Visiting Time Prediction Using Machine Learning Regression Algorithm

Indri Hapsari[1], Isti Surjandari[2], Komarudin[3]
[1,2,3]Industrial Engineering, Universitas Indonesia
[1]indri.hapsari63@ui.ac.id, [2]isti@ie.ui.ac.id, [3]komarudin@ie.ui.ac.id

***Abstract*— Smart tourists cannot be separated with mobile technology. With the gadget, tourist can find information about the destination, or supporting information like transportation, hotel, weather and exchange rate. They need prediction of traveling and visiting time, to arrange their journey. If traveling time has predicted accurately by Google Map using the location feature, visiting time has another issue. Until today, Google detects the user's position based on crowdsourcing data from customer visits to a specific location over the last several weeks. It cannot be denied that this method will give a valid information for the tourists. However, because it needs a lot of data, there are many destinations that have no information about visiting time. From the case study that we used, there are 626 destinations in East Java, Indonesia, and from that amount only 224 destinations or 35.78% has the visiting time. To complete the information and help tourists, this research developed the prediction model for visiting time. For the first data is tested statistically to make sure the model development was using the right method. Multiple linear regression become the common model, because there are six factors that influenced the visiting time, i.e. access, government, rating, number of reviews, number of pictures, and other information. Those factors become the independent variables to predict dependent variable or visiting time. From normality test as the linear regression requirement, the significant value was less than p that means the data cannot pass the statistic test, even though we transformed the data based on the skewness. Because of three of them are ordinal data and the others are interval data, we tried to exclude and include the ordinal by transform it to interval. We also used the Ordinal Logistic Regression by transform the interval data in dependent variable into ordinal data using Expectation Maximization, one of clustering algorithm in machine learning, but the model still did not fit even though we used 5 functions. Then we used the classification algorithm in machine learning by using 5 top algorithm which are Linear Regression, k-Nearest Neighbors, Decision Tree, Support Vector Machines, and Multi-Layer Perceptron. Based on maximum correlation coefficient and minimum root mean square error, Linear Regression with 6 independent variables has the best result with the correlation coefficient 20.41% and root mean square error 48.46%. We also compared with model using 3 independent variable, the best algorithm was still the same but with less performance. Then, the model was loaded to predict the visiting time for other 402 destinations.**

Keywords—visiting time; clustering, classification, machine learning; regression

I. INTRODUCTION

Having control to the journey is an obsession for tourists, because they have limited money and time. If they can reduce the uncertainty, it will be good to have a smooth journey. They want to access information anywhere and anytime, to support their decision. But now those requirements are not enough. They need another comprehensive feature to decide something as a part of decision support system. Tourists only concern to good result without thinking how to achieve the optimal journey.

To develop a journey, beside considering tourists' destination selection and their limitation in time and money, journey also needs the prediction of time. Time in a journey consist of traveling time to move one place to another and visiting time or time to stay in one destination. Traveling time is taken from Google Map and it is trustable because Google can access the location from tourists' mobile phone. All destinations have this information and when they are connected each other, Google can predict the travelling time when users need it. Because of this information, Google can combine it with the public transportation time-table. Using specific algorithm based on the purposes, application will arrange a sequence of destinations to visit or what we call as route.

For visiting time, Google use the same way to predict it by following users' location. From the users' device, Google knows the arrival time and leaving time of each location, especially for tourist destinations and restaurants. Google needs enormous data to recognize it as the pattern. The crowdsourcing data come from huge number of tourist information and not all the destination having it. For example, in this research using case study in East Java, one province in Indonesia, from 626 destinations only 224 have the visiting time. Whereas, tourists need it to support their decision or to support the application.

That is why we need the model prediction to estimate the visiting time. Visiting time become the dependent variable that is influenced by independent variables. Actually, Google Map has many features to be used as information to predict the visiting time, but none of the researchers used the information. Another reason we need to predict the visiting time is Google does not open the database like it does on coordinate location by Google API, so it is easy to measure the traveling time.

The features from Google Map mobile version will be explained with the relationship with 10As of success destination by Morrison [1]. If visiting time is a dependent

variable that we must predict, the feature on Google Map become the independent variables. Those independent variables are access, city, rating, number of reviews, number of pictures, and information (address, URL address, operating hours, number of telephone, and short description). Because the model has more than one independent variables, this research uses multi regression to solve this problem using 224 destinations from 626 destinations in East Java that have the visiting time. All the requirements will be checked statistically before continuing the development of the model.

