RESEARCH ARTICLE OPEN ACCESS

Validated Conceptual Model for Predictive Mapping of Graduates' Skills to Industry Roles Using Machine Learning Techniques

Fullgence Mwachoo Mwakondo
Institute of Computing and Informatics, Technical University of Mombasa, P.O. Box 90420 – 80100
Mombasa - Kenya

ABSTRACT

This paper presents a validated conceptual model for evaluating graduates using machine learning techniques by mapping their problem solving skills to industry jobs' competence requirements. This is because, for college graduates, knowing the right industry role that suits them based on their competences has remained critical when searching for jobs after graduation. Indeed, as thousands of university students graduate each year and enter the market to search for jobs that are limited so then are they exposed to a high risk of not only long search but also job mismatch on employment. In order to enhance both their quick employability and optimal performance in the job, evaluation of graduates' possession of relevant skills is necessary by not only employers but also training institution. In fact, problem solving is one of the skills acquired by graduates during training and strongly sought for by employers during evaluation, yet it is not clear which of its predictor attributes are related to enhanced performance in the job. Besides, evaluation is supposed to be predictive by matching skills possessed by graduates with those required by the job. Thus, predictive evaluation using data-driven techniques such as machine learning may greatly promote graduates' performance in the job. However, we do not have a validated conceptual model for machine learning-based predictive evaluation of graduates skills towards industry roles that can be used by both employers and learning institutions. As a result, there is a mismatch between skills possessed by graduates and those required in the job whose impact is evidenced by high employee turnover, poor productivity and low job motivation. This paper focusses on this gap by addressing two objectives: 1) to outline theoretical conceptual development 2) to develop experimental conceptual validation methodology. Theoretical development was approached through two cognitive dimensions, namely knowledge and skills, and were derived from three cognitive theories. A total of 13 concepts were revealed as follows: 4 independent and 9 confounding. Validity of these concepts was investigated empirically where 5 concepts were confirmed as valid, namely relevant content knowledge, cognitive skills, technical skills, academic capacity, and age. The machine learning implementation of the validated conceptual model recorded an average accuracy of 88.6% on a carefully selected benchmark dataset.

Keywords: - Employability, Mapping, Problem solving skills, Training evaluation, Trends

I. INTRODUCTION

Globally, large number of graduates hold jobs that do not make best use of their skills, 70% in Sub Saharan Africa; 35% in Europe [16]. This suggests that there is a mismatch between skills possessed by graduates and those required in the job. Consequently, this has negative impact to both graduates and employers as evidenced by reduced job satisfaction, high employee turnover, and low productivity [16]; [10]; [3]; [21]. And that is why for college graduates, knowing the right industry role that suits them based on their competences on graduation is critical especially when searching for jobs after graduation. More particularly, as thousands of university students graduate each year and enter the market to search for jobs that are limited, they are exposed to a high risk of not only a long search but also job mismatch on employment [5]. Unfortunately, this trend will continue unless a mitigation is provided to enhance both employability of graduates and their performance in the job.

Ideally, one such mitigation is evaluation by both employers and training institution of graduates' possession of skills that empower them to perform in the job. In fact, problem solving is one of the skills acquired by graduates during training and strongly sought for by employers during evaluation. Besides, evaluation is supposed to be predictive by matching the skills possessed by graduates with those required by the job. Coincidentally, due to wide availability of data predictive evaluation using data-driven techniques such as machine learning may greatly improve and promote graduates performance in the job. However, we do not have a validated conceptual model for predictive evaluation of graduates skills towards industry roles that can be used by both the employers and learning institutions. As a result, the impact is a mismatch between skills possessed by graduates and those required in the job leading to industry academia gap evidenced by high employee turnover, poor productivity and low job motivation [16]; [3]; [21].

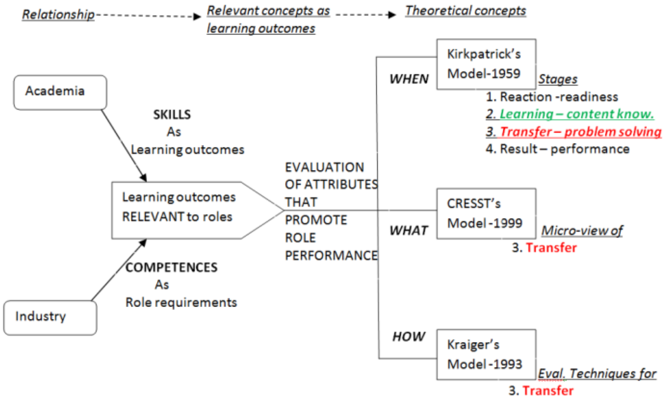
In conclusion, predictive evaluation through mapping of skills to industry roles involves matching and linking graduates' skills with those required by industry roles for job prediction purpose [3]. Surprisingly, this process is useful because it provides feedback to candidates for job suitability [1], easy way for companies to filter high quality candidates before interview [22] as well as credentials to candidates to signal employability [14]. While skill refers to ability to apply knowledge needed to perform a task so as to produce desired results [24], industry role is a well-defined job in an industry occupation that requires a certain set of skills to perform [19]. However, despite efforts to map the skills to industry roles this process has remained difficult due to lack of valid concepts for the modelling methodology in the phenomenon [1]; [21]. Therefore, our research question would be, are there valid concepts that can be used to develop a valid model for predictive mapping of graduates skills to industry roles using machine learning techniques?

