WARSAW UNIVERSITY OF TECHNOLOGY Index 351733
FACULTY OF CIVIL ENGINEERING
COMMITTEE FOR CIVIL AND WATER ENGINEERING
POLISH ACADEMY OF SCIENCES ISSN 1230-2945

DOI: 10.24425/ace.2020.134402
ARCHIVES OF CIVIL ENGINEERING
Vol. LXVI ISSUE 3 2020

PREDICTING THE LENGTH OF A POST-ACCIDENT ABSENCE IN CONSTRUCTION WITH DECISION TREES AND THEIR ENSEMBLES

A. KRAWCZYŃSKA-PIECHNA [1]

Work safety control and analysis of accidents during the construction performance are some of the most important issues of the construction management. The paper focuses on the post-accident absence as an element of the occupational safety management. The occurrence of the post-accident absence of workers can be then treated as an indicator of building performance safety. The ability to estimate its length can also facilitate works planning and scheduling in case of the accident. The paper attempts to answer the question whether it is possible and how to use decision trees and their ensembles to predict the severity of the post-accident absence and which classification algorithm is the most promising to solve the prediction problem. The paper clarifies the model of the prediction problem, introduces 5 different decision tress and different aggregation algorithms in order to build the model. Thanks to the use of aggregation methods it is possible to build classifiers that predict precisely and do not require any initial data treatment, which simplifies the prediction process significantly. To identify the most promising classifier or classifier ensemble the prediction accuracy measures of selected classification algorithms were analyzed. The data to build the model was gathered on national (Polish) construction sites and was taken from literature. Models obtained within simulations can be used to build advisory or safety management systems allowing to detect threats while construction works are being planned or carried out.

Keywords: post-accident absence, decision trees, prediction, classification ensembles

[1] PhD., Eng., Warsaw University of Technology, Faculty of Civil Engineering, Mechanics and Petrochemistry, Łukasiewicza 17 St., 09-402 Płock, Poland, e-mail: Anna.Krawczynska@pw.edu.pl

1. INTRODUCTION

An accident at work is defined in ESAW (European Statistics on Accidents at Work) methodology as a discrete occurrence during work which leads to physical or mental harm. Fatal accidents at work are those that lead to the death of the victim within one year, while non-fatal accidents at work collected within ESAW [6, 7] are those that imply at least four full calendar days of absence from work. For many years, in a majority of EU Member States, the highest incidence of accidents has been reported for persons employed in construction. In addition, for decades, more than half of all absence days registered are concentrated in only three sectors of economic activity: manufacturing, construction and wholesale and retail trade. According to ILO [10] report, averagely, every 10 minutes one construction worker bears death during his work. This puts the construction industry, which was responsible for 30% of all fatal occupational accidents across EU [1], at the head of the most hazardous professions basically due to high variability of working conditions.

Accident statistics are being gathered and analyzed in all countries - in Poland by GUS (Main Statistical Office), in Europe by Eurostat, worldwide by ILO. However, collecting accidents' statistics itself is not the essence of the problem; it is identifying cause-and-effect relationships, the most common accidents' scenarios, factors affecting severity of accidents, and occupational groups burdened with the highest risk of accident. In Poland such studies were conducted by Hoła's team, who developed computer knowledge database and a mathematical model of an accident event development in the construction industry (Hoła & Szóstak in [9], Szóstak in [17]). In turn Drozd [4,5] analyzed costs of accidents and the length of a post-accident absence, while Tehrani et al. [18] assessed the influence of different factors on the safety culture. Baradan, Dikmen and Akboga Kale in [1] proved a significant inverse relationship between the severity of the accident, measured by the fatality rate, and the HDI index (Human Development Index). A study on the above-mentioned papers and global research carried out by Saiful, Razwanul and Tarek [16], Mistikoglu et al. [13], Chua and Goh [2], or Chan, Leung and Liu [3], indicates that there is a strong need to build models of advisory systems supporting H&S management in its various aspects. Such systems should be effective, easy to use, and predict precisely with the historical data. Kim, Cho and Zhang proved in [11] that, thanks to recent advances in BIM technology and computational algorithms, as well as reliable knowledge about performed construction processes, it is possible to create a platform that automatically identifies safety hazards related to activities performed on scaffolding. Such platform outputs preventive measures that can be analysed before the construction begins.