In Section II we describe all the concept related with the statistics test for normality of the data and algorithms in machine learning. Then we continue with section III about research methodology to predict the data. In Section IV we will discuss and analyze the result for both statistical test and machine learning algorithm, especially for Linear Regression. In the end in Section V there is a conclusion about the strength and weakness of the model prediction and what we could do for the next research.

II. Literature Review

Tourist need a supporting system to do the traveling journey. Now because of the frequently use of internet, it is very easy to access all they need from mobile devices. Application developers have a strong competition to create a system that can fulfill tourist needs. A tourism mobile recommender system is developed to answer the need of the tourist with suggestions. The major requirement is information, but they need more. That is why many applications were developed in various type.

The applications are classified into three groups [2]. The first group is the recommender system that applies as a website or mobile application. Website is easier to build for the web developer but a bit difficult for the user, otherwise mobile application is difficult to build but it needs less effort for the user. Each of those option has the strengths and weakness, and sometimes it is not only the good and bad, but related to the image, a practical consideration, and the policy to manage the system. For example, Kawai et al. [3] created a system that will give the tourists system to search efficient route among several destinations and suggests the path with beautiful scenic site using the information from the Web. GUIDE is a system developed by Cheverst et al. [4] for giving information that combine personality of tourist and environmental information. The second group is about the source of recommender system. First source is collaboration among users where they can share their experience and become information for the others. This source is mining from the social media, geo-tagging or travel website. For example, Bellotti et al. [5] created context-aware mobile recommender system, Magitti. Magitti takes user activities from context and patterns of user behavior to automatically generate recommendations by content matching. Second source is content, it means the developer will create the information and update it. The last source is hybrid that combines collaborative and content to work together. This source will complete each other to get better information. MapMobyRek by Averjanova et al. [6] is an approach for integrating recommendation using electronic map technologies. Kenteris et al. [7] created DailyTRlP, a heuristic approach to give personalized recommendations of daily itinerary for visiting any tourist's favorite destinations. The third group is the use of public transportation or private vehicle. For public transportation it means that they will deal with time dependency. Time dependency is time that connected each other, and departure time in previous destination will influence arrival time in next destination. Tourist will also consider about transit time, or the time they must wait until public vehicle comes. This problem does not happen in single mode transportation. For multimodal transportation, Gavalas et al. [8] created eCOMPASS, a context aware web and mobile application which derive personalized multi-mode transportation via selected destinations. For single mode vehicle, Maruyama et al. [9] developed P-Tour to compute a semi-optimal schedule in reasonable time using genetic algorithms. Gavalas [10] also built Scenic Athens, a context-aware mobile city incorporating scenic by walking.

Most of the routing algorithm in those applications is using Orienteering Problem (OP) algorithm as a base, then it was modified to fit with the certain condition. OP is an algorithm that arranges a set of possible destinations with each score has a goal to maximize the total score of the visited destinations. Orienteering first comes as an outdoor sport that has some place with the score. Players will use compass and map to visit some places and get the points or profit. Players try to maximize their score in limited time, by visiting any places that contain more points. This algorithm becomes an extension of traveling salesman problem (TSP) with the difference is in this method they do not need to visit every place. In the beginning, Tsiligirides [11] approached two heuristics for OP, divided into stochastic algorithm and deterministic algorithm.

Even though there are many modification of OP model, stay time or visiting time become the mandatory data that the algorithm data base must have. From all the previous papers, they generated time without considering the real data. Application developer must concern about visiting time because the user will get the good or bad route arrangement. Application like Google Trip has a good data base by tracking the user GPS that including in their Android device. From those feature Google can predict the visiting time in certain place such as tourist destination, restaurant, mall and many more. The problem is not all the tourist destination has the information like visiting time. Google use crowdsourcing to process it as 'People typically spend xx minutes in here'. There is information about visiting time for popular tourist destination, but not for unpopular destination. Good route must accommodate all the possibility of tourists' choice of destination that will be very diverse. That is why application must offer as much as possible the destination list. The destination information must include the visiting time and traveling time. Traveling time is easier to get because each destination has coordinate or location, and Google Map has prediction time almost real-time.

