Employability and Related Context Prediction Framework for University Graduands: A Machine Learning Approach

Manushi P. Wijayapala, Lalith Premaratne, Imali T. Jayamanne

***Abstract*— In Sri Lanka (SL), graduands' employability remains a national issue due to the increasing number of graduates produced by higher education institutions each year. Thus, predicting the employability of university graduands can mitigate this issue since graduands can identify what qualifications or skills they need to strengthen up in order to find a job of their desired field with a good salary, before they complete the degree.**

The main objective of the study is to discover the plausibility of applying machine learning approach efficiently and effectively towards predicting the employability and related context of university graduands in Sri Lanka by proposing an architectural framework which consists of four modules; employment status prediction, job salary prediction, job field prediction and job relevance prediction of graduands while also comparing performance of classification algorithms under each prediction module. Series of machine learning algorithms such as C4.5, Naïve Bayes and AODE have been experimented on the Graduand Employment Census - 2014 data. A pre-processing step is proposed to overcome challenges embedded in graduand employability data and a feature selection process is proposed in order to reduce computational complexity. Additionally, parameter tuning is also done to get the most optimized parameters. More importantly, this study utilizes several types of Sampling (Oversampling, Undersampling) and Ensemble (Bagging, Boosting, RF) techniques as well as a newly proposed hybrid approach to overcome the limitations caused by the class imbalance phenomena. For the validation purposes, a wide range of evaluation measures was used to analyze the effectiveness of applying classification algorithms and class imbalance mitigation techniques on the dataset. The experimented results indicated that RandomForest has recorded the highest classification performance for 3 modules, achieving the selected best predictive models under hybrid approach having an area under the ROC curve interpretation as an 'Excellent' experiment, while a C4.5 Decision Tree model under Ensemble approach has been selected as the best model of the remaining module (Salary Prediction module).

***Keywords*— Machine Learning, Employability Prediction, Data Mining, Supervised Learning**

Manuscript received on 18 April 2016. Recommended by Prof. G.N. Wikramanayake on 19 Dec 2016.

Manushi Wijayapala holds a B.Sc. (Hons) First class in Statistics with Computer Science from the University of Colombo, Sri Lanka and Bachelor of Information Technology (First class) degree from the University of Colombo School of Computing, Sri Lanka. (email:manushimn@gmail.com).

H.L.Premaratne is a Senior Lecturer at the University of Colombo School of Computing. (email: hlp@ucsc.cmb.ac.lk)

I.T. Jayamanne is a Lecturer at the Department of Statistics, University of Colombo. (email: imali@stat.cmb.ac.lk)

I. Introduction

One of the main objectives of higher education is to prepare students to pursue different careers in a country. With many economies being reported as not producing adequate employment opportunities to absorb the growth in the working age population, a generation of productive young workers will have to face an uncertain future unless something is done to reverse this trend. Thus increasing the graduands' chances of obtaining decent jobs that match their education and training, equipping students in universities with the necessary competencies to enter the labour market, enhancing their capacities to meet specific workplace demands, improving the students' skills and qualifications to meet the employers' expectations are some of the essential tasks that need to be carried out in order to improve the employability of Sri Lanka [1].

In Sri Lanka, 'employability' has been a major topic among many parties in recent years. Especially, unemployable graduates and graduands are becoming a crucial issue that recent governments are facing. Conflicts between the parties involved in these matters are often experienced. Once it was difficult to find details about graduand unemployability, the census done by HETC with the guidance of the Ministry of Higher Education, proves to be a gold-mine and provided valuable insights into the main factors having a significant bearing on the employability of graduands. A systematic and scientific analysis using these data will result in a great solution for the issue of unemployment of graduands.

Machine learning (ML) has been recognized as a type of artificial intelligence which focuses on computer program development that can teach themselves to nurture without being explicitly programmed and change when exposed to new data [2]. In other words, the goal of ML is to invent or use the learning algorithms which will learn automatically without the human assistance or intervention.

The main aim of this research is to discover the plausibility of applying machine learning approach efficiently and effectively towards predicting the employability and related context of university graduands in Sri Lanka. Hence, objectives of this research can be devised as follows;

1. Propose a framework using ML based architecture to,
   a. Predict the employment status (Employed, Unemployed, Underemployed) of a university graduand at the time of official graduation
   b. Predict the salary range of an employable graduand (Very low, Low, Average, High)

c. Predict the job relevance with the degree (Relevant field, Irrelevant field) of an employable graduand
d. Predict the type of job field of an employable university graduand (Medicine field, Engineering field, Commerce field, Lecturing, Administrating field, Agriculture-Export field, and Support Staff field)

while also coping with the constraints conflated in graduand employability data.

2. Identify the most important factors for the employment status of a university graduand, for the salary range of an employable graduand and for the type of job field of an employable graduand.
3. Compare and identify the most efficient and accurate classification algorithm/s to predict the employability of university graduands under each prediction module.