The present paper also matches up with work safety modelling problems. It attempts to answer the question whether it is possible to use decision trees or their ensembles to predict the length of the post-accident absence using historical observations, and which algorithm is the most promising to apply in the prediction problem. Models obtained within simulations can be used to build advisory or safety management systems allowing to detect threats while construction works are being planned or carried out.

2. MATHEMATICAL MODEL OF THE PREDICTION PROBLEM

The paper focuses on the post-accident absence as an element of the occupational safety management. According to Drozd [4], the length of the post-accident absence can be treated as an indicator of building performance safety. It is strongly affected by the working experience of the injured and the size of the company. In Polish conditions the absence decreases with the increasing size of the company, which proves that larger construction companies pay more attention to works safety.

To solve the prediction problem decision trees and their ensembles are suggested. In contrast to other prediction methods, classifier ensembles consist in building different models of the same phenomenon and then combining their judgments [19]. The multi-model supervised learning in statistical data analysis has been used successfully in picture recognition, medical and biological sciences, customer segmentation, business modelling, etc. However, it is still less popular in project and construction planning, which is a pity, because aggregation algorithms, especially those based on decision tress or random forests, as Koronacki and Ćwik in [12] claim, are thought to be the most effective classifiers. They do not require any initial data treatment, pre-processing, or analysis, which simplifies the prediction process significantly. They can also be used in case of missing data, what happens when collecting information on the construction site.

To predict the severity of the post-accidence absence let's consider set $\boldsymbol{U}$ containing $\boldsymbol{N}$ observations which are different historical accidents at work. Each observation is described by a vector of attributes $[x_{i1}, x_{i2}, \ldots, x_{iL}, y_i]$. There are two kinds of attributes: predictors (called the input data) $X_1,\ldots, X_L$ and one target attribute Y (called the output data). Variables $x_{i1}, x_{i2}, \ldots, x_{iL}$, y_i describe attributes' values for the observation i ($i = 1,2,\ldots N$). The value of the target attribute is a class label. Therefore, set $\boldsymbol{U}$ can be defined as (Eq. 2.1):

$$[x_i, y_i]_{N\times(L+1)} = \begin{bmatrix} x_{11} \dots & x_{1L} & y_1 \\ \dots \quad \dots & \dots & \dots \\ x_{N1} \dots & x_{NL} & y_N \end{bmatrix} \tag{2.1}$$

Predictors are the circumstances of the accident, such as: size of the company (the level of employment), work experience of the injured, type of work performed by the injured just before the accident, source of the accident, mechanization of works performed by the injured just before the accident. These data are recorded to the accident reports in Poland. Drozd in [4] explained them and proved their impact on the length of the absence, which decreases simultaneously with an increase of workers' experience and company's size. On the other hand, it lengthens in case of accidents at heights and those caused by workers' improper behaviour.

The target attribute is the severity of the post-accident absence expressed through the absence's length. The goal of the classification problem is to construct, using the historical data, a mathematical model that predicts the class of the unlabeled examples. This means that the dependence between the length of post-accident absence Y and accident's circumstances $\boldsymbol{X} = [X_1,\dots, X_L]$ is being sought.

Classification can be done using a single classifier or a classifier ensemble, where a variety of classifiers (either different types of classifiers or different instantiations of the same classifier) are pooled before a final decision is made. Intuitively and mathematically, classifier ensembles provide an extra degree of freedom in the classical variance tradeoff, allowing solutions that would be difficult or impossible to reach with a single classifier (Oza & Tumer, [14]). There are several methods of combining classifiers: bagging (bootstrap aggregation), boosting, stacking, etc. In the present paper boosting and bagging techniques are being compared. In boosting technique learners are learned sequentially with early learners fitting simple models to the data. The training set used for each member of the series is chosen based on the performance of the earlier classifier in the series (Opitz & Maclin, [15]). Observations wrongly classified by a single classifier receive a higher weight in order to be chosen to the next training set, so the algorithm is "forced" to learn using them. The final classifier arises as a result of weighted component voting. Bagging method is based on a k-fold drawing with replacement of n training sets from the n-element reference set. Each individual classifier in the ensemble is generated with a different random sampling of the training set.

