The 2nd Conference on Management, Business, Innovation, Education, and Social Science (CoMBInES)

Taichung, Taiwan 3-6 March, 2022

CORRELATION ANALYSIS BETWEEN ENGLISH LANGUAGE PROFICIENCY AND EASE OF LEARNING PROGRAMMING LANGUAGE: A CASE STUDY ON STUDENTS OF THE FACULTY OF COMPUTER SCIENCE UNIVERSITY OF INTERNATIONAL BATAM

Tony Wibowo, Rosita Tandiono
Faculty of Information System, University International Batam
{tony.wibowo@uib.ac.id, 1837078.rosita@uib.edu}

ABSTRACT

Programming capabilities in one of the most sought skills among the younger generation due to its usage among best-paid jobs in the future. Besides that, communication is also one of the highest tiers in 21st-century skills. In this study, we aim to investigate the relation between English proficiency and computer programming score among computer science undergraduate students in Batam. Data are gathered from 1438 data spanning from 2007 to 2020. Statistical analysis and Machine Learning are used to prove the relation between them results in a strong correlation between English proficiency and computer programming score.
Keywords: *English proficiency, programming, correlation*

INTRODUCTION

The rapid development of technology makes us aware that competition is getting tougher. To be able to compete internationally, we certainly need a way to communicate with others. One of the most important elements in communication is language. English is one of the languages spoken in almost every country. Even some of them make it the official or main languages of their country, such as the United States, Australia, Canada, and New Zealand. One-third of the world's population of about two billion people is estimated to speak English (Welianto, 2020). English has become an international language. In a study conducted by Education First regarding the English proficiency index, Indonesia's ranks in 2017 fell from the previous year, which was ranked 32 to rank 39. The average score of English language proficiency for Indonesia is 52.15. With this rank, Indonesia is far behind Singapore with a score of 66.03 which is also ranked 5th worldwide (Moriand, 2017). In 2020, Indonesia was ranked 61st out of 100 countries with the 50 lowest English proficiency in the world.

Computer programming is the process of designing and creating executable computer programs to achieve specific computing results of calculations or to perform specific tasks. As children gain access to programming environments, the context in which they are used is also relevant to learning (Banerjee et al., 2018). Many programming languages can be used for computer programmings, such as Python, Java, and C#. According to IEEE Spectrum's interactive rank in 2017 which can be seen in Fig. 1, Python is the leading programming language, followed by C, Java, and C++. Of course, the choice of language to use depends on the type of program and the programmer's experience. (Beal, 2018).

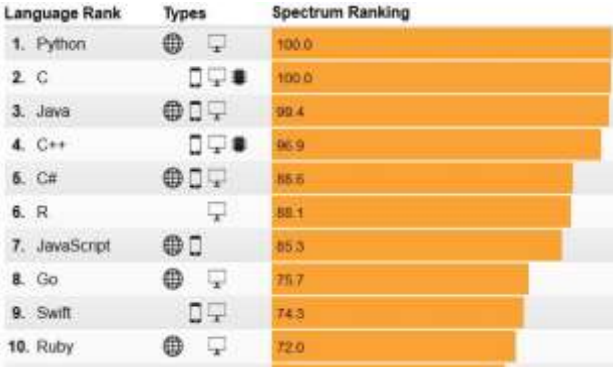

Figure 1. IEEE Spectrum Interactive Ranking (2017)