No successful prior research work has been considered fulfilling all the aforementioned research objectives. Even though researches have been carried out related to 'graduate' employability prediction [3, 4, 5], no research work has been found in literature related to predicting employability of university 'graduands'. Additionally, any kind of graduate or graduand employability prediction research has not been carried in the Sri Lankan context. Furthermore, none of the previous research work related to employability prediction of graduates have attempted to predict the job salary, job field and job relevance, which will be covered in this research. Moreover, we believe that this is the first occasion in employability prediction researches that considered the class imbalance problem and attempted this amount of class imbalance problem mitigation techniques to give more accurate results. Additionally, findings of this research can be used to reduce the overall unemployability of Sri Lanka. Even though this research focuses on graduand employability prediction, this can further extend with other sectors of the society to get a clear picture about unemployability.

This paper is organized as follows: Section II discusses related work; Section III gives an overview of the methodology; Section IV describes the experimental setup. Moreover Section V presents the results of the study and finally Section VI and Section VII present in-depth discussion together with conclusions and future work.

II. Related Work

The section exposes some of the previous research studies and results related to our research objectives.

Jantawan and Tsai [3] proposed a method to predict whether a graduate in Maejo University in Thailand will be employed, unemployed or undetermined. Thus authors try to build a graduate employment model using several classification algorithms in data mining and compare those algorithms to find the best algorithm to predict the employability in this university. The algorithms used by the authors for this comparison were Bayesian methods (AODE, WAODE, NaviveBayes, BayesNet and HNB) and tree methods (C4.5, BFTree, REPTree, NBTree, and ID3). Results showed that the WAODE algorithm, a type of Bayes algorithm has achieved the highest accuracy of 99.77%. However their framework is questionable since they have used two variables, 'work status of the graduate' and 'position of graduate' as the explanatory variables when they actually try to predict the employability status of the graduate. Because of using explanatory variables which are almost similar to the target variable, authors have gained an almost impossible accuracy of 99.8%.

Sapaat, Mustapha, Ahmad and Chamili [5], in their work focused on identifying features that influenced graduates' employability of Malaysian universities and tries to predict whether a graduate has been employed, remains unemployed or in an undecided situation in the first six months after the graduation based on actual data from the graduates. To accomplish it, authors have used data from the Tracer Study for the year 2009. The prediction has been executed through a series of classification algorithms of Bayes and decision tree methods. Results have shown that C4.5 decision tree algorithm gave the highest accuracy leading to the conclusion that a decision tree based classifier is more appropriate for the tracer data [5].

The study [4] presented a graduate employability model which uses different types of Bayesian methods to hunt the most important factors of graduate employability in Khon Kaen University, Thailand. In addition to that, authors try to compare the accuracy of all selected Bayesian algorithms. The researchers have used hold-out validation method to evaluate the models. The results have shown that the AODE and WAODE algorithms have gained the highest accuracy [4]. Furthermore, the experiment has shown that work province, the times which found the work and occupation type have a direct effect on employability.

All of the researches related to employability prediction [3, 4, 5] in literature, attempt to predict the employability status of university 'graduates'. However we could not find any research work related to predicting employability of university 'graduands'. A university graduate is a person who has already graduated whereas a university graduand is a person about to graduate after completing the degree. Furthermore, one of the limitations in their research work is they have only tried to predict the employability status of the graduate, i.e. whether a graduate is employed, unemployed or undetermined. These studies would have been more interesting if they had included modules which predict the salary range and the job field as well.

III. Methodology

In the previous section, we reviewed some existing related work and identified potential limitations in those approaches. This section outlines the proposed methodology to address the research questions by extensively describing the design aspects and the foundation of this study.

In this research, it is aimed to apply a machine learning (ML) based approach to build a framework of predictive models, which can correctly classify a university graduand into three classes first according to the employability status and then classify each employable graduand according to the job field, job salary and job relevance. The proposed framework is depicted in Figure 1. Model-1, Model-2, Model-3 and Model-4 which are depicted in this figure are the chosen best four models (best models are chosen according to the various evaluation measures that will be discussed in later sections) for four respective modules, which will be selected at the end of this study after mitigating all the conflated issues. It should be noted that second, third and fourth models will be applied only if the first model gives the output as ‘employed’.