II. RELATED WORK

[6] built a classification model for improvement of employee selection by predicting both retention and performance of new job applicants. Although performance of their model was good (80%), the target concepts for mapping were broad. For each role, graduates were mapped not only as either ‘can perform’ or ‘can’t perform’ but also as either ‘retainable or unretainable’, hence in two layered labels. Prediction label was a combination of layer1 (can perform or can’t perform) and layer2 (retainable or unretainable) labels. This way, it was possible to have more than one industry role with similar labels hence multiple label prediction problems. Also, [17] presented a classification model for predicting graduate’s employability. Although performance of their model was good (98%), the target concepts were broad and were mapping graduate’s skills as either employed or unemployed. Whereas target concepts were too broad and therefore not specific to industry roles, most of their ML attributes were not relevant to problem solving skills. [23] presented a model to map graduate’s skills to programmer competences. Although their ML attributes were relevant to problem solving skills, they were just too specific for programmers only and hence domain dependent. Equally, [22] developed a classification model to predict employability by mapping graduate’s skills to software engineer’s role. Although performance of their model was good (82%) and their ML attributes were relevant to problem solving skills, their target concepts for mapping were broad and were mapping graduate’s skills as either satisfactory or unsatisfactory. Besides, it was possible to have more than one industry role with similar labels hence multiple label prediction problems. The current study is an extension of a proposed model for predictive mapping of graduate’s skills to industry roles using machine learning techniques by [18].

III. THEORETICAL DEVELOPMENT

Theoretical development was approached through two cognitive dimensions, namely knowledge and skills, and were derived from three cognitive theories [26]; [2]; [15]. Skills and knowledge were perceived to have been acquired in the Academia as learning outcomes during training, while competences were perceived to be those learning outcomes that were relevant and required to perform jobs in the industry [12]. Further, there was need to understand from theoretical literature those attributes that promote improved performance in the job. Thus, three theories were identified and helped to give insight on when and at what stage was evaluation done during training as well as what attributes and how these attributes were evaluated. This helped to understand these attributes as concepts acquired from content knowledge learned during training and concepts that promote transfer of knowledge and skills outside training environment such as job environment. Fig 1 and 2 illustrate our approach towards theoretical development.



Relationship
Relevant concepts as learning outcomes
Theoretical concepts
Academia
SKILLS As Learning outcomes
Learning outcomes RELEVANT to roles
COMPETENCES As Role requirements
Industry
EVALUATION OF ATTRIBUTES THAT PROMOTE ROLE PERFORMANCE
WHEN
Kirkpatrick’s Model-1959
Stages
1. Reaction -readiness
2. Learning – content know.
3. Transfer – problem solving
4. Result – performance
WHAT
CRESST’s Model-1999
Micro-view of
3. Transfer
HOW
Kraiger’s Model-1993
Eval. Techniques for
3. Transfer

Fig. 1 Towards Theoretical Development

Besides, it was revealed from literature that the adequacy of these concepts to promote performance in the job varied according to some influence from some demographic factors. As a result a total of 13 concepts were revealed as follows: 4 independent and 9 confounding and were used to develop our conceptual model. Figure 3 shows our conceptual model derived from the theoretical development. This, therefore confirms that a model captures relevant features of a phenomenon and these features are derived from theoretical literature as elaborated by [20].

