A PRODUCTIVITY DECISION SUPPORT SYSTEM FOR CONSTRUCTION PROJECTS THROUGH MACHINE LEARNING (ML)

Tarek Mahfouz, Assistant Professor, tmahfouz@bsu.edu
Department of Technology, College of Applied Science & Technology, Ball State University, Muncie, Indiana, USA

ABSTRACT

Productivity rate estimating is usually based on experience. It entails taking into account different factors, some of which are project specific while others are activity related. Conventional methods integrate the experience with stored data from previous projects. Such a process is often labor intensive and inaccurate. Consequently, accurate estimation of productivity rates through lessons learned is essential for efficient planning of projects. This paper proposes an automated decision support tool for productivity rate estimating through Machine Learning (ML). The adopted methodology (1) utilizes data from a set of completed projects; (2) defines a list of factors -that affect productivity estimates- based on comprehensive literature review and previous experiences; and (3) develops and compares the outcomes of automated Support Vector Machines (SVM) and Naïve Bayes (NB) models for the assessment of productivity rates. The research methodology focuses on steel structures related activities including fabrication, delivery, and assembly. The models retrieve the closest case to a newly encountered one, including project description, project's attributes, and activity's attributes, and reports its estimated productivity and duration. This paper defines a pilot study aiming at developing a comprehensive automated estimator for the construction industry. The outcomes of the current research illustrate the potential of ML modeling to be adopted for assigning productivity rates, making it a powerful tool for decision making.

Keywords: Decision support, Machine Learning (ML), Support Vector Machines (SVM), Naïve Bayes (NB), Productivity Estimating

1. INTRODUCTION

Construction endeavors are described as complex undertakings that produce nonstandard components (Clough et al. 2000). Consequently, productivity estimates of construction activities, like many other aspects of construction projects, are based on a number of integrated and associated parameters. Depending on the nature of the activity, these parameters include, but not limited to, type of material, capacity of the used equipment if any, efficiency of the crew, weather conditions, and experience of the estimator. Adding to these, Dunlop and smith (2003) illustrated that productivity estimates adopted from a previous project must be adjusted for specific site factors. Productivity estimates of certain types of construction activities, like excavation, can be accurately estimated through mathematical equations. However, in construction activities that are dependent on human performance, the accuracy of the estimate degrades due to inconsistency in the human behavior (Stensrud 1998). Consequently, conventional methods of productivity estimating rely mainly on two components, namely personal experience and previously stored data. Such a process is often labor intensive and inaccurate (Karshenas and Tse 2002). In addition, as illustrated by Graham and Smith (2004), due to the complexity of construction activity, there is a difference between actual productivity rates and estimated ones.

Consequently, accurate estimation of productivity rates through lessons learned is essential for robust and efficient planning of construction projects.

In an attempt to mitigate these drawbacks, a number of researches developed models to facilitate productivity estimating in an efficient manner. These methodologies ranged from mathematical modeling (Ibbs 2011, and Jarkas 2012), to Artificial Intelligence (AI) techniques like Case-Based Reasoning (CBR) (Dzeng and Tommelien 1997, and Yau and Yang 1998), to hybrid mechanisms (Graham and Smith 2004). These studies have made major accomplishments in their fields; however, none of them have utilized the induction learning feature of Support Vector Machines (SVM) and Bayesian theory (Naïve Bayes NB) to capture the associations between the different parameters affecting productivity estimating of construction activities. To fill this research gap and complement previously developed model, this paper presents an initial investigative step of the potentials of the aforementioned two Machine Learning (ML) techniques to estimate productivity of construction activities. To that end, the adopted research methodology focuses on steel structure related activities. It is instigated that this research task will provide insight on the performance of SVM and NB for construction estimating. In addition, it will pave the way for the development of a comprehensive automated estimator for the construction industry though identification of the pitfalls and area that requires further investigation.

The rest of the body of this paper describes the followings.