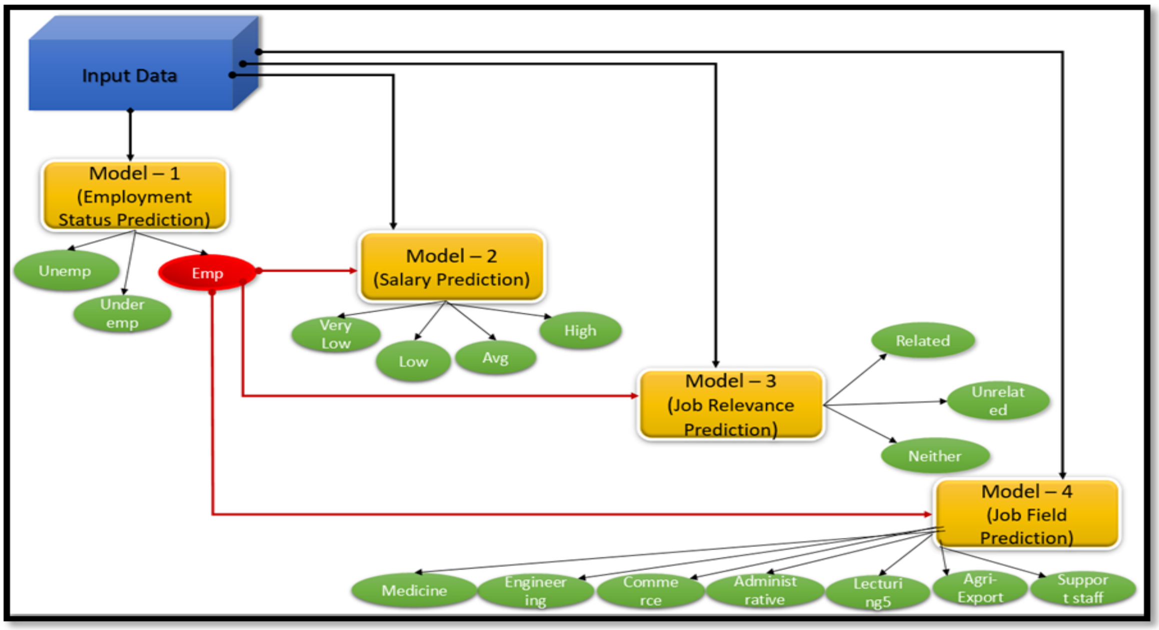

Fig. 1 Employability prediction framework proposed by the study

Reflecting the notion, the proposed comprehensive methodology for employability prediction of university graduands under four modules, explore decision tree algorithms, Bayesian algorithms and combination of multiple classifiers (ensemble methods) which are supervised learning models. Supervised learning would be the ideal solution to obtain a better classification, since there is a reasonable amount of annotated data already. In order to achieve a stable model, each classification algorithm was tried with a range of model parameters and compared them based on different evaluation metrics. Based on the constraints identified in university graduands’ employment data and classification models (e.g.: class imbalance problem), several mitigation mechanisms were presented to overcome these limitations.

Figure 2 presents an architectural view of the proposed methodology. The proposed methodology basically consists of four main modules namely employment status prediction, salary prediction, job relevance prediction and job field prediction while the methodology of each of these four modules consist of four main phases namely pre-processing, feature selection, applying classifiers/ training models and selecting the best classifier as depicted in Figure 2. It should be noted that module 2, module 3 and module 4 depends on the output of module 1.

Original data which was taken from Graduand Employment Census – 2014, went through a series of pre-processing steps. This pre-processing stage consists of data cleaning (handling missing data, correcting inconsistent data and classifying detailed data into categories) and data transformation (generalization and attribute construction). Most probably the original data set can have certain incompatibilities that will affect the performance of the

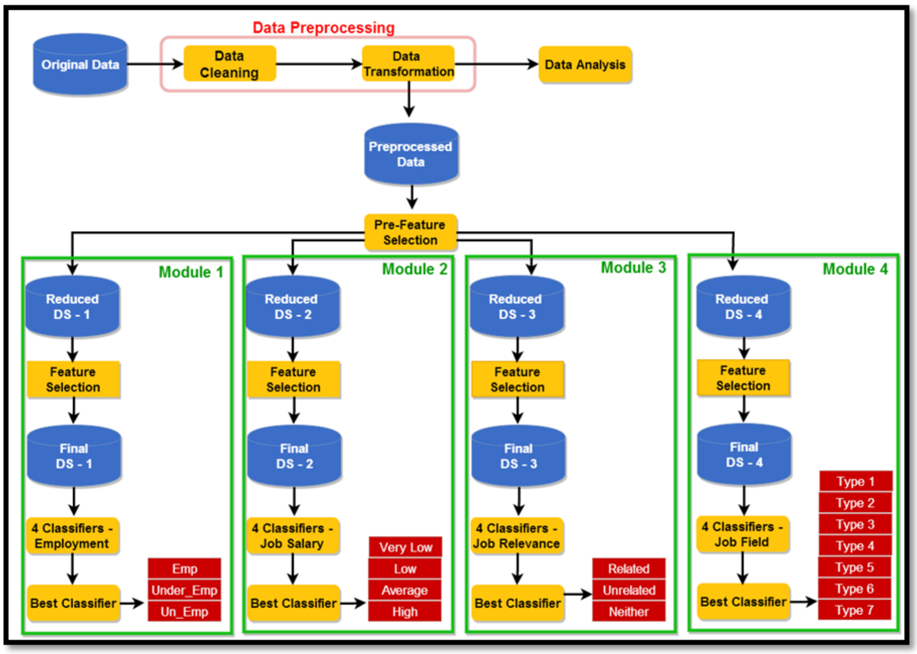

Fig. 2 Overall System Architecture

final model. Therefore, a careful pre-processing would be vital to achieve better results in the next phase

After pre-processing data, initial feature selection will be carried out using the domain knowledge, expert opinions and using previous literature. The reason for doing this initial feature selection was, the original data set contained additional variables which are not related to employability prediction at all, since employability prediction was not the only objective of the graduand employment census. For four different modules four sets of features will be selected from the original set of features. After doing the initial (manual) feature selection, automated feature selection will be carried out under each four modules separately as described in Section A.

After finalizing the initial steps, the next stage will be applying four different classifiers under each module to the final four data sets. Parameters of these classifier algorithms will be fine-tuned to suit the training models.