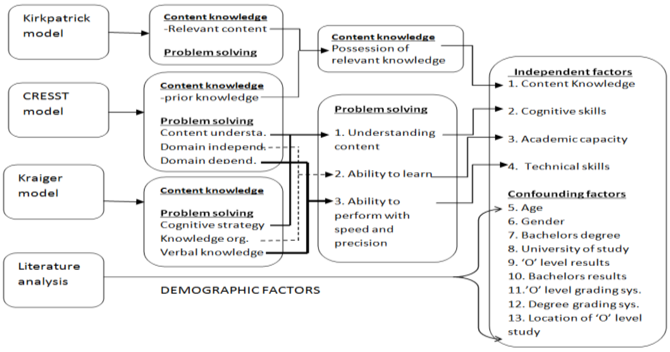


Fig. 2 Theoretical Conceptual Development

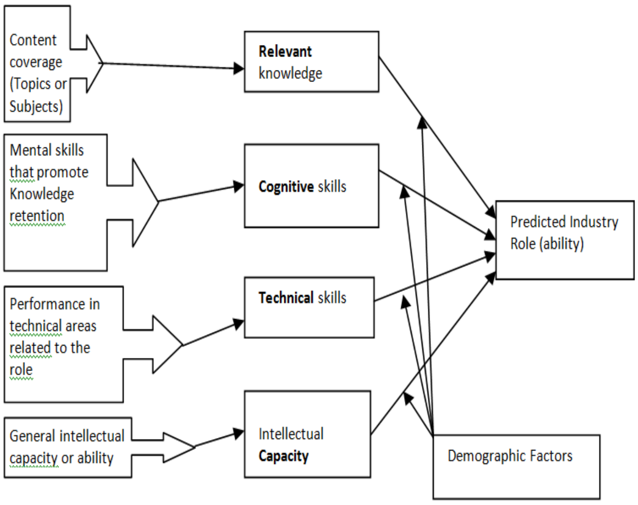

Content coverage (Topics or Subjects)
Relevant knowledge
Mental skills that promote Knowledge retention
Cognitive skills
Performance in technical areas related to the role
Technical skills
General intellectual capacity or ability
Intellectual Capacity
Predicted Industry Role (ability)
Demographic Factors

Fig. 3 Conceptual Model

IV. RESEARCH METHODOLOGY

Initially, theoretical literature analysis provided concepts that characterized the research problem before experimental evaluation validated them. The concepts were operationalized by identifying appropriate indicators and then variables through which data would be collected or recorded. Table 1 illustrates the operationalization of each concept through carefully selected indicators and variables.

Table 1 Operationalization of Conceptual Framework concepts

Concept	Indicator	Variable	Measure-ment
1.Relevant content knowledge	Domain Body of Knowledge	Topic areas of Body of knowledge	Scale
2.Cognitive skills	Cognitive skills areas	Skills areas of Bloom's Taxonomy	Scale
3.Technical skills	Domain technical subjects	Domain technical subjects	Scale
4.Academic capacity	School GPA	Average GPA high school and university	Scale
5.Industry role	Occupational industry roles	Occupational industry roles	Nominal
6. Demographic factors as Confounding factors	Environmental factors:	university of study	Nominal
		Bachelor's Degree type	Nominal
		Location of 'O' level study	Nominal
	Physiological factors:	age	Nominal
		gender	Nominal
	Psychological factors:	'O' level grading system	Nominal
		'O' level results	Nominal
		Degree grading system	Nominal
		Bachelor's Results	Nominal

After carefully searching for a dataset that would suit the purpose of this method, AMEO2015, one of the datasets listed by [1] was selected to validate our conceptual model. The dataset was downloaded from the web link http://research.aspiringminds.com/resources/. The dataset contains data related to entry level engineers, including software engineers. The dataset has 38 attributes and 3998 instances. AMEO2015 is a dataset comprising cognitive skills test scores (AMCAT test scores), bio data details and employability outcomes of job seekers.

AMEO is an acronym for Aspiring Minds Employability Outcomes which is a research affiliated group with the following research objectives: 1) to determine combination of skills needed for various jobs in the market, 2) to provide feedback to candidates on their job suitability, gaps in their skill set for a particular job, and ways for them to improve upon, 3) to provide job credentials to candidates to signal employability, 4) to provide an easy way for companies to filter high quality candidates and provide interview opportunities for them.

In our study, the dataset was carefully analysed to produce a benchmark dataset. This included the following steps:

1) Filtering out all non-software engineers' data records. Specialization column of the data set was used where all non-Computer Science and non-Information Technology data records were removed.

2) Filtering out all trainees and senior software engineers' data record. Designation column was used to remove any data record that implied a trainee or senior software developer/engineer.

3) Filtering out columns or attributes that were not relevant to our study, such as date of joining, job city, personality attributes, salary, etc. Attributes that correlated to our data variables/features in the conceptual model were retained.

4) Deriving data values for variables that were not directly represented in the dataset, such as age was derived from date of birth and date of joining columns, Relevant content knowledge was derived from domain column, cognitive skills was derived from average of English, Logical, and Quant columns, Technical skills was derived from computer programming column, academic Capacity was derived from average of 12percentage (High school exam grade) and collegeGPA columns.

5) Selecting industry roles whose names clearly indicated a well-defined software engineer's role. General names such as software engineers and software developers were ignored.

6) Computing the weights for each of the independent variables for all the industry roles selected.

Table 2 describes the main sources of the benchmark dataset attributes relative to the original secondary dataset.

Table 2: Description of the benchmark dataset

NO	ATTRIBUTES	DESCRIPTION	SOURCE (Column name in the original dataset)
1	GENDER	Gender	GENDER
2	AGE	Age	DOB (Date of Birth)
3	LOLE	Place of O-level Study	CollegeCityTier (2=1, 0=1)
4	GSOLE	Grading System of O-level	12 Grade Exam Board (High School Exam. Board)
5	ROLE	Results for O-level	12 Grade Exam Results (High School Results- 4 classes)
6	BDGREE	Type of Bachelor's Degree	Specialization
7	UNIVERSITY	University of Study for Bachelors	CollegeID
8	GSBDEGREE	Grading System for Bachelors	CollegeTier
9	RBACHELORS	Results for Bachelors	CollegeGPA (grouped into 4 classes)
10	R	Relevant Content Knowledge	Domain (converted to out of 12 points = x12)
17	D	Cognitive Skills	Average (Logical, English, Quant) X12/1000
12	A	Technical skills	ComputerProgramming (X12/1000)
13	C	Intellectual Capacity	Average (12 Grade Exam Result, CollegeGPA)
14	Class	Target Industry Role	Designation (Job title)