- Background;
    - Literature Review;
    - Support Vector Machines (SVM)
    - Naive Bayes (NB);
- Methodology;
- Results and discussion; and
- Conclusion, Limitations, and Future Work.

2. BACKGROUND

2.1 Literature review

The construction industry represents a perfect example of a knowledge-intensive industry where decision making is extensively based on experience (Kovacevic et al. 2008, and Yua and Yang 1998). Estimating productivity rates of construction activity is one of those examples were experiences gained in previous projects are employed. For newly encountered projects, estimators utilize their gathered expertise to modify stored data from previous projects according to specific project attributes (Graham and Smith 2004). These processes are often cumbersome and time consuming. Consequently, facilitating the use of this knowledge through lessons learned to assist decision making in relation to productivity estimating became a necessity to enhance project control and to focus the professional expertise on value-adding tasks. As a result, a number of research studies have attempted to address this need. Through mathematical modeling, Jarkas (2012) utilized multiple categorical regression to capture the effects of rebar diameter, quantity of reinforcement, wall thickness and geometry, as well as curvature intensity on rebar installation activities. Jarkas research provided recommended mechanisms to improve labor productivity in the aforementioned activity. Furthermore, Ibbs (2011) evaluated the impact of change orders on cost, schedule, and productivity in construction projects. Other researchers utilized CBR to tackle this problem. Dzeng and Tommelien (1997) developed the “CasePlan” system to facilitate schedule generation in relation to power plant boiler erection. The system utilized project attributes and annotated cases to select a similar previously developed schedule for a new project. Yau and Yang (1998) developed a CBR system to estimate costs and durations of construction projects in the preliminary design stage. In 2004, Graham and Smith developed and compared productivity estimators for concrete activities through mathematical modeling (similarity measures) and decision trees. The current research task augments the aforementioned developments through investigating the possibility of adapting two ML

methodologies, namely Support Vector Machines (SVM) and Naïve Bayes (NB), for estimating productivity rates of steel structure related activities.

2.2 Support Vector Machines (SVM)

"SVM are learning systems that use hypothesis space of linear functions in high dimensional space, trained with a learning algorithm from optimization theory that implements a learning biased derived from statistical learning theory" (Shawe-Taylor and Cristianini 2000). In other words, SVM algorithms aim at separating a set of instances into two groups based on specific pre-defined characteristics "attributes". In its simplest form, Linear SVM (LSVM), SVM defines a separation surface (hyperplane) in accordance with equation 1 that best separates the analyzed instances into groups, where ones includes a set of positively classified cases and the other defines a set of negatively classified ones as shown in Figure 1a (Mahfouz et al. 2010). Due to the existence of more than one separation plane, SVM finds the optimum hyperplane that separates the input data with maximum geometric margin (γ) (Figure 1b).

$$y=<\mathbf{w}.\mathbf{x}>+b \tag{1}$$

where vector **w** (weight vector), **b** (functional bias) are the parameters that control the function of the separation hyperplan, and **x** is the feature vector which may have different representations based on the nature of problem.

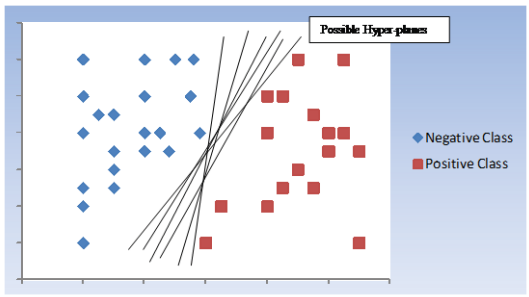

Figure 1a. Possible Hyperplanes in SVM

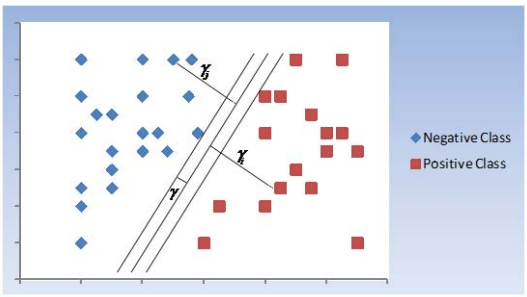

Figure 1b. Hyperplane Representation in SVM

Where γ_i and γ_j defines the Euclidean distance of two points from the decision boundary in the input space; and γ is the maximum geometric margin over all possible hyperplanes