To build prediction models the data collected on 87 different construction sites in Poland and presented by Drozd in [4] was used. The information about accidents was obtained from the statistical accident reports made after each accident. The observed predictors and the range of their historical values are collected in Table 1.

Tab. 1. Predictors and their historical values. Author's elaboration.

Predictor's name	Type of predictor's value	Observed predictors' values
company's size (the level of employment)	numeric (number of company's employees)	from 4 to 241
experience of the injured workers	numeric (number of years of work in construction)	from 1 to 16
type of work performed by the injured just before the accident	binary	1 in case of work at heights, 0 in other cases
the source of the accident	binary	1 – if the accident was caused by the worker himself (worker's psycho-physical state or improper behaviour, incl. not using protective equipment), 0 – if the accident was caused by inappropriate workplace organisation and improper protection of the workplace
the state of mechanization of works performed by the injured just before the accident	binary	1 – if work was performed with equipment being moved or with machine in motion, 0 – in other cases

In the examined data collection the target value (the length of the post-accident absence) varied between 2 and 29 days, while its mean value was 24.4 days and standard deviation was 5.9 days. The problem that is being discussed and proposed methodology of its solution are of a discriminatory type. Therefore, the length of absence, in the first run of calculations, was divided into 6 separate classes (intervals), which at the same time determine the severity of the worker's absence: 0-5 days (very low severity), 6-10 days (low severity), 11-15 days (medium severity), 16-20 days (medium severity), 21-25 days (high severity) and 26-30 days (more than high severity). In the second run of simulations, the severity classes were merged into 3: 0-10 days (low severity, marked L), 11-20 days (medium severity, marked M) and 21-30 days (high severity, marked H).

3. SIMULATIONS' RESULTS

To answer the question whether it is possible to build classifiers to predict the post-accident absence length, and what is the prediction accuracy, two aggregation algorithms were examined: AdaBoost.M1 and Logit Boost and five different classifiers were tested. The most time-effective (the main aspect of selection) classification trees were selected for the comparison, both weak and strong:

- Decision Stump – a weak, one-node decision tree,
- random tree classifier – a class for constructing a tree that considers randomly chosen attributes at each node and performs no pruning,
- J48 tree – a Java implementation of the C4.5 algorithm in Weka data mining tool; it generates an unpruned or a pruned C4.5 decision tree,
- LMT (Logistic Model Tree) – a classifier for building logistic model trees, which are classification trees with logistic regression functions at the leaves; the algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values;
- REP (Reduced-Error Pruning) Tree – a fast decision tree learner, which builds a decision tree using information gain and prunes it using reduced-error pruning (with backfitting).

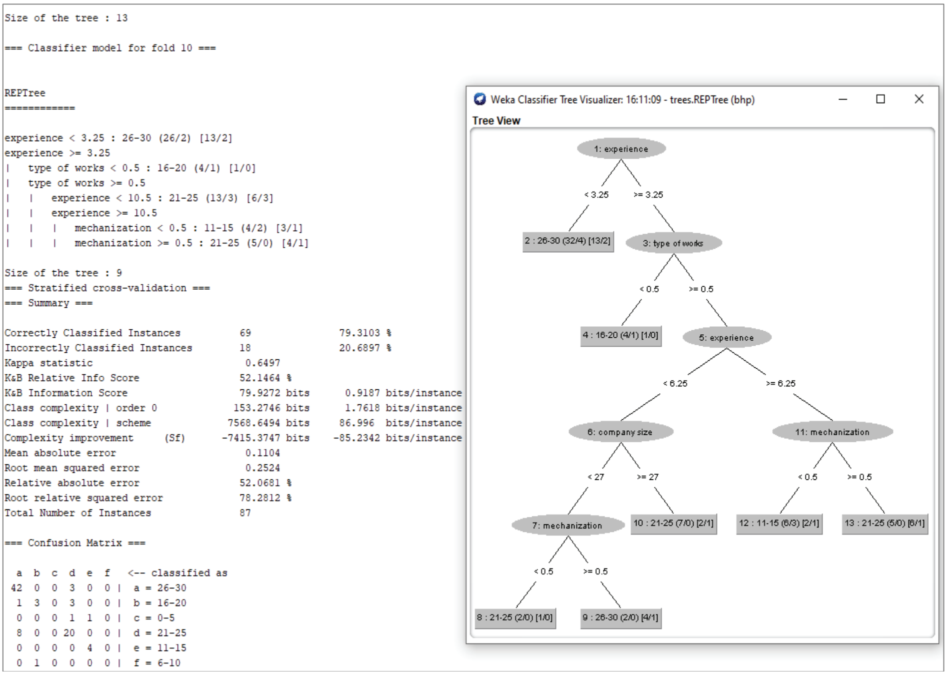
To obtain training sets a 10-fold, 15-fold and leave-one-out cross-validations were performed. The maximum number of iterations was set to 100 and no resampling was allowed. All the calculations were performed in WEKA 3.8. environment.