From the beginning, the programming language has always been linked with pure mathematical logic which does not depend on messy human language. A systematic approach to creating a programming environment relies on the implementation of various mathematical tools and methods (Lazebna et al., 2019). But in fact, programming languages are very much tied to the English language (Guo, 2018). Many people think that as long as we are good at mathematical logic, it will be easy for us to learn programming. Then how about the English language which is closely related to all programming languages? Is it also make us easier to learn programming? Today, information technology is constantly developing and grows. Various companies or organizations are competing with each other. Data is not just a statement or fact of a certain thing. Huge amounts of data can be processed to help companies or organizations achieve their goals. In a study by Jawad & Mughal (2018), the process of using various patterns, intelligent methods, algorithms, and tools to analyze useful information and extract data from large data warehouses is called data mining. According to Joseph (2019), data mining is a multi-disciplinary subfield of computer technology, the process of computing to discover patterns in big data sets, including intelligent methods to extract patterns from data. Based on the conclusions obtained from data mining, a company or organization will be able to make good decisions to change work patterns to improve the progress of the company or organization. Linear with the research conducted by Idris & Ammar (2018) entitled 'The Correlation between Arabic Student's English Proficiency and Their Computer Programming Ability at the University Level'. This research examines the Correlation between Arabic Students' English Proficiency and Computer Programming Ability using a correlation test to their English level with their GPA in programming learning. On other hand, Guo (2018) used an international online survey for non-native English speakers learning computer programming. During their research, they discovered many barriers faced by non-native speakers of English and a desire to improve teaching materials. The research developed by Lestari (2019) entitled "Student Data Clustering Using K-Means Algorithm to Supporting Promotion Strategy (Case Study: STMIK Bina Bangsa Kendari)". The method used by researchers in this study is by performing data mining, using clustering techniques and the K-means algorithm, which produce grouping student based on data which used so that could help STMIK Build Nation Kendari determine strategy promotion which appropriate target. The research conducted by Syahra (2018) entitled 'Application of Data Mining in Grouping Student Value Data for Determining Student Majors at SMA Tamora Using the K-Means Clustering Algorithm'. This research uses data mining with clustering techniques and the K-means algorithm to classify student scores in determining majors.

PROPOSED INNOVATION

In this research, we aim to determine the relationship between English proficiency and programming capabilities using data of computer science undergraduate English proficiency score and computer programming related score using machine learning. The data are taken based on data from Batam International University, which contains data from computer science faculty students from the 2007 to 2020 class. The data will go through several processes, such as data cleaning, data integration, data selection, data transformation, and lastly the data will be used

for clustering. Data analysis was done using the K-means algorithm with Weka 3.8.5 application and SPSS application. In this step, the data clustering for all students using all variables, which are: English proficiency score (TOEIC) and computer programming score. The clustering will be carried out in 2 clusters, the first using 5 clusters and the second using 10 clusters.

METHODOLOGY

The research flow of this study is shown in Fig. 2:

Figure 2. Research Flow
(Sibarani & Omby, 2018)

The first step is writing a literature review. At this stage, we are looking for all sources of knowledge to support our research. We will look for research that is similar to the research we are currently doing. In the second step, we will identify the problem. At this stage, we write down the background, problem formulation, and the objectives and benefits of the research, so that the readers can find out the urgency of this research. The third step is research instrument design. At this stage, we will design what research instruments we will apply in this research to obtain the data we need. In the fourth step, we will implement a research instrument. The research instruments that we have designed and defined will be implemented into the population samples that we have determined. After we implemented the research instrument, we will do data collection. At this stage, we collect sample data from the population that we have determined so that we can proceed to the analysis stage. After the data are collected, then we will do data cleaning. The data cleaning stage is the stage where the process is carried out cleaning of inconsistent, incomplete, and/or duplicate data.

After the data are cleaned, we will do data integration. Data integration or data merging is the stage where data from several sources, for example from several databases or files, are combined into one. Data integration must be carried out to ensure all data can be processed at once in the clustering process in stages next. After doing data integration, we will do data selection. Data selection is the stage of selecting the data to be used in the data mining stage, from all the data that has been collected. Selection data usually takes into account, for example, certain attributes that will be used in the clustering process of all collected data. After we select the data, we will do data transformation. In the data transformation stage, the format or form is changed from the data that has been collected into a form that can be processed by data mining applications. Data transformation is generally often done in data mining because the data collected is usually not directly in the which can be processed by data mining applications. After all data processes are done, then we will analyze the data by clustering data. This stage is the stage where the data that has been collected is selected and transformed into a form suitable for processing for producing certain clusters with the

appropriate members. In the clustering process, we use a data mining application, namely Weka, and the algorithm used is the K-means algorithm. WEKA implements most machine learning algorithms and also visualizes the results (Hussain et al., 2018). After we get the result from the application, we will do data analysis. At this stage, the results of data mining obtained are further evaluated to find patterns that meaning. The output results of the clustering process are data models that contain information from each cluster that has been formed. The patterns that are found will clarify related information such as whether members from each cluster. At the final stage of the research, we will write a report containing the results of this research.

The population in this research were students of the Faculty of Computer Science of University International Batam. The sample of this research is all students consisting of information systems and information technology study programs from 2007 to 2020. The method of selecting the sample is known as the sampling technique. In this research, the sampling technique that we will use is the purposive sampling technique. This technique is widely used because the sample taken is a special population of people who are directly involved in the problem to be studied. The use of purposive sampling in this study aims to determine the correlation between English language proficiency and the ease of learning a programming language.