SVM are applicable not only to problems of binary nature but also to multiclass classification nature. For a sample space X and output space Y, a binary classification problem will have Y= {-1, 1} while a multiclass one will have Y= {1, 2 …, m}.

In fact, not all classification problems are linearly separable. As the complexity of the analyzed problem increases, the representation of the defined attribute values becomes more complex. This mandates the representation of the cases in a higher dimension space in order to be linearly separate them. The literature of the ML domain defines Kernel representation as a solution to this problem (Shawe-Taylor and Cristianini 2000, Shawe-Taylor and Cristianini 1999, Platt 1999, and Mangasarian and Musicant 1999). By introducing the appropriate Kernel function to equation 1, the data set can be transformed and plotted in a higher feature space (equation 2). Such transformation converts the problem from being linearly inseparable to a linearly separable one (Figure 2).

$$\mathbf{x}=(x_1,\ldots,x_n)\rightarrow\emptyset(\mathbf{x})=\left(\emptyset_1(\mathbf{x}_1),\ldots,\emptyset_n(\mathbf{x}_n)\right) \text{ or } k(\mathbf{x},\mathbf{y})=[\emptyset(\mathbf{x}).\emptyset(\mathbf{y})] \tag{2}$$

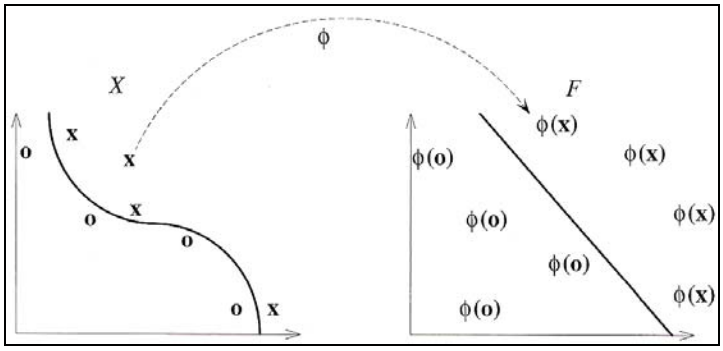

Figure 2: Kernel Transformation (Mangasarian and Musicant, 1999)

Within the current research task, (1) cases are defined as specific steel activity; (2) attributes are defined as the characteristics pertinent to each steel activity. The mechanism of generating these characteristics and their values is discussed in the methodology section; and (3) classes are defined as the average productivity rates related to each steel activity.

2.3 Naïve Bayes (NB)

The Naïve Bayes classification methodology is based on conditional probability. In the training step of NB model development, the likelihood of a specific outcome is calculated based on existing conditions in previously encountered instances. Consequently, it finds the most likely possible classification for an a newly encountered instance among all available ones taking into consideration the presence of prior knowledge its attributes. It is based on the work of Reverend Thomas Bayes (1702 – 1761), who was an English Presbyterian and Mathematician that is considered to be the first to apply Probability Theory, the basis of Naïve Bayes Classifiers, in an inductive manner. For more illustration, the NB classifier calculates the odds of a newly encountered steel activity being classified to a specific productivity rate while having prior knowledge of the activity related characteristics. A decision is made based on the highest calculated probability for all productivity classes. NB assumes that all steel activities are mutually exclusive and exhaustive. Consequently, it assumes that any newly analyzed activity falls into one productivity class and cannot be classified to more than one, hence the Naïve component of the name.

3. METHODOLOGY

The following section of the paper describes the different steps of the adopted research methodology. To that end, the proposed research plan (Figure 3) is fourfold (1) problem identification; (2) factors identification; (3) model development and implementation; and (4) model testing and validation.