The best predicting algorithms for the first run of simulations (with 6 classes) are gathered in Table 2. The prediction accuracy is a percentage of correctly classified instances, while MAE stands for a mean absolute error. Fig. 1 presents the best fitting single REP Tree classifier.

The best promising classifiers (J48 and REP Tree) were investigated in the second run of simulations, in which 6 output classes were merged into just 3. The accuracies of the classifiers and their ensembles are gathered in table 3, while Fig. 2 shows the best fitting tree.

Tab. 2. Comparison of the prediction accuracy obtained with different trees and their ensembles. Author's elaboration.

Classifier	Max. prediction accuracy of a single classifier	Aggregation algorithm	Max. prediction accuracy of a classifier ensemble	MAE
Decision Stump	58,6%	AdaBoost.M1	65,3%	0,23
		LogitBoost	77,0%	0,08
LMT	74,7%	AdaBoost.M1	77,0%	0,08
J48	78,1%	AdaBoost.M1	80,5%	0,07
Random tree	73,6%	LogitBoost	77,0%	0,08
REP Tree	79,3%	AdaBoost.M1	79,3%	0,12
		LogitBoost	79,3%	0,16

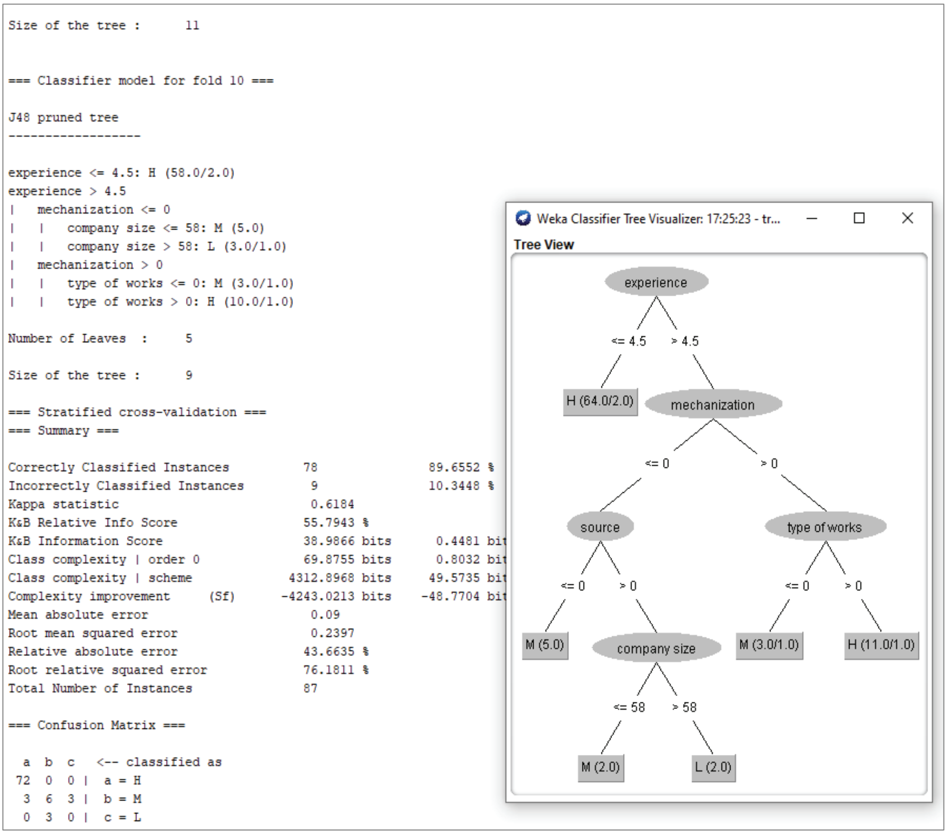


```
Size of the tree : 13

=== Classifier model for fold 10 ===


REPTree
============

experience < 3.25 : 26-30 (26/2) [13/2]
experience >= 3.25
|   type of works < 0.5 : 16-20 (4/1) [1/0]
|   type of works >= 0.5
|   |   experience < 10.5 : 21-25 (13/3) [6/3]
|   |   experience >= 10.5
|   |   |   mechanization < 0.5 : 11-15 (4/2) [3/1]
|   |   |   mechanization >= 0.5 : 21-25 (5/0) [4/1]

Size of the tree : 9
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          69               79.3103 %
Incorrectly Classified Instances        18               20.6897 %
Kappa statistic                          0.6497
K&B Relative Info Score                 52.1464 %
K&B Information Score                   79.9272 bits      0.9187 bits/instance
Class complexity | order 0             153.2746 bits      1.7618 bits/instance
Class complexity | scheme             7568.6494 bits     86.996  bits/instance
Complexity improvement     (Sf)      -7415.3747 bits    -85.2342 bits/instance
Mean absolute error                      0.1104
Root mean squared error                  0.2524
Relative absolute error                 52.0681 %
Root relative squared error             78.2812 %