Lastly, design of experiments to determine the most relevant attributes or features using the datasets was conducted. Generally, in computing there is a predefined way of carrying out experiments. [13] has elaborately defined the six steps to follow as: 1) conception, 2) design 3) preparation

4) execution 5) analysis 6) dissemination and decision making. Besides, [25] outlines basic principles that should be observed before an experiment is conducted. Following these guidelines, laboratory experiments to validate the proposed conceptual model were carefully developed. Table 3 outlines the experimental design for the experiments. Three algorithms (Logistic Regression, K-Nearest Neighbours, and SVM) were used as experimental subjects. Out of the features generated by each of the three algorithms, features that appeared in at least two of these algorithms were selected.

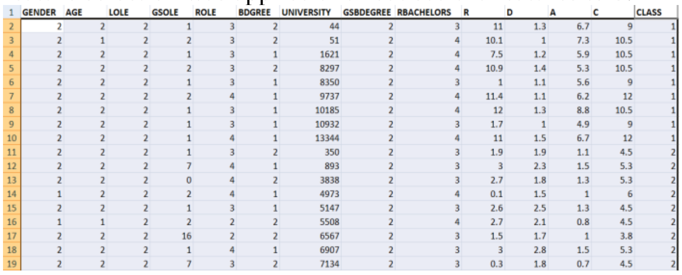
Table 3 Characterization of research experimental design (*adapted from [13])*

Step	Design element	Design Details
1. Conception	Research question	What concepts are appropriate as machine learning attributes for mapping graduates' skills to occupational industry roles?
	Experiment objective	To select relevant features for the model
2. Design	Hypothesis	H_{0A}: All features are equally relevant for better performance of the model
	Experimental unit	Graduate Employees Skills (Software Engineering Domain Literature Benchmark)
	Experimental subjects	ML Algorithms (3 filter algorithms)
	Dependent variable	Performance (accuracy)
	Independent variables	All Features for the conceptual model
3. Preparation & Execution	Data preprocessing	Split (ratio 80:20) dataset into Training dataset, test dataset
	Randomization	6-10 random trials
	Local control	Apply 5-fold cross-validation and sequential backward selection method
	Execution Procedure	❑ Split (ratio 80:20) dataset into 2: train, test sets ❑ Divide features into subsets using combinations of 2 to all features ❑ Train and test 3 filter algorithms on each subset using 5-fold cross-validation ❑ Get the best subset for each algorithm using sequential backward selection method ❑ Select features that appear in at least two of these 3 best subsets
4. Analysis	Analysis criterion	Out of the features generated by each of the three algorithms, select the one that appears in at least two of these algorithms
	Analysis of results (Model Evaluation)	**Evaluation of model accuracy using benchmark (dataset2):** Approach : Hypothesis testing Technique :ANOVA, Paired sample T Test Significance value: 0.05

V. EXPERIMENTAL RESULTS

A. Findings of Secondary Data

A total of 13 variables or features were derived from the dataset with 279 data records (instances) and 12 well defined industry roles. Figure 4a shows a snapshot of the benchmark dataset where the codes adopted in the class columns represented the following industry roles extracted: 1:ios developer(9), 2:data analyst (14), 3:android developer(23), 4:java developer(40), 5:programmer(12), 6:software test engineer(42), 7:systems administrator(9), 8:network engineer(8), 9:php developer(19), 10:web developer(32), 11:programmer analyst(51), 12:test engineer(19). In conclusion, a benchmark dataset with a total of 13 features excluding the target class label was extracted after which feature selection was applied and reduced the features to 5.