3.1 Problem Identification

The current research task focuses on steel structure related activities including fabrication, delivery, and installation. To that end, four completed projects were utilized for the model development. Steel structures under consideration are restricted to hangers, warehouses, and parking sheds. The total number of activities were adopted based on their breakdown in the schedules of the analyzed projects. In other words, if the scheduled detailed the assembly activities into modules, columns and bracing of zones (i=1…m), cladding of zones (c=1…n), roof trusses (r=1…p), and miscellaneous (m=1…j), the total number of activities pertinent to this project is m+n+p+j. Similar analogy was followed for the fabrication, and delivery activities. The total number of identified activities are 59.

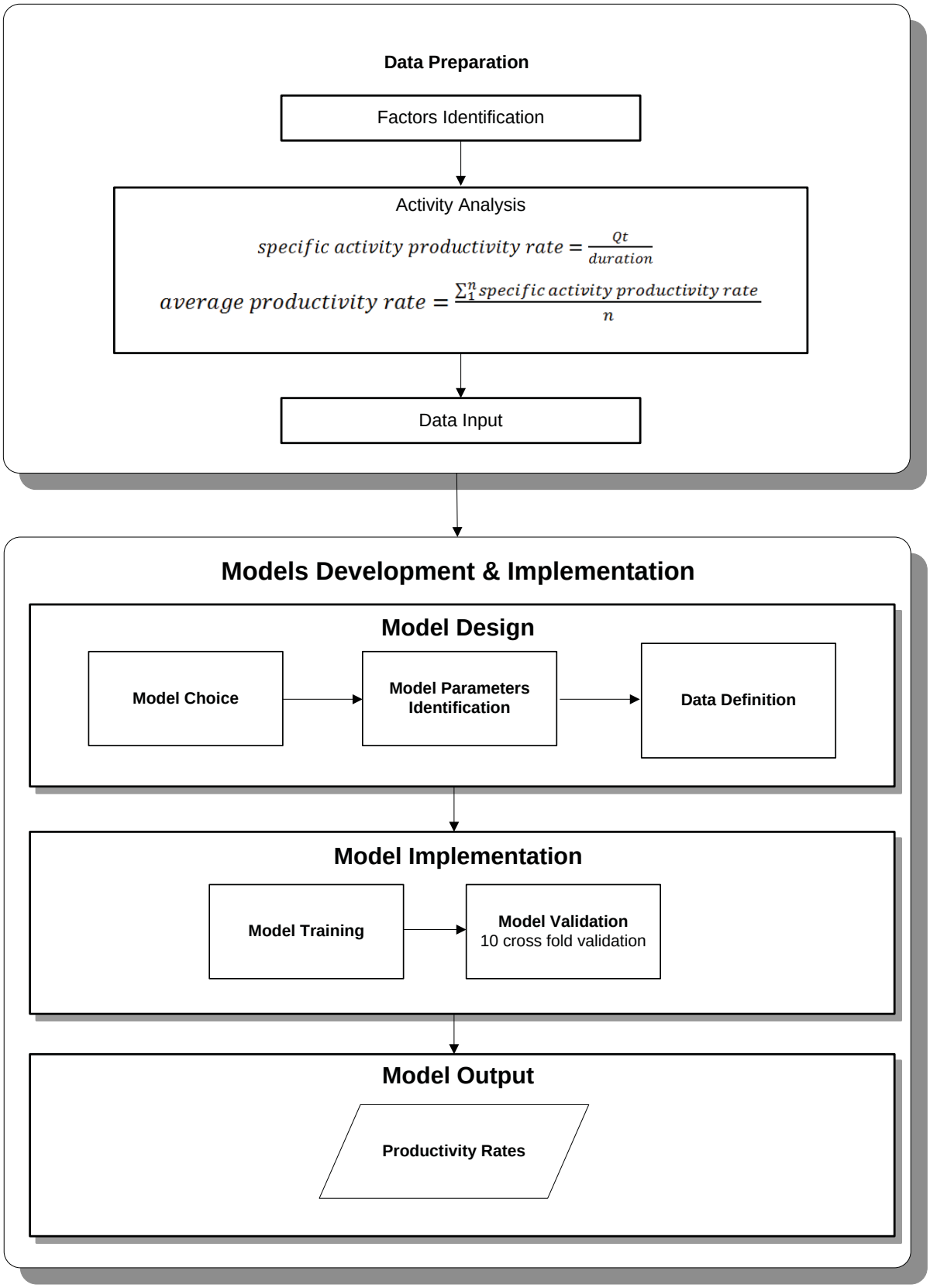

Figure 3: Research Methodology

3.2 Factors Identification

Since this paper represents an initial step in a study aiming at finding a robust methodology for automating estimating processes for the construction industry, the focus of this sub-task is to define an inclusive list of factors that affect the prediction accuracy of the ML models. Consequently, the factors under consideration for the current research task were defined in two steps. First, a compressive literature review of factors identified in similar researches activities (Skitmore 1991, Akintoye and Fitzgerald 2000, Trost and Oberlender 2003, Serpell 2004, Graham and Smith 2004, An et al. 2007, and Mahfouz 2011) were gathered. Second, these factors were evaluated by the author, to eliminate unrelated ones and augment missed factors, yielding a total of 28 factors. Table 1 illustrates the full list of factors definitions and their types.

Table 1: List of Utilized Factors

Item	Factor Definition	Type of Factor
1	Size of the building	Ft^2
2	Type of the building	Shed, Warehouse, Simple hangar, Hangar with attachments
3	Type of the activity	Procurement, Fabrication, Delivery, Installation
4	Building design	Pre-engineered, fully engineered, not designed
5	Experience of the estimator with similar projects	Ordinal 5:high–1:low
6	Experience of the contractor with similar projects	Ordinal 5:high–1:low
7	Experience of the fabricator with similar projects	Ordinal 5:high–1:low
8	Type of the Client	Local, Foreign