In order to overcome the limitations caused by a possible class imbalance phenomena, several mitigation techniques will be applied such as sampling, ensemble and hybrid approaches on these training models. Thereby the suitability of these approaches was measured on experimented algorithms and data. Employed different models were compared based on different evaluation metrics (accuracy, precision, recall, specificity, F-ratio, G-means and ROC AUC; area under the curve of receiver operating characteristic curve) to achieve a stable model. After the training, the model should be capable of predicting the employment status, job salary, job relevance and job field of the university graduands.

A. Feature Selection

Feature selection which is also known as subset selection, is the process of choosing a small subset of features that is sufficient enough to predict the targets easily and more accurately. The biggest of pros of applying feature selection techniques are being able to avoid over-fitting and being able to reduce the computational cost[6].

Even though decision trees have the ability to select features on their own, experiments with C4.5 decision trees have shown that adding a random binary attribute to standard datasets impacts classification performance, causing it to weaken (generally by 5% to 10% in the situations tested) [7]. The reason for this is, at some point in the trees that are learned, the unrelated attribute is invariably chosen to branch on, producing random errors when test data is processed. Even though one may think that how can this happen when decision trees are cleverly designed to pick the best attribute for splitting at each node, the reason is as proceed further down the tree, very less data is available to support making the selection decision [7]. Thus the feature selection step will be done before training the C4.5 classifier. However, in the case of RandomForest (RF), it is different since at each step, voting mechanism is used after randomly choosing the splitting attributes. Thus feature selection is not needed for the RF since it automatically chooses the best features.

Due to the negative effect of irrelevant attributes on most of the machine learning algorithms, it is common to do the learning after a feature selection step that attempts to eradicate all but the most relevant attributes. The finest way to select relevant attributes is by manually, based on the understanding of the learning problem and what the features actually mean [7]. This is the reason that initial (pre) feature selection was done as described in previous sections. However, automatic methods also can be useful. Reducing the dimensionality of the dataset by removing unsuitable attributes improves the classification performance of learning algorithms. Furthermore, it also speeds them up, although this may be compensated by the computation involved in feature selection. More importantly, reducing the number of features yields a more easily interpretable representation of the target concept, focusing only on the most relevant attributes.

Fundamentally there are two types of feature selection algorithms; Filter and Wrapper methods. Due to the expensive computational time taken when wrapper methods are applied, only filter methods will be used in this study for the feature selection. Two popular filter feature selection methods (Chi squared attribute evaluation and Gain ratio attribute evaluation) were used with the Ranker search method which ranks the attributes by their importance. After applying these feature selection methods on the training data, the results of these two methods will be analyzed and the common features which have given lowest ranks in both two methods will be removed.

B. Classification Algorithms explored

1) *C4.5 Algorithm:* Among decision tree algorithms, C4.5 is the most commonly used algorithm. C4.5 was originally proposed as a successor of ID3 algorithm [8]. It is also capable of handling pruning, missing values and numeric values. At each node of the tree, C4.5 chooses the attributes of the data that most effectively splits the training dataset into subsets of one class or the other. The splitting criteria are based on the normalized information gain.

2) *Naïve Bayesian (NB) Algorithm:* Naïve Bayes is simple probabilistic classifier, based on applying Bayes Theorem which provides a way to calculate posterior probability P(Ci|X) using prior probability of class, prior probability of data and likelihood of the data given the class [2]. Naïve Bayes uses a strong assumption of conditional independence where it assumes attributes are independent from each other. Naïve Bayes can be trained very efficiently[9].

3) *AODE Algorithm:* Averaged one-dependence estimator (AODE) is a probabilistic classification learning technique. AODE was developed to solve the attribute independence problem of the famous Naive Bayes algorithm. Even though it usually develops considerably more accurate and reliable classifiers than Naive Bayes, its computational complexity is relatively high. Similar to NB, AODE also does not use tuneable parameters and does not perform model selection. As a result of that, AODE also has low variance. It supports incremental learning where the classifier can be updated efficiently with information from new examples as they become available [7]. Instead of predicting the single class, it predicts class probabilities, allowing the user to recognize the confidence which each classification can be made. Moreover, AODE can directly handle some situations where some data are missing.

C. Ensemble Algorithms explored

The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm for improving generalizability and robustness of a single estimator [2]. Usually ensemble methods work well with unstable base classifiers like decision trees and on stable classifiers like Bayesian classifiers, these methods do not work very well [7].

1) *Bagging:* Bagging produces multiple versions of the same predictor and combines the numerical prediction of these versions using plurality vote to identify the prediction class [9]. Multiple versions of the same base algorithm will be applied on each replicated bootstrap, where each bootstrap is created based on random sampling with replacement technique.

2) *Boosting:* Boosting change the weights for incorrectly classified data with the previous model used. Boosting involves incrementally building an ensemble by using all data to train each learner, but instances that were misclassified by the previous learners are given more weight so that subsequent learners give more focus to them during training [10]. In this study, AdaBoost.M1 algorithm will be used for Boosting.

3) *Random Forests:* Random Forests operate by building a multitude of decision trees at training time and outputting the class that is the mode of the classes (for classification prediction) or mean of the classes (for regression prediction) of the individual trees. Random Forests correct the overfitting problem of decision trees. The algorithm for inducing a Random Forest was developed by Breiman and Cutler and "RandomForests" is their trademark [11]. When splitting a node while constructing the tree, the split which is chosen is no longer the best split amid all features [11]. Instead, the best split among a random subset of the features is picked as the split. Because of the randomness, the bias of the RF usually increases a little relative to the bias of a single non-random tree [11]. But because of averaging, its variance also drops, typically more than compensating for the increase in bias, thus, yielding an overall enhanced model [2].