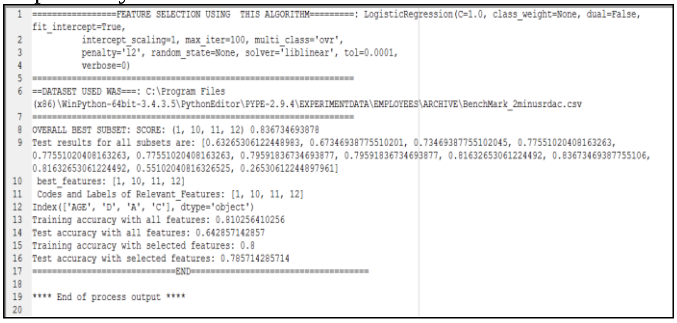


1	GENDER	AGE	LOLE	GSOLE	ROLE	BDGREE	UNIVERSITY	GSBDEGREE	RBACHELORS	R	D	A	C	CLASS
2	2	2	2	1	3	2	44	2	3	11	1.3	6.7	9	1
3	2	1	2	2	3	2	51	2	4	10.1	1	7.3	10.5	1
4	2	2	2	1	3	1	1621	2	4	7.5	1.2	5.9	10.5	1
5	2	2	2	2	3	2	8297	2	4	10.9	1.4	5.3	10.5	1
6	2	2	2	1	3	1	8350	2	3	1	1.1	5.6	9	1
7	2	2	2	2	4	1	9737	2	4	11.4	1.1	6.2	12	1
8	2	2	2	1	3	1	10185	2	4	12	1.3	8.8	10.5	1
9	2	2	2	1	3	1	10932	2	3	1.7	1	4.9	9	1
10	2	2	2	1	4	1	13344	2	4	11	1.5	6.7	12	1
11	2	2	2	1	3	2	350	2	3	1.9	1.9	1.1	4.5	2
12	2	2	2	7	4	1	893	2	3	3	2.3	1.5	5.3	2
13	2	2	2	0	4	2	3838	2	3	2.7	1.8	1.3	5.3	2
14	1	2	2	2	4	1	4973	2	4	0.1	1.5	1	6	2
15	2	2	2	1	3	1	5147	2	3	2.6	2.5	1.3	4.5	2
16	1	1	2	2	2	2	5508	2	4	2.7	2.1	0.8	4.5	2
17	2	2	2	16	2	2	6567	2	3	1.5	1.7	1	3.8	2
18	2	2	2	1	4	1	6907	2	3	3	2.8	1.5	5.3	2
19	2	2	2	7	3	2	7134	2	3	0.3	1.8	0.7	4.5	2

Figure 4a: SE Benchmark dataset

B. Feature selection

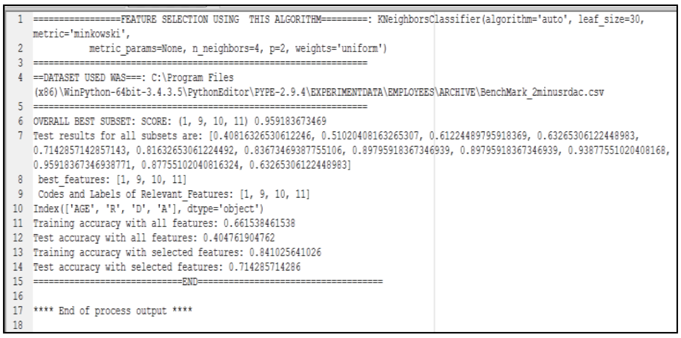
Initially, features were selected that other evidence, including more general models fitted into the full dataset, suggest would be important predictors of industry roles as applied by [7]. In this paper, three machine learning algorithms, namely logistic regression (LR), K-Nearest Neighbour (KNN) and Support Vector Machines (SVM) were used for this process. Through sequential backward selection method the three algorithms, (LR, KNN, and SVM(kernel='gamma', C=1.0, random_state=0)) were applied on the benchmark dataset (see Figure: 4a,b,&c) and resulted into a range of 4 feature subsets for each of the respective algorithms that gave an optimal performance accuracy (validation =0.80%, test=0.78%) , (validation =0.84%, test=0.71%) and (validation =0.90%, test=0.85%) respectively.



```
================FEATURE SELECTION USING  THIS ALGORITHM=========: LogisticRegression(C=1.0, class_weight=None, dual=False,
fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr',
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0)
=================================================================
==DATASET USED WAS===: C:\Program Files
(x86)\WinPython-64bit-3.4.3.5\PythonEditor\PYPE-2.9.4\EXPERIMENTDATA\EMPLOYEES\ARCHIVE\BenchMark_2minusrdac.csv
=================================================================
OVERALL BEST SUBSET: SCORE: (1, 10, 11, 12) 0.836734693878
Test results for all subsets are: [0.63265306122448983, 0.67346938775510201, 0.73469387755102045, 0.77551020408163263,
0.77551020408163263, 0.77551020408163263, 0.79591836734693877, 0.79591836734693877, 0.81632653061224492, 0.83673469387755106,
0.81632653061224492, 0.55102040816326525, 0.26530612244897961]
 best_features: [1, 10, 11, 12]
 Codes and Labels of Relevant_Features: [1, 10, 11, 12]
Index(['AGE', 'D', 'A', 'C'], dtype='object')
Training accuracy with all features: 0.810256410256
Test accuracy with all features: 0.642857142857
Training accuracy with selected features: 0.8
Test accuracy with selected features: 0.785714285714
=============================END=====================================

**** End of process output ****
```