To predict the visiting time without the crowdsourcing, this research will propose the variables that may influence the visiting time. With this research, researcher can use another

data than Google database if they need to predict visiting time. Until now, there is no offering from Google to use their database of visiting time. It is only the traveling time that we can predict based on coordination from Google API. From Morisson [1] there are 10 key success of a tourist destination (10As), and the more success means the tourist will stay longer in a destination. Because Google Map has reliable data to use, we assume there is relationship between the 10As and Google Map features. Google Map features become independent variable, while visiting time become dependent variable. The 10 keys to success as Morisson [1] stated are Awareness, Attractiveness, Availability, Access, Appearance, Activities, Assurance, Appreciation, Action, and Accountability. Awareness will be related with the tourist knowledge about those destinations and it will affect to the information that is received by the tourist. To measure this, we can use number of review because more review means more popular destination. Attractiveness can be measured by Google Earth to see the landscape and density neighborhood around the destination. For wide or narrow road can be related with factor Access using Google satellite. The high access score will be given if the destination is on the side of main road with a lot of building around it, followed by neighborhood with many buildings, until natural environment without building seen from scale 1:100 meters. Availability show the easiness to book and reserve to visit destination, and how many channels are provided for it. This variable can be represented by the information provided on Google Map like address, phone number, website address, operating hours, and short description about destination. Appearance is measuring the beauty of destination for the first-time tourist or for experienced tourist. Rating will represent tourist score for the destination. More rating means better appearance. Moreover, number of pictures also show the willingness of tourists of its beauty and want to share it. Activities means number of activities that is provided in destination. Number of rating will indicate the tourist's satisfaction for various activities that can be enjoyed in destination. Assurance will be related to safety and security for the tourist. It can be shown by number of rating that represent the tourist satisfaction for the guarantee. Appreciation is tourist feeling to the acceptance and warmness that offered by destination. It is also shown by the number of rating they give. Action is how the city government is planning the tourism for a long-term. From the formal website that is controlled by Dinas Pariwisata, it shows how good a government seriously take the tourism in its city by preparing and keeping the good performance. Accountability is evaluation of the tourism performance by Destination Marketing Organization (DMO). Even though there is no formal DMO scores, number of rating given by the tourists hopefully can represent the performance of destination.

In the beginning we assumed the relationship between dependent and independent variable was linear. Model of predicting visiting time will use multiple regression [12] because there are some independent variables from Google Map feature that will predict one dependent variable. For dependent variable x_j and m independent variables $x_1, x_2, \ldots, x_k, x_m$ is the linear multiple regression of x_j on the independent variables is given by the equation:

$$\hat{x}_j = \alpha_\tau + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \ldots + \beta_m x_m \quad (1)$$

Where the regression coefficient for any of the variables, say β_k for x_k is the linear regression coefficient for the least squares fit of x_i to x_k, holding the other independent variables constant. However, the model must be valid first before it will be used to predict the dependent variable. Some tests must be done before the developing of model prediction. The requirements are [13]:

1. The dependent variable is a linear function of the independent variables or linearity
2. Each person (or other observation) should be drawn independently from the population.
3. The variance of the errors is not a function of any of the independent variables. This assumption is referred to as homoscedasticity.
4. The errors are normally distributed.
5. The dependent variable does not influence any of the independent variables.
6. The independent variables are measured without error, with perfect reliability and validity.
7. The regression must include all common causes of the presumed cause and the presumed effect

The first assumption (linearity) is the most important. If it is not fulfilled, then all estimation like R^2, the regression coefficients, standard errors, tests of statistical significance will bias. If assumptions 2,3 and 4 are not fulfilled, regression coefficients will not be biased, but standard errors will not be accurate.

If the data cannot fulfill the linear regression requirement, it must use machine learning to deal with nonlinear. Machine Learning produce a model that can usually be expressed in terms of a few support vectors and can be applied to nonlinear problems. The basic idea is to find a function that approximates the training points well by minimizing the prediction error. When minimizing the error, the risk of overfitting is reduced by simultaneously trying to maximize the flatness of the function.