9	Type of Material Specified	Local, Imported
10	Fabricator Location	Local, Foreign
11	Details of existing data	Ordinal 5:high–0:none
12	Level of details of the project drawings	Ordinal 5:high–1:low
13	Level of details of the project technical specification	Ordinal 5:high–1:low
14	Financial capacity of the company	Ordinal 5:high–1:low
15	Financial capacity of the client	Ordinal 5:high–1:low
16	Project Duration	Numerical days
17	Difficulty of the estimating procedures	Ordinal 5:high–1:low
18	Estimator's career experience	Numerical years
19	Estimator's field work experience	Numerical years
20	Estimator's experience with field work in similar projects	Ordinal 5:high–0:none
21	Capacity of the estimating team	Ordinal 5:high–1:low
22	Capacity of the procurement team	Ordinal 5:high–1:low
23	Capacity of the technical office team	Ordinal 5:high–1:low
24	Capacity of the quality control team	Ordinal 5:high–1:low
25	Level of construction difficulty	Ordinal 5:high–1:low
26	Location of the Project	Within isolated area, intermediate density, densely populated area
27	Weather Condition	Dry, Humid, Wet
28	Temperature while construction	High, intermediate, Low

3.3 Model Development and Implementation

Within the current sub-task, the 59 identified activities are utilized for the model development. In addition, the set 28 factors illustrated in Table 1 are used as the input features of the model, while the productivity rates represent the predicted output. To that end, the proposed research methodology developed and compared the outputs of 5 ML models, namely (1) Linear Support Vector Machine (LSVM) model, (2) 1st degree polynomial kernel SVM model; (3) 2nd degree polynomial kernel SVM model; (4) 3rd degree polynomial kernel SVM models; and (5) Naïve Bayes (NB) model. Productivity rates adopted for the model training are the average rates of all activities required to perform a specific task. The specific and average rates are calculated as per equation 3 and 4 respectively.

$$specific\ activity\ productivity\ rate = \frac{Qt}{duration} \tag{3}$$

$$average\ productivity\ rate = \frac{\sum_1^n specific\ activity\ productivity\ rate}{n} \tag{4}$$

Where *Qt* is the quantity of steel related to a specific activity, *duration* is the actual duration in days taken to finish this activity, and *n* is the total number of activities needed to finish a specific task.