This research variable is the object of research or what is the point of research. In this study, two variables were used: the independent variable and the dependent variable. In this study, the independent variable is the result of English proficiency (GPA) and the dependent variable is the result of programming language learning (GPA).

The type of research used in this research is qualitative research with research methods, named correlational research. A correlation study is a study that involves data collection activities to determine whether and the level of a relationship between two or more variables. Correlation studies are performed when a researcher wants to know about the presence of a related variable in a subject or subject and the strength of the relationship.

Result and Discussion

In this research, the total number of samples was 1,438 people who were students of the computer science faculty at University International Batam. Based on the data obtained, 0.2% are from 2007 students, 4.2% are from 2008 students, 6.9% are from 2009 students, 7.6% are from 2010 students, 6.5% are from 2011 students, 5.4% are from 2012 students, 5.8% are from 2013 students, 6.4% are from 2014 students, 8.2% are from 2015 students, 5.9% are from 2016 students, 6.8% are from 2017 students, 12.2% are from 2018 students, 12% are from 2019 students, and 11.9% are from 2020 students (See Fig. 3).

Figure 3. Percentage of Student Class Year

From the class year data, there is an average value of programming courses, where 42.2% of students get an average value greater than 80, 43% of students get an average value from 65 to 79, 10.1% of students get an average value from 51 to 64, 2.0% of students get an average value from 40 to 50, and 2.6% of students get an average value below 40 (See Fig. 4).

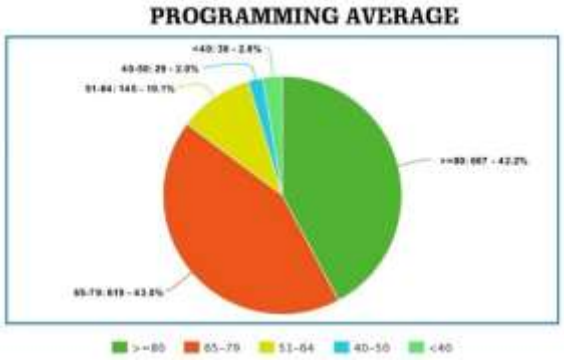

Figure 4. Percentage of Students Programming Average

From the class year data, there are also English scores from the TOEIC test results, where 30.3% of students scored from 785 to 990, 44.6% of students scored from 605 to 780, 15.6% of students scored from 405 to 600, 6.7% of students scored from 255 to 400, and 2.7% of students scored below 250 (See Fig. 5).

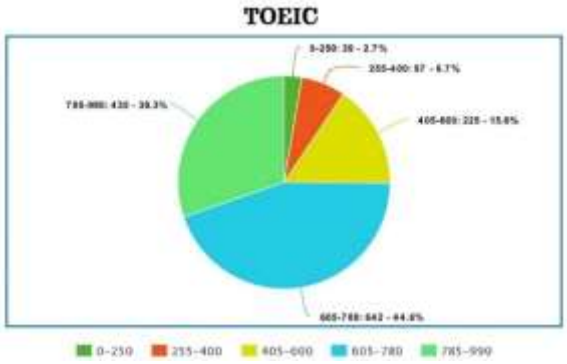

Figure 5. Percentage of Students TOEIC Score

Data correlation testing shows that the significance value for the relationship between the average programming value and the TOEIC value is 0.000, where the significance value is less than 0.05, it can be said that there is a relationship between the average programming value and the TOEIC value (See Fig. 6).

Correlations

		AVG_PEM	TOEIC
AVG_PEM	Pearson Correlation	1	.520**
	Sig. (2-tailed)		.000
	N	1438	1438
TOEIC	Pearson Correlation	.520**	1
	Sig. (2-tailed)	.000	
	N	1438	1438

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 6. SPSS Correlation Test Results

The Pearson Correlation value resulting from this correlation test is 0.520, which shows that the relationship between the average programming value and the TOEIC value has a positive relationship. A positive correlation relationship indicates that the higher the average value of programming, the higher the TOEIC value, and conversely, the lower the average value of programming, the lower the TOEIC value. The clustering process will use the K-means algorithm and the Weka 3.8.5 application. In this step, data clustering for all data will be carried out using the TOEIC value and programming value, where the number of clusters to be used is 5 clusters and 10 clusters. Fig. 7 shows the results of clustering with 5 clusters and Fig. 8 shows the results of clustering with 10 clusters.