Brownlee [14], a machine learning specialist stated that there are 5 top algorithms for data prediction. They are Linear Regression, k-Nearest Neighbors, Decision Tree, Support Vector Regression, and Multilayer Perceptron. Linear Regression works by predicting coefficient for line or curve that fits with the data. Linear Regression is simple, fast and has great performance. k-Nearest Neighbors can process classification and regression. This algorithm will store the data for training, then trying to locate the k most similar pattern to make a prediction. As another simple and great performance algorithm, k_Nearest Neighbors has the strengths because it considers the distance between the data to make predictions. Next algorithm is Decision Tree that also can classify and regression like the previous algorithm. This algorithm creates a tree based on evaluating from the root of tree and goes to the leaves until it can give prediction. Decision tree will select the best split point to make predictions, repeat the process until the fixed tree is developed. The next algorithm is Support Vector Regression

that fit the data into the line to minimize the error of cost function. This algorithm will use a optimization process for training data. If the data is not fit with the line, a margin will be added around the line to relax the constraint, that will allow bad predictions to be tolerated as a better prediction. If the data is projected to the higher dimensional space, it can make prediction. Different kernels can be used for projection control and make it more flexible. The last algorithm is Multi-Layer Perceptron or simple neural network that also support classification and regression. It is complex algorithm that has many configuration parameters that maybe only be fit with the pattern through intuition and need a lot of trial and error. In the classification this algorithm will discriminate among the classes, while in regression it will fit the real value output.

III. Methodology

Data will be collected from Google both for dependent variables and independent variables. For dependent variables which is visiting time, we used the feature from Google Map mobile version. The feature that will be used comes from 'People typically spend xx time in here' as dependent variables. If the visiting time come in ranges, such as 45 - 90 minutes, we will use the average. From 626 destinations in East Java only 224 destinations have information about visiting time. For independent variables, model will use Government, Rating, Review, Pictures, Information, and Access. As explained in previous section about 10As of successful destination, Government will represent Action, Rating will represent Appearance, Activities, Assurance and Accountability. Another variable, such as number of Pictures will represent Appearance. Information that consist of address, URL address, operating hour, phone number, and short description will represent Availability. The view from Google Earth will represent Attractiveness and Access. The three independent variables will use ordinal data because the value based on the judgement. For example in Government variables as the representation of how the city government the tourism problem in there area, we will give score 1 if there is no information about it, and 5 for the detail information about tourism that the government publish it in their organization website. If the interval data is needed, the ordinal can be transformed into interval with Method of Successive Interval. Otherwise, the interval data can be transformed into ordinal data by clustering.

For the first model must be pass the statistically tests for linearity, normality, autocorrelation and homoscedasticity. The first two is important so if the data cannot fulfill the test, it would not continue for the remain. For linearity we used Anova to find the significant value and for normality we used one sample Kolmogorov Smirnov. If the data is passed all the test, it continues with multiple linear regression model that needs interval data. If the data is not normal, it will continue with transform the data related to the original plot, negative or positive skewness. Then, the normality test will be run again. If the transformation data is still not normal, we will use another model that is not need normality assumptions, like Ordinal Logistic Regression. The model must pass the Parallel Line Test. Another alternative way is using the classification algorithm in machine learning. Weka will be used to develop the model because it has various regression algorithm as classifier. The top 5 regression algorithm are Linear Regression, k-Nearest Neighbors, Decision Tree, Support Vector Machines, and Multi-Layer Perceptron. The best algorithm is achieved if the output is maximum correlation coefficient and minimum root mean square error. We will try it for 3 and 6 independent variables. We run the data as the training set with 10 cross validation for the whole algorithms and compared them. After the model was chosen, it predicted the 402 destinations as the testing set.

IV. Results and Discussion

The first model must be tested for linearity, normality, autocorrelation and homoscedasticity, before it is decided will be continued to multi linear regression model or other. The test was for 224 destinations that have visiting time information. The first requirements that must be fulfilled is normality. From One Sample Kolmogorov Smirnov test we found all the Asymp. Sig. (2-tailed) was 0 or less than $\alpha = 0.05$, so it means all variables are not normal. Then we continued with linearity test using Anova. From Anova Output Table it was shown that value significant deviation for linearity were more than $p = 0.05$ except for variable Access and Pictures. It means not all the variables are linear. From measure of association output, the R squared values were almost 0. Thus, there was no correlation among those variables. The next steps are transform the data to be normal by checking the plot skewness. From the plot, there were moderate negative skewness for variable Access, Government, Info, Pictures, and Rating, and moderate positive skewness for variable Review and Visiting Time. We transformed each variable, run the normality test again using one sample Kolmogorov Smirnov, but the result was still the same. The Asymp. Sig. (2-tailed) was less than $\alpha = 0.05$, so it means those variables are still not normal. Then we continued with Logistic Ordinal Regression because it does not need the normal assumption, but we must transform the dependent variable from interval to ordinal data. We used Expectation Maximization as clustering algorithm in machine learning that made the interval data into 3 clusters. This algorithm was chosen because it gives maximum cluster based on data characteristics. After parallel line test in Logistic Ordinal Regression there are not function that suitable with the model.