For the model development and training, the input features were prepared using Excel and the models were developed through Weka algorithms. The training of each model was performed on 10 fold cross validation scheme, where the data set is broken into two subsets containing 90% and 10% of the instances respectively. The model is first trained on the 90% subset and tested over the 10% one. The process is repeated for 10 times until the model is trained and tested over all instances.

3.4 Testing and Validation

To that end, the performance of the models is evaluated through 4 parameters defined as Prediction Accuracy (*A*), Precision (*P*), Recall (*R*), and F-Measure (refer to equations 5, 6, 7, and 8 respectively). Within each training steps, the 4 parameters are reported and the final measure of performance is calculated as the average values of all 10 steps (Mahfouz and Kandil 2011, and 2010). The following is an elaboration on the 4 measures. The Prediction Accuracy (*A*) of is the ratio between the total number of correctly predicted productivity rates out of all tested activites. Whereas, the Precision (*P*) of the model is a representation of the ratio of instances in which the productivity rate was correctly predicted to the total number of instances predicted with the same productivity rate, whether correct or wrong. On contrast, the Recall (*R*) is the ratio of instances in which the productivity rate was correctly predicted to the total number of instances that should have been predicted with the same productivity rate. Examining equations 6 and 7, and the aforementioned definitions, it becomes clear that there is always a tradeoff between the two performance measures. Consequently, an average measure is reported, which is defined as F-Measure.

$$\text{Accuracy} = t_p / n \tag{5}$$

$$\text{Precision} = t_p / (t_p + f_p) \tag{6}$$

$$\text{Recall} = t_p / (t_p + f_n) \tag{7}$$

$$\text{F- Measure} = 2PR / (R+P) \tag{8}$$

Where t_p is the true positive prediction of the model, f_p is the false positive prediction of the model, and f_n is the false negative prediction of the model.

4. RESULTS AND DISCUSSION

The results of the application of the aforementioned research methodology is illustrated in Table 2 and Figure 4. A closer examination of the table shows:

- The highest prediction rate of 70% was achieved through Naïve Bayes (NB) model. In addition, the model attained an increase in prediction accuracy amounting to 47%, 43%, 18%, and 17% over the Linear (LSVM), 1st degree Polynomial, 2nd degree polynomial, and 3rd Degree Polynomial Support Vector Machines (SVM) models respectively. This could be attributed to the number of instances analyzed. It was noticed in previous researches that as the number of instances exceeds twice the number of features, the induction learning feature of the SVM model enhances. In comparison, the conditional probability applied by the NB allows the model to predict an outcome at higher accuracy, even within a small sample (Oracle 2009, Mahfouz and Kandil 2010).
- The trend of enhanced performance of the NB over the developed models extends to the other performance measures.
    - The NB model recorded an increase of 30%, 19%, 18%, and 9% over the LSVM, 1st degree Polynomial, 2nd degree polynomial, and 3rd Degree Polynomial SVM models respectively, with respect to recall.
    - In regards to the average model precision, NB achieved an increase of 14%, 8%, 7%, and 5% over the LSVM, 1st degree Polynomial, 2nd degree polynomial, and 3rd Degree Polynomial SVM models respectively.
    - Consequently, an increase of 22%, 13%, 12%, and 7% were scored in the F-Measure of NB over LSVM, 1st degree Polynomial, 2nd degree polynomial, and 3rd Degree Polynomial SVM models respectively.