Total Number of Instances               87

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 42  0  0  3  0  0 |  a = 26-30
  1  3  0  3  0  0 |  b = 16-20
  0  0  0  1  1  0 |  c = 0-5
  8  0  0 20  0  0 |  d = 21-25
  0  0  0  0  4  0 |  e = 11-15
  0  1  0  0  0  0 |  f = 6-10
```

Fig. 1. The REP Tree classifier – classification statistics and decision tree visualisation. Author's elaboration.

Tab. 3. Comparison of the prediction accuracy obtained with J48 and REP Tree classifiers and their ensembles in the second run of calculations. Author's elaboration.

Classifier	Max. prediction accuracy of a single classifier	Aggregation algorithm	Max. prediction accuracy of a classifier ensemble	MAE
J48	89,6%	AdaBoost.M1	93,1 %	0,05
		Bagging	93,1 %	0,09
REP Tree	89,6%	AdaBoost.M1	90,8 %	0,08
		LogitBoost	90,8 %	0,10



```
Size of the tree :      11


=== Classifier model for fold 10 ===

J48 pruned tree
------------------

experience <= 4.5: H (58.0/2.0)
experience > 4.5
|   mechanization <= 0
|   |   company size <= 58: M (5.0)
|   |   company size > 58: L (3.0/1.0)
|   mechanization > 0
|   |   type of works <= 0: M (3.0/1.0)
|   |   type of works > 0: H (10.0/1.0)

Number of Leaves  :      5

Size of the tree :      9

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          78               89.6552 %
Incorrectly Classified Instances         9               10.3448 %
Kappa statistic                          0.6184
K&B Relative Info Score                 55.7943 %
K&B Information Score                   38.9866 bits      0.4481 bit
Class complexity | order 0              69.8755 bits      0.8032 bit
Class complexity | scheme             4312.8968 bits     49.5735 bit
Complexity improvement     (Sf)      -4243.0213 bits    -48.7704 bit
Mean absolute error                      0.09
Root mean squared error                  0.2397
Relative absolute error                 43.6635 %
Root relative squared error             76.1811 %
Total Number of Instances               87

=== Confusion Matrix ===

  a  b  c   <-- classified as
 72  0  0 |  a = H
  3  6  3 |  b = M
  0  3  0 |  c = L
```

Fig. 2. The J48 Tree classifier – classification statistics and decision tree visualisation. Author's elaboration.