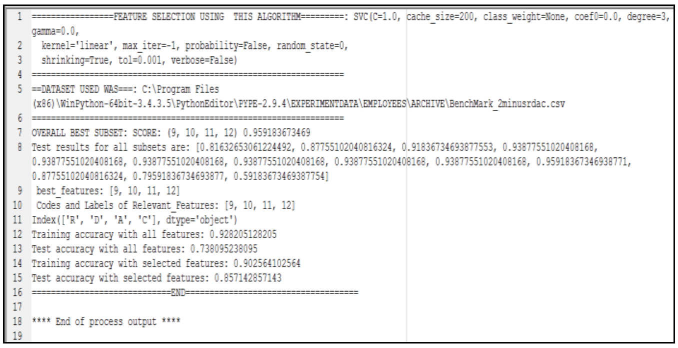
Figure 4a: Logistic Regression algorithm run results



```
================FEATURE SELECTION USING  THIS ALGORITHM=========: KNeighborsClassifier(algorithm='auto', leaf_size=30,
metric='minkowski',
           metric_params=None, n_neighbors=4, p=2, weights='uniform')
=================================================================
==DATASET USED WAS===: C:\Program Files
(x86)\WinPython-64bit-3.4.3.5\PythonEditor\PYPE-2.9.4\EXPERIMENTDATA\EMPLOYEES\ARCHIVE\BenchMark_2minusrdac.csv
=================================================================
OVERALL BEST SUBSET: SCORE: (1, 9, 10, 11) 0.959183673469
Test results for all subsets are: [0.40816326530612246, 0.51020408163265307, 0.61224489795918369, 0.63265306122448983,
0.7142857142857143, 0.81632653061224492, 0.83673469387755106, 0.89795918367346939, 0.89795918367346939, 0.93877551020408168,
0.95918367346938771, 0.87755102040816324, 0.63265306122448983]
 best_features: [1, 9, 10, 11]
 Codes and Labels of Relevant_Features: [1, 9, 10, 11]
Index(['AGE', 'R', 'D', 'A'], dtype='object')
Training accuracy with all features: 0.661538461538
Test accuracy with all features: 0.404761904762
Training accuracy with selected features: 0.841025641026
Test accuracy with selected features: 0.714285714286
=============================END=====================================

**** End of process output ****
```

Figure 4b: K-Nearest Neighbor algorithm run results



```
================FEATURE SELECTION USING  THIS ALGORITHM=========: SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0, degree=3,
gamma=0.0,
  kernel='linear', max_iter=-1, probability=False, random_state=0,
  shrinking=True, tol=0.001, verbose=False)
=================================================================
==DATASET USED WAS===: C:\Program Files
(x86)\WinPython-64bit-3.4.3.5\PythonEditor\PYPE-2.9.4\EXPERIMENTDATA\EMPLOYEES\ARCHIVE\BenchMark_2minusrdac.csv
=================================================================
OVERALL BEST SUBSET: SCORE: (9, 10, 11, 12) 0.959183673469
Test results for all subsets are: [0.81632653061224492, 0.87755102040816324, 0.91836734693877553, 0.93877551020408168,
0.93877551020408168, 0.93877551020408168, 0.93877551020408168, 0.93877551020408168, 0.93877551020408168, 0.95918367346938771,
0.87755102040816324, 0.79591836734693877, 0.59183673469387754]
 best_features: [9, 10, 11, 12]
 Codes and Labels of Relevant_Features: [9, 10, 11, 12]
Index(['R', 'D', 'A', 'C'], dtype='object')
Training accuracy with all features: 0.928205128205
Test accuracy with all features: 0.738095238095
Training accuracy with selected features: 0.902564102564
Test accuracy with selected features: 0.857142857143
=============================END=====================================

**** End of process output ****
```

Figure 4c: SVM algorithm run results

Thus, the best features that gave optimal results to each algorithm as evidenced by Figures 4a, b, & c, in increasing order of importance, were:

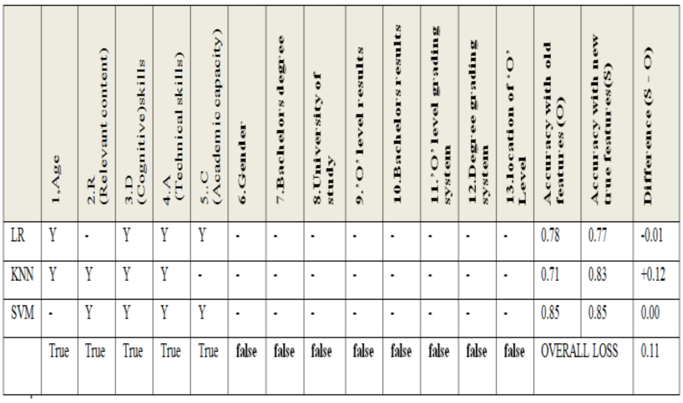
LR= {Age, D, A, C}; KNN = {Age, R, D, A}; SVC = { R, D, A, C}

Eventually, comparison was conducted and features that were popular in at least two algorithms were selected as true candidates for the best features while the rest were marked as false. Table 4a presents results of comparative analysis of features' subsets for the three algorithms where Y (yes) was used to mark a feature selected by an algorithm, otherwise a dash (-). A true/false score was used to analyse the features along the columns where a feature with at least two Ys was scored true otherwise false.