Then, to find out the regression model that will be fit with the data we used classification algorithm in machine learning by WEKA [15]. The classification algorithm will learn from data without relying on rules-based programming, because we have done with the statistic test procedures. We predicted the model based on 224 destinations that have visiting time, to predict 402 destinations without visiting time. Using cross-validation, the algorithm will train the data first, then we continued with the data testing.

WEKA [15] uses 10-fold cross validation provides an average accuracy of the classifier. Weka takes 100% labeled data, it produces 10 equal sized sets. Each set is divided into two groups: 90% labeled data are used for training and 10%

labeled data are used for testing. It produces a classifier with an algorithm from 90% labeled data and applies that on the 10% testing data for set 1. It does the same thing for set 2 to 10 and produces 9 more classifiers. it averages the performance of the 10 classifiers produced from 10 equal sized. k-fold validation reduces the variance and stabilizes the accuracy by averaging over k different data sub-sets. Therefore, the generalization estimate is usually better than just using only one training set and one test set.

We were using 224 destinations with the visiting time as the training data and 6 numeric attributes (Access, Rating, Review, Information, Pictures, Government). There are 5 top algorithms [14] like Linear Regression, k-Nearest Neighbors, Decision Tree, Support Vector Regression, and Multi-Layer Perceptron. Linear Regression is the first algorithm that should be tried because of its simplicity. Support Vector Regression works by finding a line of best fit that minimizes the error of a cost function. Multilayer Perceptron or Neural Networks is a complex algorithm for prediction model to handle so many configuration parameters that usually can only be tuned effectively through intuition and a lot of trial and error. k-Nearest Neighbors also a simple algorithm, but it does not assume very much about the problem except the distance between data instances, so it is meaningful to make predictions. Decision Tree will select the best split point in order to make predictions. Actually, the first three top algorithm is come from linear function. This function will give a score to each category by combining vector of weights. Then the predicted category is the highest score. In equation (1) the β_k for x_k will variative based on the algorithm we used like Support Vector Regression, Multilayer Perceptron, or Linear Regression.

After running the data, we compared in Table I the correlation coefficient, and root mean squared error for 5 regression algorithms.

TABLE I
TOP 5 ALGORITHM PERFORMANCE COMPARISON

Method	Correlation	RMSE
LinearRegression	20.41	48.47
lazy.IBk	19.74	61.25
REPTree	13.36	51.39
SMOreg	14.85	49.66
MultilayerPerceptron	2.61	63.93

It was found that Linear Regression has the lowest error and highest correlation after 10 folds cross-validation. The correlation among variables can be seen as it was stated below.

$$VisitingTime = -8.68Access + 9.45Information - 0.18Pictures + 127.69 \quad (2)$$

We load the model as a prediction for 402 destinations that do not have visiting time. Visiting time has a range between 63.17 minutes to 152.22 minutes. For example, Mountain Wilis at Madiun has score 1 for Access, Information and Picture, With the linear regression model it will give visiting time 128.81 minutes. With the same logic all destinations will have its visiting time by prediction.

Linear Regression function on Weka gives different value with Linear Regression on SPSS. With the same input, the correlation coefficient on SPSS is smaller than WEKA, it is only 7.8 %, while the error is bigger which is 48.17%. WEKA gives the better result because it can detect and remove high correlation among attributes. On WEKA we can set eliminateColinearAttributes to True to run this function. WEKA can select relevant attributes using attributeSelectionMethod to avoid unrelated attributes that make worse performance. WEKA also uses a ridge regularization technique to reduce the complexity of the learned model, that usually need some assumptions to process it on SPSS. The technique will minimize the square of the absolute sum of the learned coefficients to prevent from larger coefficient (Brownlee, 1998).