4. CONCLUSIONS

Accuracy measures for different ensembles and models built in Weka indicate that, whatever the loss function is, logit (LogitBoost algorithm) or exponential (AdaBoost.M1), classifier ensembles, as expected, provide slightly better prediction accuracy than a single decision trees. The best improvement is observed for the weak Decision Stump classifier, while for J48 and REP Tree the difference is minor. Better prediction for six classes of severity of the post-accident absence (first run of calculations), reaching 81,6%, was obtained only with Random Committee aggregation algorithm and Random Tree classifier being aggregated, which was a result of supplementary calculations. Similar prediction accuracies, at ca. 80%, were achieved within Random Forest method and for different bagging ensembles. It means that value of ca. 80% accuracy is a ceiling value for the data being analysed, whatever the aggregation method is. It's not bad but it's not excellent either.

As expected, the merging of classes (into 3) improved the quality of prediction. However, when analysing the confusion matrix and detailed accuracy measures, such as the F-measure or PRC Area, it should be noted that the prediction accuracy for classes 0-5 / 6-10 and 0-10 (low severity) is very low, regardless of the aggregation algorithm. Thus, in the course of further research and in order to build a reliable decision model it is necessary to expand the knowledge database with observations for which the post-accident absence varies between 0 and 10 days.

The research proved that there is a strong dependence between the length of the post-accident absence and selected accident's circumstances, especially workers' experience. It also proved that time-effective J48 and REP Tree classifiers and their ensembles based on boosting algorithms are a valuable tool to support prediction of the length of post-accident absence, even if the set of historical data used to learn the classifier is relatively small or there are data missing. However, in the course of further research, it is intended to build single classification tress in order to solve the problem. They are, due to their visualization, simpler to interpret and more useful for construction managers, and their prediction accuracy is just slightly (and acceptably) worse comparing to their ensembles.

REFERENCES

1. S. Baradan, S.U. Dikmen, O. Akboga Kale, „Impact of human development on safety consciousness in construction", International Journal of Occupational Safety and Ergonomics 25(1): 40-50, 2019
2. I.Y.S. Chan, M.Y. Leung, A.M.M. Liu, "Occupational health management system: A study of expatriate construction professionals", Accident Analysis & Prevention 93: 280-290, 2016.
3. D.H.K. Chua, Y.M. Goh, "Incident Causation Model for Improving Feedback of Safety Knowledge", Journal of Construction Engineering and Management 130(4): 542-551, 2004.
4. W. Drozd, "Regression analysis of accident absenteeism and variables describing working conditions", Monography 480 – Recent advances in civil engineering: Construction management, Cracow, Cracow University of Technology Publishing, 2015.
5. W. Drozd, "Analysis of Cost Regression and Post-Accident Absence", AIP Conference Proceedings 1863 (1), 230004, 2017.
6. European Statistics on Accident at Work ESAW, 2013. Summary methodology. Eurostat Methodologies and Working Papers. Luxembourg: Publications Office of the European Union.
7. European Statistics on Accident at Work ESAW, 2016. Accidents at work statistics [online]. Available from: http://ec.europa.eu/eurostat/statistics-explained/ index.php?title=Accidents_at_work_statistics
8. E. Gatnar, "Multi-model approach in discrimination and regression", Warsaw, PWN, 2009.
9. B. Hoła, M. Szóstak, "Analysis of the State of the Accident Rate in the Construction Industry in European Union Countries", Archives of Civil Engineering 61(4): 13-34, 2015.
10. International Labor Organization ILO, "A Vision for Sustainable Prevention", XX World Congress on Safety and Health at Work, Global Forum for Prevention, 24-27 August 2014, Frankfurt, Germany, 2014.
11. K. Kim, Y. Cho, S. Zhang, „Integrating work sequences and temporary structures into safety planning: automated scaffolding-related safety hazard identification and prevention in BIM", Automation in Construction 70: 128-142, 2016
12. J. Koronacki, J. Ćwik, "Statistical learning systems", Warsaw, Akademicka Oficyna Wydawnicza Exit, 2008.