D. Overcoming class imbalance problem

Chawla states that “A dataset is imbalanced if the classification categories are not approximately equally represented” [12]. The problem with imbalanced data is that in classification problems with such data, the minority class instances are more likely to be misclassified than the majority class instances, due to the design principles of most machine learning algorithms.

In a literature review [13] done by Longadge, Dongre, and Malik, they have stated that according to the existing literature all the methods which can be used to rectify data imbalance problem can be categorized in to three main approaches; data pre-processing approach, the algorithmic approach and feature selection approach. In data pre-processing technique, sampling is applied on data in which either new samples are added or existing samples are removed. Algorithmic approach includes the cost-sensitive method, recognition-based approaches, and ensemble based approaches. The aim of cost sensitive classification is to minimize the cost of misclassification that can be realized by choosing the class with the minimum conditional risk. Since the cost of each class was not known at the learning time, cost sensitive approach will not be used to mitigate imbalance problem in this study. Feature selection methods are also vital since the data imbalance problem is commonly accompanied by the problem of high dimensionality of the data set.

To mitigate the data imbalance problem, this study proposes 4 approaches as mentioned below.

1) *Undersampling:* This approach attempts to balance the distribution of class by randomly removing a majority class sample. The biggest drawback with this method is loss of valuable information.

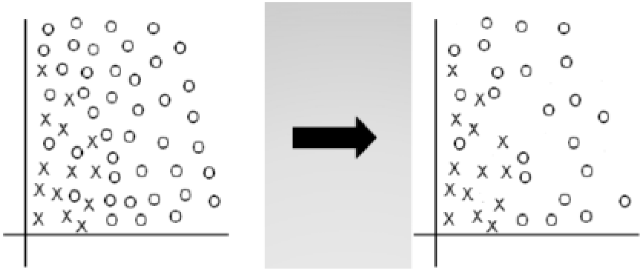

Fig 3 Overview of undersampling technique [13]

2) Oversampling: This approach attempts to balance the distribution of class by replicating minority class instances. The main drawback with this method is it can overfit the data.

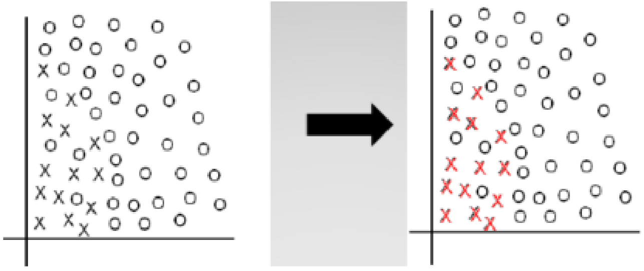

Fig 4 Overview of oversampling technique [13]

3) *Ensemble based approach:* Under this approach previously mentioned 3 ensemble algorithms will be used; Bagging, Boosting and RandomForest

4) *Hybrid approach:* This study proposes this approach which combines the aforementioned two methods; sampling technique and ensemble based technique. In this proposed hybrid approach, first, sampling technique will be applied to the training dataset and then generated new training dataset will be fed to the base algorithms along with the previously mentioned ensemble methods.

IV. Experimental Setup

A. Dataset

This study was carried out using a secondary dataset which included 15,726 records of detailed information about the state university graduands who were graduated in year 2014. This data was gathered for the 'Graduand Employment Census - 2014', under the HETC project with the guidance of the Ministry of Higher Education.

Originally, data set contained 82 variables. Table I shows chosen variables in this research after initial-feature selection.

TABLE I: INITIALLY SELECTED VARIABLE SET

No	Variable name	Variable Description
1	University	Name of the university
2	Gender	Gender of the graduand
3	Ethnicity	Ethnicity of the graduand
4	Faculty	Faculty the graduand studied in
5	Degree Type	Type of the degree
6	Stream	Stream of the degree specialized in
7	Medium	Medium of the studying degree
8	Class	Received class from the degree
9	English Proficiency (Written)	Written English skills of the graduand
10	English Proficiency (Oral)	Oral English skills of the graduand
11	O/L English results	O/L results for English subject
13	Browsing web	Whether graduand has ever browse web
15	Using office-packages	Whether graduand can use the office packages well
16	Writing Programs	Whether graduand can wrote computer programs
17	Extra activities	Whether graduand has done extra-curricular activities in university life
18	Extra activities : Detail	If response is yes, for the above variable, extra activities he has done
19	Vocational activities	Whether graduand has done vocational activities in university life
20	Vocational activities : Detail	If response is yes for the above variable, vocational activities he has done
21	Other education	Whether graduand has additional educational qualifications
22	Other education : Detail	If response is yes for the above variable, such qualifications he has
23	Lived Area	Type of area which the graduand has lived in most of his/her life
24	District	District which the graduand has lived in most of his/her life
25	GCE A/L	The type of school graduand have attended for GCE A/L
26	Parents' education	Highest level of education achieved by either graduands' father or mother
27	Expected salary	Expected salary of the graduand
28	Expected sector	Expected job sector of the graduands
29	Employment status	The employment status of the graduand
30	Employed sector	Employed job sector
31	Position hold	Position hold in the job
32	Economic sector	Economic sector of the job
33	Actual salary	Actual job salary
34	Job relevance	Whether job is related with the degree of the graduand

Three out of four target variables are highlighted in Table I in blue colour while the other target variable is a combination of two variables which are highlighted in green colour. Furthermore, variables which are numbered after 29 are only relevant for the graduands who have already employed at the time when this census was carried out. Moreover, the attributes which are coloured in yellow were concatenated into one single variable named computer literacy since all these three variables are binary.