V. CONCLUSION

Predicting the visiting time is a mandatory for applications that have function to develop a good route for the traveler. Even though we can use information from Google crowdsourcing in the future, it still becomes a requirement to analyze all the factors that make tourist will stay longer than other destination. The 10As of successful destination become a clear guidance that must be followed by the destination management. Because there are six independent variables that influence the dependent variable which is visiting time, we will use multiple linear regression to predict it. To make sure the data is linear, we did the linearity and normality test. From those test, the result is this data is not linear and normal. Then we used ordinal logistic regression that has no requirement in normality, but the model still not significant. Machine learning come as the solution to predict this data without any assumption. For the first we tried with 5 top algorithms in classification, compared the result and found Linear Regression in Function group has the lowest error and highest correlation. The model helps us to predict the 402 destinations that has no visiting time. The weakness of the model is the error is still high and the correlation is still low. It needs more process to do with the input data and algorithm. The input data could be better if has more instances to make a better fitting, has other attributes to give more characteristic for the model, and should be a change in the way to give value to the instances. For algorithm improvement, it is better to find out the characteristic of each algorithm to get right algorithm and good parameter to give a better result. However, this research has offered the way to predict the dependent variable for unpredicted independent variable distribution data and become alternative besides the crowdsourcing method.

REFERENCES

[1] A. M. Morrison, Marketing and Managing Tourism Destinations, Routledge, 2013

[2] I. Hapsari and I. Surjandari, "Tourism Mobile Recommender Systems: A Survey," 6th IEEE International Conference on Advanced Logistics and Transport (ICALT), 2017, pp. 75-80.

[3] Y. Kawai, J. Zhang, and H. Kawasaki, "Tour Recommendation System Based On Web Information and GIS," IEEE International Conference on Multimedia and Expo, 2009, pp. 990–993.

[4] K. Cheverst, N. Davies, K. Mitchell, A. Friday, and C. Efstratiou, "Developing A Context Aware Electronic Tourist Guide," Proceedings of The SIGCHI Conference on Human Factors in Computing Systems, 2000, pp. 17–24.

[5] V. Bellotti, B. Begole,,E. H. Chi, N. Ducheneaut, J. Fang, E. Isaacs, T. King, M. W. Newman, K. Partridge, B. Price, P. Rasmussen, M. Roberts, D. J. Schiano, and A. Walendowski, "Activity-based Serendipitous Recommendations with The Magitti Mobile Leisure Guide," Proceedings of The 26th Annual SIGCHI Conference on Human Factors in Computing Systems, 2008, pp. 1157–1166.

[6] O. Averjanova, F. Ricci, and Q. N. Nguyen, "Map-based Interaction with A Conversational Mobile Recommender System," Proceedings The 2nd International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2008, pp. 212–218.

[7] M. Kenteris, D. Gavalas, G. Pantziou, and C. Konstantopoulos, C. "Near-optimal Personalized Daily Itineraries for A Mobile Tourist Guide," Proceedings IEEE Symposium on Computers and Communications, 2010, pp. 862–864.

[8] D. Gavalas, V. Kasapakis, C. Konstantopoulos, G. Pantziou, N. Vathis, and C. Zaroliagis, "The eCOMPASS Multimodal Tourist Tour Planner," Expert System Application, 2015, Vol. 42(21), pp. 7303–7316.

[9] A. Maruyama, N. Shibata, Y. Murata, K. Yasumoto, and M. Ito, "P-TOUR : A Personal Navigation System," World Congress on ITS, 2004, Vol. 1, pp. 18–21.

[10] D. Gavalas, V. Kasapakis, G. Pantziou, C. Konstantopoulos, N. Vathis, K. Mastakas and C. Zaroliagis, "Scenic Athens: A Personalized Scenic Route Planner for Tourists," IEEE Symposium on Computers and Communication, 2016, pp. 1151–1156.

[11] T. Tsiligirides, "Heuristic Methods Applied to Orienteering," Journal Operation Research Society, 1984, Vol. 35(9), pp. 797–809.

[12] T. Z. Keith, Multiple Regression and Beyond: An Introduction to Multiple Regression and Structural Equation Modeling, Second Edition, Routledge, 2015.

[13] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

[14] J. Brownlee, "How to Use Regression Machine Learning Algorithms in Weka," 2016. [Online] Available: http://machinelearningmastery.com/use-regression-machine-learning-algorithms-weka/ [Accessed 30/01/2018]

[15] I. Witten, E. Frank, M. Hall. Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, Elsevier, 2011.