13. G. Mistikoglu et al., “Decision tree analysis of construction fall accidents involving roofers”, Expert Systems and Applications 42(4): 2256-2263, 2015.
14. N.C. Oza, K. Tumer, “Classifier ensembles: Select real-world applications”, Information Fusion 9(1): 4-20, 2008.
15. D. Opitz, R. Maclin, “Popular Ensemble Methods: An Empirical Study”, Journal of Artificial Intelligence Research 11: 169-198, 1999.
16. M. Saiful, I. Razwanul, M. Tarek, “Safety Practices and Causes of Fatality in Building Construction Projects: A Case Study for Bangladesh”, Jordan Journal of Civil Engineering 11(2): 267-278, 2017.
17. M. Szóstak, “Modelling of the development of an accident situation in the construction industry” PhD Thesis, Wroclaw University of Technology, Poland, 2017.
18. V. Z. Tehrani, O. Rezaifar, M. Gholhaki, Y. Khosravi, “Investigating factors of safety culture assessment in construction industry projects”, Civil Engineering Journal 5(4): 971-983, 2019
19. M. Walesiak, E. Gatnar, „Data analysis in R”, Warsaw, PWN, 2009.

LIST OF FIGURES AND TABLES:

SZACOWANIE CZASU TRWANIA POWYPADKOWEJ NIEOBECNOŚCI W PRACY W BUDOWNICTWIE Z ZASTOSOWANIEM DRZEW DECYZYJNYCH I ICH RODZIN

Słowa kluczowe*: powypadkowa nieobecność w pracy, drzewa decyzyjne, predykcja, rodziny klasyfikatorów*

STRESZCZENIE

Produkcja budowlana jest jedną z najbardziej wypadkowych – zarówno w kraju, jak i na całym świecie, o czym świadczą badania naukowe oraz liczne statystyki i raporty. O ile liczne statystyki powypadkowe są cennym źródłem danych o wypadkach, o tyle znacznie cenniejsze dla zarządzających bezpieczeństwem na budowie i zajmujących się planowaniem robót są proste w interpretacji modele, pozwalające przewidywać zagrożenia i oceniać ich skutki. Badania w tym obszarze prowadzą m.in. [1,2,3,4,5,17,18]. W pracy skoncentrowano się na zagadnieniu długości nieobecności powypadkowej pracownika. Jest ona bowiem nie tylko uciążliwa dla pracodawcy z przyczyn organizacyjnych, ale także świadczy, co potwierdza [4,5], o poziomie bezpieczeństwa na budowie.

W artykule skupiono się na analizie możliwości predykcji czasu trwania powypadkowej absencji pracownika przy użyciu drzew decyzyjnych i ich rodzin. Przedmiotem rozważań jest zatem N-elementowy zbiór U obserwacji – tj. odnotowanych wypadków w pracy. Każdą obserwację należącą do zbioru U charakteryzuje wektor [x_{i1}, $xi2$, …, x_{iL}, y_i] tzw. atrybutów obserwacji. Wyróżniamy L atrybutów objaśniających (tzw. predyktorów): X_1,…, X_L oraz 1 atrybut objaśniany Y. Zmienne x_{i1}, x_{i2}, …, x_{iL} , y_i opisują wartości atrybutów i-tej obserwacji. Reprezentację zbioru U można zatem zapisać jako równanie (2.1). Dysponując określonym zbiorem obserwacji U, chcemy znaleźć relację pomiędzy długością powypadkowej absencji pracownika Y a okolicznościami wystąpienia wypadku X=[X_1,…, X_L] w postaci modelu. W analizowanym problemie decyzyjnym predyktorami (okolicznościami wypadku) są: wielkość przedsiębiorstwa, w którym pracował poszkodowany, doświadczenie zawodowe poszkodowanego, charakter i stopień mechanizacji prac wykonywanych przed zdarzeniem wypadkowym, źródło wypadku, por. Tab. 1. Do zbudowania drzew użyto dane z 87 budów i zaprezentowane przez Drozda w [4].