Table II shows the final feature sets after applying all the pre-processing steps and feature selection algorithms for all four modules. However as mentioned in previous sections, for RF algorithm, the full datasets (module 1 – 22 features, other modules - 23 features each) will be used since RF itself can choose the relevant attributes very well.

TABLE II: SUMMARY OF FEATURES SELECTED IN ALL 4 MODULES

Rank	Module 1 : Employment Status Prediction	Module 2 : Job Salary Prediction	Module 3: Job Field Prediction	Module 4 : Job Relevance Prediction
1	Medium	Discipline	Discipline	Medium
2	Discipline	Faculty	Faculty	Discipline
3	Gender	Medium	Degree Type	Faculty
4	Faculty	Gender	Employed Sector	Stream
5	Stream	Expected Salary	Stream	Employed Sector
6	Degree Type	Stream	Medium	Degree Type
7	Lived Area	Degree Type	Vocational Training	Expected Salary
8	GCE O/L English	Employed Sector	Preferred Sector	English Proficiency(Oral)
9	Expected Salary	University	University	Lived Area
10	University	Lived Area	Computer Literacy	GCE O/L English
11	Preferred Sector	Computer Literacy	Professional Education	English Proficiency(Written)
12	Vocational Training	Vocational Training	Gender	University
13	English Proficiency (Oral)	Preferred Sector	Lived Area	School - GCE A/L
14	School - GCE A/L	GCE O/L English	Expected Salary	Parents Education
15	Parents Education		Class	Class
16	District			Computer Literacy
17				District

B. *Evaluation Procedure*

In this study, Nested Stratified K-Fold Cross Validation is used to evaluate the algorithms explored. In this approach, both the parameter optimization and the evaluation of the algorithm are done together. In order to understand this complex evaluation technique more clearly, first K-Fold Cross Validation is explained and then 'stratified' version of cross validation will be explained. 'Nested' stratified k-fold cross validation will be explained afterwards.

In the usual k-fold Cross Validation procedure, the data set is randomly split into k mutually exclusive subsets (the folds) of approximately equal size [7]. Then the model is learned using (k-1) folds, and the fold left out is used for testing. This process is then repeated k times, with each subsample is used exactly once as the validation data.

Finally, the average value computed in the loop will be the performance measure reported by k-fold Cross Validation.

Stratified k-Fold is a variation of k-fold which returns stratified folds: each set contains approximately the same percentage of samples of each target class as the complete set. This approach can be computationally bit expensive, but without wasting much data, it is known to produce reliable results [7].

Nested variation of Stratified k-Fold Cross Validation comes into play when one needs to tune the parameters. As the name suggest, in nested variation, there is an outer cross validation as well as an inner cross validation as shown in Figure 5. Nested k-fold cross validation encapsulates one layer of cross-validation inside another one. The inner layer is used to try out different parameters and pick the ones that work best for the given distribution while the outer layer is used to evaluate the best parameters found in the inner layer [14].

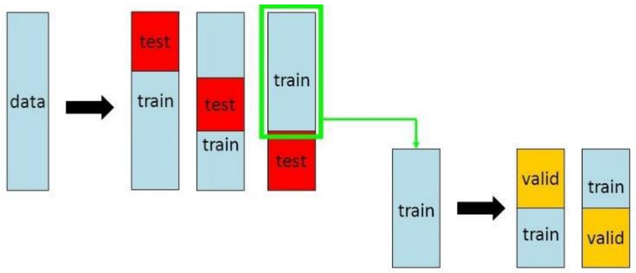

Fig 5 Inner and outer cross validations [14]

In nested variation, there is an outer cross validation as well as an inner cross validation. Nested k-fold cross validation encapsulates one layer of cross-validation inside another one. The inner layer is used to try out different parameters and pick the ones that work best for the given distribution while the outer layer is used to evaluate the best parameters found in the inner layer [14]. In short, the inner cross-validation is only used to find the "optimal" parameter settings by finding those settings that maximize estimated predictive performance in the inner cross-validation. Once those settings have been found, the model is rebuilt with those settings from the full training set (i.e. the particular training set of the outer cross-validation) and that single model is used for prediction.

This Nested Stratified k-fold Cross Validation method is quite powerful for detecting over fitting and estimating the generalization error conservatively [14]. Hence in this study, Nested Stratified 10-fold Cross Validation was used. The reason for choosing k=10 is that it is empirically proven that for majority of the datasets, 10-fold schema gives better training model with lesser possibility of having over-fitting scenarios [7].