Those algorithms whose features had been scored false, hence marked for removal, were further analysed to study performance impact of removing each feature both in isolation and in combination. Caution was taken to ensure core features of the model were carefully removed and analysis was conducted on the impact of adding a feature in other algorithms where it was not selected, especially the core features marked for removal. Popular features that did not exist in other algorithms, were added unconditionally into the subsets of these algorithms. As a result, all three algorithms were affected through adding popular features, namely LR (feature 'R'), KNN (feature 'C') and SVM (feature 'age'). The overall impact in performance for removing or adding new features was determined.

For logistic regression (LR), the impact of adding 'R' was a loss in performance of -0.01 (0.78 to 0.77). For KNN, the impact of adding 'C' was a gain in performance of +0.12 (0.71-0.83). For SVC, the impact of adding 'age' was 0.00 (0.85-0.85). In conclusion, the addition of these popular features would result to a total gain in performance of +0.11 as shown in Table 4. As a result, a total of five features from the original 13 were selected as optimal features for further analyses, namely: age, R (relevant knowledge), D (cognitive skills), A (technical skills), C (capacity). Table 4a. shows cross analysis of features selected by the three algorithms.

Table 4a: Analysis of relevant features using SE benchmark dataset



	1.Age	2.R (Relevant content)	3.D (Cognitive)skills	4.A (Technical skills)	5..C (Academic capacity)	6.Gender	7.Bachelors degree	8.University of study	9.'O' level results	10.Bachelors results	11.'O' level grading system	12.Degree grading system	13.location of 'O' Level	Accuracy with old features (O)	Accuracy with new true features(S)	Difference (S – O)
LR	Y	-	Y	Y	Y	-	-	-	-	-	-	-	-	0.78	0.77	-0.01
KNN	Y	Y	Y	Y	-	-	-	-	-	-	-	-	-	0.71	0.83	+0.12
SVM	-	Y	Y	Y	Y	-	-	-	-	-	-	-	-	0.85	0.85	0.00
	True	True	True	True	True	false	false	false	false	false	false	false	false	OVERALL LOSS		0.11

Further experiments were conducted using the SE benchmark dataset where two induction algorithms for our machine learning model were fitted with all features, then with only the 5 selected features and the results were as shown in Table 4b. Further analysis was conducted to test whether model's performance difference was significant.

Table 4b: Model performance with all and only selected features in SE benchmark dataset

	Validation Test (naïve Bayes) %		*Validation Test (SVM) %*	
	All features (t_a)	*Selected features (t_s)*	*All features (t_a)*	*Selected features (t_s)*
Fold1	36.59	85.37	75.61	85.37
Fold2	51.43	74.29	74.29	88.57
Fold3	38.24	79.41	85.29	88.24
Fold4	40.63	75.00	87.50	87.5
Fold5	53.33	80.00	93.33	93.33
Mean	**44.04**	**78.81**	**83.20**	**88.60**

C. Testing whether the difference of group means (folds) was significant using ANOVA

Table 4b presents validation test results showing a trade-off between model's performance with all features and selected features both under naïve Bayes and SVM based constructs of the model. Two groups were defined, namely all features' and selected features' groups. The results reveal a possible difference between the two scenarios under both constructs of the model. The mean difference under naïve Bayes construct of the model was 34.77 (78.81-44.04) while under SVM was 5.4 (88.60-83.20). To be sure the difference was not due to any other factor but only difference in number of features, ANOVA test was conducted to rule out the effect of group (fold) to group (fold). For this type of test to be valid, conditions for ANOVA that must be satisfied, homogeneity of group variance and normality of data, were checked.

Table 4c presents results for ANOVA analysis for both kinds of model constructs investigated through 10 trials of 5-fold cross-validation experiments. The results indicate the feature sets variances were equal for naïve Bayes based model while not equal for SVM based model and, in fact, means of the two feature sets scores were different in either case and, therefore, the seemingly difference between the two models in Table 4b was real, was due to effect of variation of feature set. For SVM based model Welch and Brown-Forsythe values are 0.000 for both.

Table 4c: ANOVA results (effect of feature selection) in SE benchmark dataset

Type of validity	**Type of test**	**Model**	**p-value**	**Decision**
1.Homogeneity of group variances Accept if p>0.1	**(Levene test -** Equality of variances) **Hypothesis:** Are variances between the groups equal?	**naiveBayes**	0.250	*ACCEPT*
		SVM	0.021	*REJECT*
2.Difference of group means Accept if p>0.05	**(F test -** Equality of group means) **Hypothesis:** Are group means equal?	**naiveBayes**	0.000	*REJECT*
		SVM	0.000	*REJECT*

Table 4c reveals that reduction of features improved the performance of our model. The change in performance was significant. Slightly better performance could be achieved with fewer features, hence reducing the computational demand in terms of time and computational power. For this dataset, out of 13 features only 5 features produced optimal results, namely Age, R, D, A, C.

VI. CONCLUSION

This paper has presented theoretical development of concepts to tackle the problem of mapping graduates' skills to industry role as well as the methodology for validating these concepts. For purpose of clarity, the results have been presented using tables. The statistical analysis procedures have been carefully selected based on preliminary tests results for data validity.