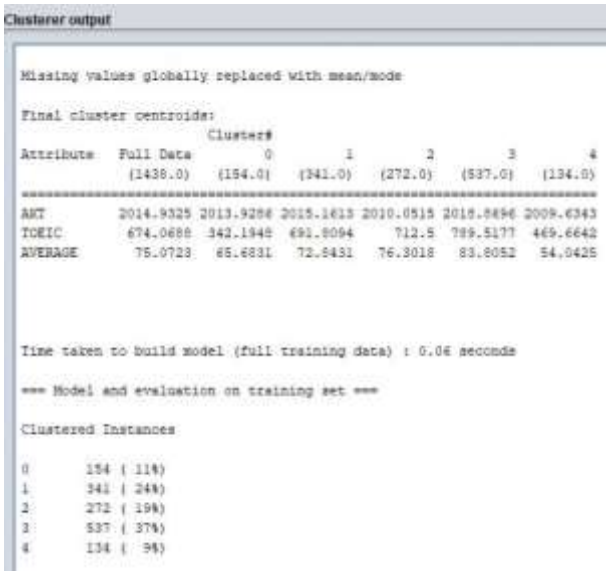

Figure 7. WEKA Clustering

with 5 Cluster

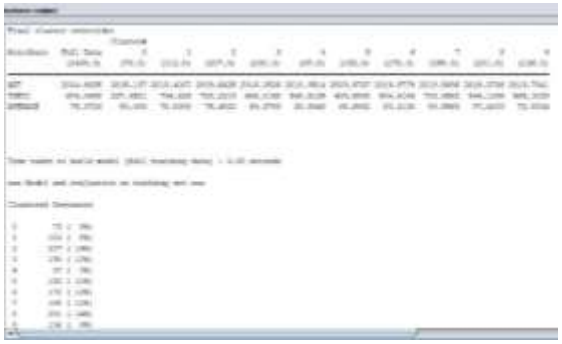

Figure 8. WEKA Clustering with 10 Cluster

After the clustering process is carried out, pattern evaluation is carried out to find meaningful patterns, certain largest clusters will be explained because small clusters will not be too relevant. In the results of data clustering all students (1,438 students) with 5 clusters which are shown in fig. 7, there are 2 largest clusters as follows:

a. The second cluster with 341 students, where the average student has a TOEIC score of 691 and a programming score of 72.8.
b. The fourth cluster with 537 students, where the average student has a TOEIC score of 789 and a programming score of 83.8.

In the results of data clustering for all students (1,438 students) with 10 clusters which are shown in fig. 8, there are 4 largest clusters as follows:

a. The seventh cluster with 172 students, where the average student has a TOEIC score of 804 and a programming score of 80.4.
b. The fourth cluster with 190 students, where the average student has a TOEIC score of 686 and a programming score of 69.3.
c. The third cluster with 207 students, where the average student has a TOEIC score of 728 and a programming score of 75.4.
d. The ninth cluster with 201 students, where the average student has a TOEIC score of 844 and a programming score of 87.4.

As we often encounter, all programming tools and learning materials are delivered in English. Such as workshops, courses, and videos. Even YouTube videos often come from foreign people, so they use English. So basically, people who want to learn programming must understand English because everything will be found in English. For programming learning books, it is also more often written in English because many terms use English in programming. From the evaluation results of the four clusters, it can be concluded that the average student of the faculty of computer science at University International Batam has a fairly high TOEIC score. Almost all computer science students at University International Batam get TOEIC scores above 700. In that way, the average final score in programming courses also gets a fairly high score. Therefore, TOEIC scores and programming course scores are directly proportional. In the correlation test using SPSS, it appears that English language proficiency has a positive relationship. A positive correlation relationship indicates that the higher the average value of programming, the higher the TOEIC value, and conversely, the lower the average value of programming, the lower the TOEIC value. In other words, English proficiency influences us in learning programming. So, the result is that good English proficiency can make it easier for us to learn programming. To make it easier for us to learn programming, we can improve our English proficiency. Several ways to improve our English proficiency are taking English courses, watching more English movies, listening to English songs, speaking English more, or we can also making friends with people with good English skills. People who have good English proficiency will tend to find it easier to learn programming in the future.We can hone our English skills by watching English TV or movies with subtitles or has subtitles. The more often you watch English films, the more vocabulary you will know complete with its meaning. Listening to songs in English is guaranteed to be easy to find because the marketing itself is almost all over the world. We can access English songs starting from songs that are played on television, radio, song applications such as Spotify. The more often you listen to English songs, the easier it

will be to increase your vocabulary. If conditions allow, you also need to build relationships with people who are fluent in English. Because they are often close to people who are used to using English, it has a good impact on their foreign language skills. With friends whose English skills are better, you can learn more things. Your skills are further honed thanks to the many new vocabularies that are known and remembered as well as spoken. So that you have nothing to lose by making friends with people whose English is already outstanding.