C. Setting Up The Parameters

Prior to process of applying base classification algorithms, it is essential to find the optimal parameters of those algorithms that suits the training datasets. Here, the evaluation procedure (nested stratified cross validation) described in above section will be used to evaluate the parameter configurations. Since there are four modules, each consist of training with all aforementioned classification algorithms, the general procedure of setting up those parameters for a single module will be discussed, since the method of tuning parameters for each module is similar.

1) *Tuning C4.5 Decision Tree Algorithm:* For C4.5 algorithm two parameters out of the six parameters given in Table III were considered to tune, while keeping the values of other four parameters to the recommended values. The parameters tuned are confidence factor (C) and minimum number of instances per leaf (M). Root Mean Squared Error (RMSE) was used to tune the parameters in C4.5 (J48) algorithm. Smaller the error, better the classifier will be. We choose the optimal parameter pair from the all possible parameter pairs where the value of C goes from 0.05 to 1 by 20 steps and M value goes from 2 to 3 by 2 steps. i.e we chose the optimal values from 40 (C, M) pairs; (0.05, 2), (0.10, 2), … , (1, 2), (0.05, 3), (0.10, 3), … , (1, 3) by running nested cross validation.

TABLE III: Parameter configuration for C4.5 algorithm

Parameter description	**Recommended default value**
Confidence factor used for pruning (C)	0.25
Minimum number of instances per leaf (M)	2
Whether reduced-error pruning is used instead of C.4.5 pruning.	False
Whether to consider the subtree raising operation when pruning.	True
Whether pruning is performed.	False
Whether counts at leaves are smoothed based on Laplace.	False

2) *Tuning RandomForest Algorithm:* For this algorithm two parameters out of the three parameters given in Table IV were considered to tune, while keeping the values of other four parameters to the recommended values. The parameters tuned are the number of features to be used and the number of trees to be generated.

TABLE IV:Parameter configuration for RandomForest

Parameter description	**Default Value**
The maximum depth of the trees	Unlimited
The number of attributes to be used in random selection	$\log_2$(numOfattributes)+1
The number of trees to be generated.	100

Here in number of trees parameter, the larger the better, but also the longer it will take to compute. In addition, note that results will stop getting significantly better beyond a critical number of trees. Thus tuning this parameter will make the computation time less without affecting the accuracy. Therefor we choose the values from 80 to 200 by 13 steps as the number of trees parameter values to be optimized (i.e, 80, 90, 100, …, 190, 200). The number of features value is the size of the random subsets of features to consider when splitting a node. The lower, the greater the reduction of variance, but also the greater the increase in bias. For classification tasks, empirical good value for this parameter is the square root of number of features used [11]. However Breiman, the co-developer of RandomForest algorithm, has said that it is better to try up the half and the twice of the square root of number of features also, to find the most optimal parameter [11]. Therefor we choose the values 5 (since there are 22 features for the first module and $\surd 22=4.7$), 2 and 10 as the number of features parameter values to be optimized. Thus taking all the combination of aforementioned values for the number of trees and number of features, 13 x 3 pairs will be tested to find the most optimal pair.

3) Tuning other algorithms: For AODE, Naïve Bayesian, Bagging and Boosting algorithms there are no concrete parameters to tune. Yet for Bagging and Boosting algorithms, number of iterations to be performed have to be assigned. Therefore for Bagging, 50 iterations were assigned according to Breiman's recommendations [9] and for Boosting, 100 iterations were assigned. However, in Boosting it will stop at less than 100 iterations if the necessary goal is reached. Furthermore the size of each bag has to be specified, as a percentage of the training set size for Bagging algorithms. So 100 was specified for this option since the bag size needs to be the same as the size of training set.

V. Results

For each of the four modules, four approaches will be considered as depicted in Figure 6. First approach is the traditional approach of classification, which is, applying a single classification algorithm (NB, AODE and C4.5 Decision Tree) for the original dataset and then, according to the results of the first approach, second, third and fourth approaches will be considered respectively. Second approach is the use of multiple classifiers (Bagging, Boosting and RandomForest) in order to increase the accuracy as well as to overcome the data imbalance problem if exists, whereas the third approach is the use of sampling techniques (oversampling and undersampling techniques) for the training data set in order to overcome the data imbalance problem if the problem of data imbalance exists. Final and the newest approach proposed by this research is the use of hybrid method which combines the second and third approaches.

Results of model evaluation for first module (employment status prediction) are extensively described below and a summary of the results of the other 3 modules will be given in conclusion section.

Figure 7 shows the dispersion of target class (employment status). It signifies the fact that there is a higher probability of having class imbalance problem in training datasets (since employed: unemployed: underemployed class ratio is close to 6:3:1). In subsequent sections, how this problem has affected the prediction outcomes of each class will be analysed.