Table 2: Average Performance Measures

	Developed Models				
Performance Measure	LSVM	1st Degree SVM	2nd Degree SVM	3rd degree SVM	NB
Accuracy	23%	27%	52%	53%	70%
Recall	33%	44%	45%	54%	63%
Precision	44%	50%	51%	53%	58%
F-Measure	38%	47%	48%	51%	60%

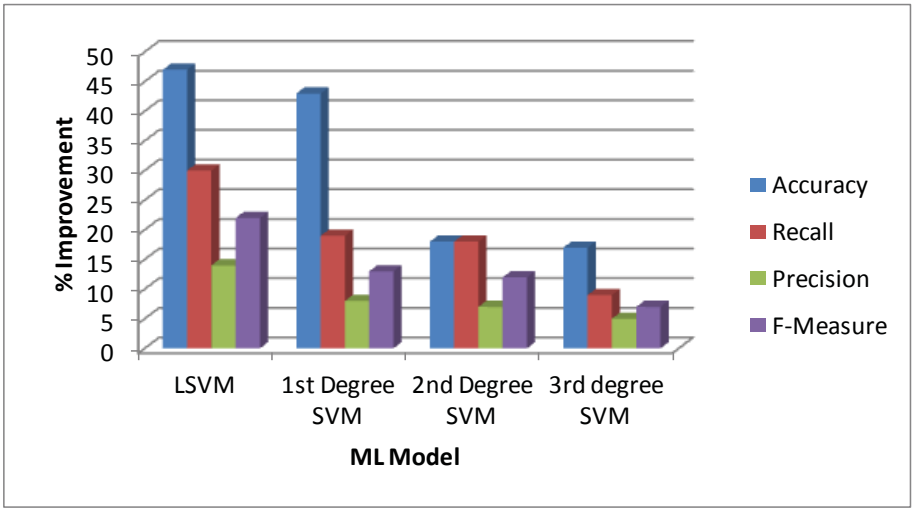

Figure 4: NB Model Improvement

5. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

The paper presented a study that represents an initial step in developing a robust productivity estimating methodology for the construction industry through Machine Learning (ML). For the proof of the concept, the current research focused on steel structures related activities including fabrication, delivery, and assembly. To that end, the adopted research methodology (1) utilized data from 4 projects, where 59 isolated activities were analyzed; (2) identified 28 features of these activities; (3) developed 5 ML productivity prediction models; and (5) used the 59 activities and 28 features for the models training and validation. The outcomes of the implementation of the aforementioned methodology yielded that a Naïve Bayes (NB) model is the most suited among the developed ones, for it has attained the highest performance measures namely, Prediction Accuracy of 71%, average recall of 64%, average precision of 59, and average F-Measure of 61%. This research is instigated to complement previously performed research activities in this field and to pave the way for further development and expansion of the current adopted methodology.

However, the presented research has a set of limitation that should be acknowledged. First, the problem analyzed is complicated due to the interaction and associations between a wide range of factors. Consequently, a more comprehensive list of factors should generated through interviews and opinion assessment of experts in the field. In addition, a Golden Standard of human performance in relation to the current analyzed problem should be established and compared to the model performance. Second, these factors should be analyzed statistically to define a list of features that are statistically significant for the prediction of productivity rates. Third, a larger number of projects should be included in the analysis to increase the learning capabilities of the ML models and in return increase the prediction accuracy. These items are currently under investigation by the author and will be addressed under future works.

REFERENCES

Akintoye, A., and Fitzgerald, E. (2000). "A survey of current cost estimating practices in the UK." Constr. Manage. Econ., 18, 161–172.

An, S., Park, U., Kang, K., Cho, M., and Cho, H. (2007). "Application of Support Vector Machines in Assessing Conceptual Cost Estimates." J. of Comp. in Civ. Eng., 21 (4), 259-264.