LIMITATIONS

In this research, we cannot conclude the data very well because the distribution of student data in the study program is less normal. For the collected data, there are some incomplete data so the results obtained do not represent all existing conditions. The data that we have collected also cannot be displayed, due to confidential data from the university. The data we have is original data so it is the privacy of every University International Batam student.

FUTURE WORK

The analysis of English language proficiency and ease of learning programming was only tested at University International Batam. Therefore, we hope that for further research there can be further development in another university. So that we can know whether it is also proven in other universities or not. In further research, we also hope that other researchers can find out another variable that affects learning programming. In this research, we can improve our English proficiency to make us easier to learn programming. But it is also possible for other researchers to find other variables that we can improve as well to make it easier for us to learn programming.

REFERENCES

Banerjee, R., Liu, L., Sobel, K., Pitt, C., Lee, K. J., Wang, M., Chen, S., Davison, L., Yip, J. C., Ko, A. J., & Popović, Z. (2018). Empowering Families Facing English Literacy Challenges to Jointly Engage in Computer Programming. *Conference on Human Factors in Computing Systems - Proceedings*, 1–13.

Beal, V. (2018). *Programming Language*. Webopedia.Com. https://www.webopedia.com/TERM/P/programming_language.html#:~:text=A programming language is a,FORTRAN%2C Ada%2C and Pascal.

Guo, P. J. (2018). Non-native English speakers learning computer programming: Barriers, desires, and design opportunities. *Conference on Human Factors in Computing Systems - Proceedings*, 1–14.

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational Data Mining and Analysis of Students Academic Performance Using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447–459.

Idris, M. Ben, & Ammar, H. (2018). The Correlation between Arabic Student's English Proficiency and Their Computer Programming Ability at the University Level. *International Journal of Managing Public Sector Information and Communication Technologies*, *9*(1), 1–10.

Jawad, M., & Mughal, H. (2018). Data Mining : Web Data Mining Techniques , Tools and Algorithms : An Overview. *International Journal of Advanced Computer Science and Applications*, *9*(6), 208–215.

Joseph, S. I. T. (2019). Survey of Data Mining Algorithms for Intelligents Computing System. *Journal of Trends in Computer Science and Smart Technology*, *1*(1), 14–23.

Lazebna, N., Fedorova, Y., & Kuznetsova, M. (2019). Scratch Language of Programming vs English Language: Comparing Mathematical and Linguistic Features. *EUREKA, Physics and Engineering*, *1*(6), 34–42.

Lestari, W. (2019). Clustering Data Mahasiswa Menggunakan Algoritma K-Means Untuk Menunjang Strategi Promosi (Studi Kasus : STMIK Bina Bangsa Kendari). *Jurnal Sistem Informasi Dan Sistem*

Komputer, *4*(2), 35–48.

Moriand, A. (2017). *Menurut Riset, Kemampuan Bahasa Inggris Orang Indonesia Masih Rendah*. Kumparan.Com. https://kumparan.com/millennial/menurut-riset-kemampuan-bahasa-inggris-orang-indonesia-masih-rendah/full

Sibarani, R., & Omby. (2018). Algorithma K-Means Clustering Strategi Pemasaran Penerimaan Mahasiswa Baru Universitas Satya Negara Indonesia. *Jurnal Algoritma, Logika, Dan Komputasi*, *1*(2), 44–50.

Syahra, Y. (2018). Penerapan Data Mining Dalam Pengelompokkan Data Nilai Siswa Untuk Penentuan Jurusan Siswa Pada SMA Tamora Menggunakan Algoritma K-Means Clustering. *Sains Dan Komputer*, *17*(2), 228–233.

Welianto, A. (2020). *Kenapa Bahasa Inggris Jadi Bahasa Internasioanl ?* Kompas.Com. https://www.kompas.com/skola/read/2020/02/29/140000369/kenapa-bahasa-inggris-jadi-bahasa-internasional?page=all