In summary, the results findings in this paper have literally provided answers to the research question posed in this study: Are there valid concepts that can be used to develop a valid model for predictive mapping of graduates' skills to industry roles using machine learning techniques? Based on the findings in this study, it is important to note when developing classifier models for mapping skills to industry roles that appropriate attributes that are valid for machine learning are content knowledge, cognitive skills, technical skills, academic capacity, and age.

REFERENCES

[1] Aggarwal, V., Srikant, S., & Nisar, H. (2015). A dataset comprising AMCAT test scores, biodata details, and employment outcomes of job seekers.

[2] Baker, E.L & Mayer, R.E.(1999). Computer-Based Assessment of Problem Solving

[3] Bharthvajan, R. (2013). Competency Mapping. International Journal of Innovative Research in Science,Engineering and Technology Vol. 2, Issue 11, November 2013

[4] Bloom, B.S. (Ed.), Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain. New York: David McKay.

[5] Chang, H.(2009).Employee Turnover: A Novel Prediction Solution with Effective Feature Selection Wseas Transactions on Information Science and Applications Issue 3, Volume 6, March 2009

[6] Chien C., Chen L.(2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. Expert Systems with Applications 34 (2008) 280–290 .Retrieved February 17, 2016 from: http://www.sciencedirect.com

[7] Clare A. and King R. D. (2003).Predicting gene function in Saccharomyces cerevisiae". Bioinformatics Vol. 19 Suppl. 2, pp. ii42–ii49, 2003

[8] Dodig-Crnkovic, G. (2002). Scientific Methods in Computer Science.

[9] Proc. Conf. for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, (2002)

[10] Kaminchia, S.(2014). Unemployment in Kenya: Some economic factors affecting wage

[11] Employment. African Review of Economics and Finance Vol. 6, No. 1, June 2014

[12] Kellaghan, T. & Greaney, V. (2003). Monitoring Performance: Assessment and Examinations in Africa. Association for the Development of Education in Africa ADEA Biennial Meeting 2003 (Grand Baie, Mauritius, December 3-6, 2003)

[13] Kitchenham, B., Pickard, L. & Pfleeger, S.L. (1995). Case Studies for Method and Tool Evaluation

[14] Korte W., Husing T., Hendriks, L. & Dirkx, J. (2013). Towards a European Quality Label for ICT Industry Training and Certification. Final Report. 2013.

[15] Kraiger, K., Ford, K. & Salas, E. (1993).Application of Cognitive, Skill-Based, and Affective Theories of Learning Outcomes to New Methods of Training Evaluation. Journal of Applied Psychology Monograph, 1993, Vol. 78,Pg 311-328.

[16] ILO. (2015). World Employment and Social Outlook 2015: The Changing Nature of Jobs 2015 Report

[17] Jantawan, B. & Tsai, C.(2013).Application of Data Mining to Build Classification Model for Predicting Graduate Employment. International Journal of Computer Science and Information Security, Vol. 11, No.4, October, 2013.

[18] Mwakondo, F.M., Muchemi, L., & Omwenga, E.I.(2016). Proposed Model for Predictive Mapping of Graduate's Skills to Industry Roles Using Machine Learning Techniques. The International Journal Of Engineering And Science (IJES) Vol.5, Issue 4, PP -15-24, 2016

[19] NOC. (2011). Human Resource & Skills Development Canada. National Occupational Classification 2011 Report

[20] Onwuegbuzie, A.J, Leech, N.L, & Collins, K.M.T. (2012). Qualitative Analysis Techniques for the Review of the Literature . Qualitative Report 2012, Vol 17.

[21] Quintin G. (2011). Right for the job: Overqualified or Underqualified.

[22] Shashidhar, V., Srikant, S., Aggarwal, V.(2015). Learning Models for Personalized Educational Feedback and Job Selection. Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015

[23] Srikant, S. & Aggarwal, V.(2014).A System to Grade Computer Programming Skills using Machine Learning

[24] Winterton, J., Delamere, F. & Stringfellow, E. (2005). Typology of Knowledge, Skills and Competences: Clarification of the concept and prototype.

[25] Wohlin, C. & Regnel, B.(1999).Achieving Industrial Relevance in Software Engineering Education, Proceedings Conference on Software Engineering Education & Training, pp. 16-25, New Orleans, Lousiana, USA, 1999.

[26] Winfrey, E.C. (1999). Kirkpatrick's Four Levels of Evaluation. In B. Hoffman (Ed.),Encyclopedia of Educational Technology. Retrieved March 24, 2015, from http://coe.sdsu.edu/eet/Articles/k4levels/start.htm

BIOGRAPHY

The author of this paper, Dr. Fullgence M. Mwakondo, is working at the Technical University of Mombasa (TUM) as a lecturer in the Institute of Computing and Informatics. He has over 10 years' experience in teaching at higher institutions of learning. He completed his BSc. Mathematics and Computer science at Jomo Kenyatta University of Agriculture and Technology, and MSc. (Information Technology) at Masinde Muliro University of Science and Technology. He is a PhD holder in Computer science from the School of Computing & Informatics at the University of Nairobi.