Clough SJ, Fengler KA, Yu IC, Lippok B, Smith RK Jr, Bent AF (2000) Proc Natl Acad Sci USA 97: 9323–9328

Dunlop, P., and Smith, S. (2003). "Estimating Key Characteristics of the Concrete Delivery and Placement Process Using Linear Regression Analysis." Civil Engng Environ Syst, 20(4), 273-290.

Dzeng, R., and Tommelien, I. (1997). "Boiler Erection Scheduling Using Product Models and Case-Based Reasoning." J. Constr. Eng. Manage, 123(2), 338-348.

Graham, D., and Smith, D. (2004). "Estimating the Productivity of Cyclic Construction Operations Using Case-Based Reasoning." Advanced engineering Informatics, 18, 17-28.

Ibbs, W. (2011). "Construction Change: Likelihood, Severity and Impact on Productivity." Journal of Legal Affairs and Dispute Resolution in Engineering and Construction.

Jarkas, A. (2012). "Analysis and Measurement of Buildability Factors Influencing Rebar Installation Labor Productivity of In Situ Reinforced Concrete Walls." J. Archit. Eng. 18, 52-61.

Karshenas, S., and Tse, J. (2002). "A Case-Based Reasoning Approach to Construction cost Estimating." Congress Comp Civil Engng, 113-123.

Kovacevic, M, Nie, J, and Davidson, C (2008). "Providing Answers to Questions from Automatically Collected Web Pages for Intelligent Decision Making in the Construction Sector." J. Comput. Civ. Eng., Volume 22, Issue 1, pp. 3-13.

Mahfouz, T. and Kandil, A. (2010). "Construction Legal Decision Support Using Support Vector Machine (SVM)" In the Proceedings of 2010 Construction Research Congress CRC2010, ASCE, Banff, Canada.

Mahfouz, Tarek and Kandil, Amr (2011). “Litigation Outcome Prediction of Differing Site Condition Disputes through Machine Learning Models.” Manuscript Accepted by Journal of Computing in Civil Engineering

Mahfouz, T., Jones, J., and Kandil, A. (2010). "A Machine Learning Approach for Automated Document Classification: A Comparison between SVM and LSA Performances." International Journal Of Engineering Research & Innovation (IERI) 2(2): 53 - 62.

Mahfouz, Tarek (2011). “Unstructured Construction Document Classification Model through Support Vector Machine (SVM)” in the Proceedings of 2011 ASCE Workshop of Computing in Civil Engineering, Miami, FL, USA. P. 126 - 133.

Mangasarian, O. L., and Musicant, D. R., (1999). “Massive Support Vector Regression.” Technical Report Data Mining Institute TR-99-02, University of Wisconsin.

ORACLE. Available online: www.oracle.com/index.html, Last accessed: May 2009.

Platt, J. (1999). “Fast training of support vector machines using sequential minimal optimization.” MIT Press.

Serpell, A. F. (2004). “Towards a knowledge-based assessment of conceptual cost estimates.” Build. Res. Inf., 32 (2), 157–164.

Shawe-Taylor, J. and Cristianini, N. (2000). “Support vector machines and other kernel-based learning methods.” Cambridge University Press.

Sirca, G. F., and Adili, H. (2005). “Case-based reasoning for converting working stress design-based bridge ratings to loaf factor design-based ratings.” Journal of Bridge Engineering, 10(4), 450-459.

Skitmore, M. (1991). “Early stage construction price forecasting: A review of performance.” Occasional paper, Royal Institution of Chartered Surveyors.

Stensrud, E. (1998). “Human Performance Estimating with Analogy and Regression Models: An Empirical Validation.” Software Metrics Symposium,. Proceedings Fifth International, 205 – 213.

Trost, S. M., and Oberlender, G. D. (2003). “Predicting accuracy of early cost estimates using factor analysis and multivariate regression.” J. Constr. Eng. Manage., 129 (2), 198–204.

Yau, N., and Yang, J. (1998). “Case-Based Reasoning in Construction Management.” Comput-Aid Civil Infrastructure Engng, 13(2), 143.