, 2021. doi: 10.1007/s42979-021-00592-x.
6. S. Reddy, S. Nalluri, S. Kunisetti, S. Ashok, and B. Venkatesh, "Content-based movie recommendation system using genre correlation," in Smart Innovation, Systems and Technologies, Springer Science and Business Media Deutschland GmbH, 2019, pp. 391–397. doi: 10.1007/978-981-13-1927-3_42.
7. D. Jannach and M. Jugovac, "Measuring the business value of recommender systems," ACM Transactions on Management Information Systems, vol. 10, no. 4. Association for Computing Machinery, Dec. 01, 2019. doi: 10.1145/3370082.
8. Y. Afoudi, M. Lazaar, and M. Al Achhab, "Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network," Simul Model Pract Theory, vol. 113, Dec. 2021, doi: 10.1016/j.simpat.2021.102375.
9. M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, "Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms," Scitepress, May 2022, pp. 15–20. doi: 10.5220/0010727500003101.
10. G. Jain, T. Mahara, and K. N. Tripathi, "A Survey of Similarity Measures for Collaborative Filtering-Based Recommender System," in Advances in Intelligent Systems and Computing, Springer, 2020, pp. 343–352. doi: 10.1007/978-981-15-0751-9_32.
11. J. Joy and V. G. Renumol, "Comparison of generic similarity measures in E-learning content recommender system in cold-start condition," in 2020 IEEE Bombay Section Signature Conference, IBSSC 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 175–179. doi: 10.1109/IBSSC51096.2020.9332162.
12. R. H. Singh, S. Maurya, T. Tripathi, T. Narula, and G. Srivastav, "Movie Recommendation System using Cosine Similarity and KNN," Int J Eng Adv Technol, 2020, doi: 10.35940/ijeat.E9666.069520.
13. C. C. Aggarwal, Recommender Systems, vol. 1. Cham: Springer International Publishing, 2016.
14. K. Falk, Practical recommender systems. Simon and Schuster, 2019.
15. J. Patterson and A. Gibson, Deep Learning A Practitioner's Approach. O'Reilly Media, Inc, 2017. [Online]. Available: http://oreilly.com/safari
16. A. Bhatia and B. Kaluza, Machine Learning in Java Second Edition Helpful techniques to design, build, and deploy powerful machine learning applications in Java. Packt Publishing Ltd, 2018.
17. P. Dangeti, Statistics for machine learning. Packt Publishing Ltd, 2017.
18. F. Berisha, "Quality of the predictions: mean absolute error, accuracy and coverage," Sep. 2017