Tab. 1. Predyktory i przyjmowane przez nie wartości

Nazwa predyktora	Typ predyktora	Przyjmowane wartości
Wielkość przedsiębiorstwa (poziom zatrudnienia)	Numeryczny	od 4 do 241
Doświadczenie zawodowe poszkodowanego	Numeryczny	od 1 do 16
Typ prac, podczas których nastąpił wypadek	Binarny	1 w przypadku prac na wysokości, 0 w innych przypadkach
Źródło zagrożenia	Binarny	1 – wypadek spowodowany był zachowaniem pracownika (stanem psychofizycznym, zachowaniem, włączając brak właściwych środków ochrony osobistej i wyposażenia), 0 – wypadek spowodowany był niewłaściwą organizacją i zabezpieczeniem miejsca pracy
Stan mechanizacji wykonywanych robot tuż przed wypadkiem	Binarny	1 – praca wykonywana przy użyciu maszyn w ruchu lub podczas ich przemieszczania, 0 – w pozostałych przypadkach

W celu dokonania oszacowania czasu trwania nieobecności powypadkowej pracownika wprowadzono, w 1-wszym toku symulacji, 6 klas (przedziałów) uciążliwości nieobecności: 0-5 dni (bardzo niska), 6-10 dni (niska), 11-15 dni (średnia), 16-20 dni (średnia), 21-25 dni (duża) i 26-30 dni (wyraźnie duża). W 2-gim toku obliczeń, scalono klasy do trzech: 0-10 dni (niska, ozn. L), 11-20 dni (średnia, ozn. M) i 21-30 dni (duża, ozn. H).

W toku symulacji obliczeniowych, analizowano zasadniczo 2 algorytmy agregujące: AdaBoost.M1 i Logit Boost oraz 5 różnych drzew decyzyjnych: typu Decision Stump, drzewo losowe Random tree, drzewo typu J48, LMT i REP. Klasyfikatory budowano w środowisku WEKA 3.8. na zbiorach testowych uzyskanych metodą kroswalidacji (10-cio, 15-to krotnej oraz typu leave-one-out). Najlepsze wyniki predykcji poszczególnych drzew oraz ich rodzin w 1-wszym toku symulacji przedstawiono w Tab. 2. Jakość predykcji po scaleniu klas (2 tok symulacji) dla najbardziej rokujących drzew, tj. J48 i REP Tree, pokazano w Tab. 3, zaś na Rys. 1 i Rys. 2 przedstawiono wizualizacje zbudowanych drzew.

Analiza miar dokładności dopasowania zbudowanych modeli wskazuje, że niezależnie od tego, jaka jest funkcja straty, logitowa (algorytm LogiBoost), czy też wykładnicza (algorytm AdaBoost.M1), zespoły klasyfikatorów zapewniają lepszą jakość predykcji niż pojedyncze klasyfikatory. Ponadto podejście wielomodelowe daje możliwość zastosowania danych wprost, bez analizy rozkładów obserwacji w klasach, co znacząco upraszcza proces wspomagania decyzji. Najbardziej zauważalną różnicę widać dla słabych drzew, podczas gdy dla drzewa J48 i REP, poprawa jakości predykcji jest stosunkowo mała. Najlepszą, wynoszącą 81,6% jakość predykcji dla 6 klas osiągnięto przy użyciu agregacji drzewa losowego Random Tree algorytmem Random Committee. Podobne rezultaty osiągano przy zastosowaniu Lasu Losowego oraz różnych algorytmów typu bagging. To wartość wysoka, jednak nie znakomita. Dla poprawy jakości szacowania scalono klasy z sześciu do trzech. Nie mniej jednak, analiza szczegółowych miar dopasowania, macierzy błędów klasyfikacji itp. wyraźnie wskazują na kłopoty wszystkich budowanych drzew i ich rodzin z klasyfikacją do klasy niskiej uciążliwości nieobecności w pracy (0-10 dni). Wynika to głównie z małej ilości danych wejściowych w tym obszarze. W toku dalszych badań planuje się rozbudowę bazy danych o obserwacji z niską nieobecnością powypadkową, jak również planuje się skupić na samodzielnych drzewach decyzyjnych. Są one łatwiejsze do interpretacji i bardziej przydatne dla kierowników budowy i zarządzających procesem budowlanym, a ich dokładność predykcji może być tylko nieznacznie (i akceptowalnie) gorsza w porównaniu do ich zespołów.

Received: 18.02.2020 Revised: 23.06.2020