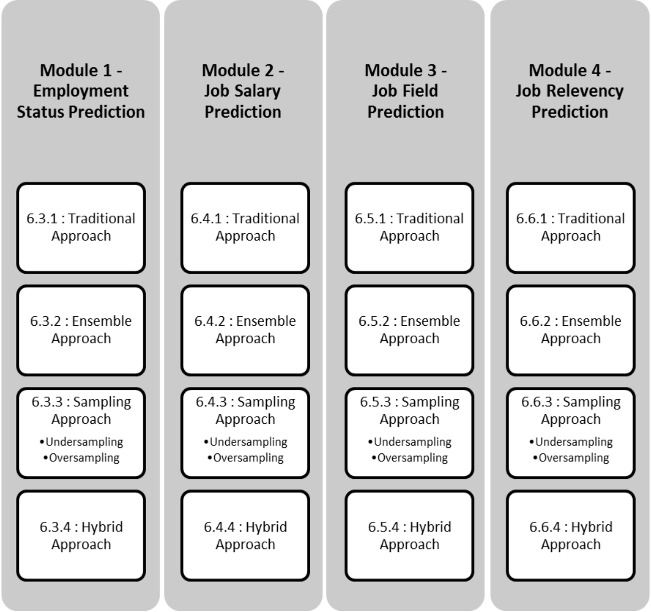
Module 1 - Employment Status Prediction
6.3.1 : Traditional Approach
6.3.2 : Ensemble Approach
6.3.3 : Sampling Approach
•Undersampling
•Oversampling
6.3.4 : Hybrid Approach
Module 2 - Job Salary Prediction
6.4.1 : Traditional Approach
6.4.2 : Ensemble Approach
6.4.3 : Sampling Approach
•Undersampling
•Oversampling
6.4.4 : Hybrid Approach
Module 3 - Job Field Prediction
6.5.1 : Traditional Approach
6.5.2 : Ensemble Approach
6.5.3 : Sampling Approach
•Undersampling
•Oversampling
6.5.4 : Hybrid Approach
Module 4 - Job Relevency Prediction
6.6.1 : Traditional Approach
6.6.2 : Ensemble Approach
6.6.3 : Sampling Approach
•Undersampling
•Oversampling
6.6.4 : Hybrid Approach

Fig 6 Overview of evaluation procedure

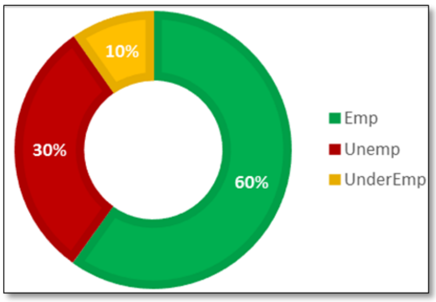

Fig 7 Dispersion of Employment status classes

A. *Traditional Approach – applying single algorithms*

The three diagrams shown in Figure 8(a),(b),(c) are the colour coded confusion matrices for each 3 classifier models. The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabelled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions (in the colour coded matrices, more blue implies higher number of correct classifications while less blue implies lesser number of correct classifications). Thus, from Figure 8 it can be seen that in all 3 classifiers, majority class (upper row of matrices) has performed really well compared to other two classes while the minority class (last row of matrices) has performed poorer than the other two classes, illustrating the symptoms of data imbalance phenomena.

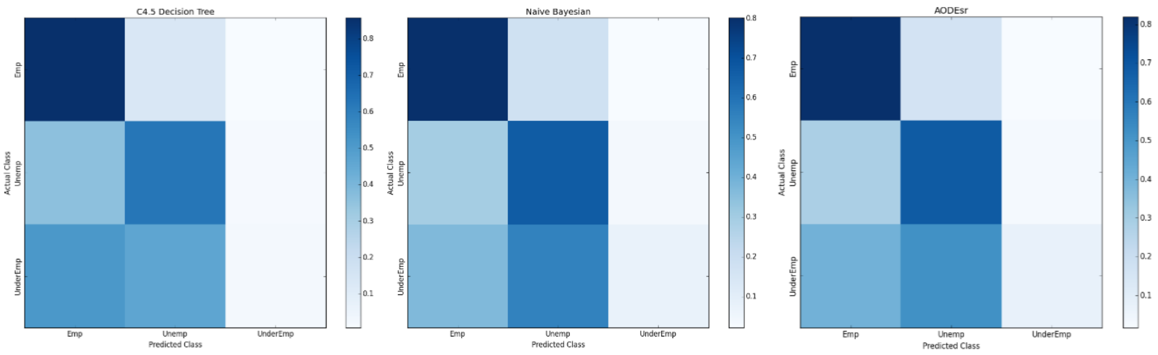

Fig 8(a),(b),(c) : Confusion Matrices for C4.5, NB and AODE (Traditional)

1) *Individual Evaluation Measures:* When the class wise performance measures were analyzed by individual evaluation measures illustrated in Table V, it can be seen that the prediction outcomes of minority class ('UnderEmp' class) has given unsatisfactory results in all three classifier models. Yet, the evaluation measure, accuracy, which is insensitive to class imbalance phenomena, shows the best results on minority class. Evaluation measures which are sensitive to class imbalance phenomena, such as Precision and Recall gives reflective figures to notify that the class imbalance problem exists in the dataset.

TABLE V: PREDICTION PERFORMANCE ON INDIVIDUAL MEASURES (*E-EMPLOYED, N-UNEMPLOYED, U- UNDEREMPLOYED*)

Class	Accuracy			Precision			Recall		
	C4.5	**NB**	**AODE**	**C4.5**	**NB**	**AODE**	**C4.5**	**NB**	**AODE**
E	0.7	0.7	0.8	0.76	0.79	0.80	0.86	0.80	0.82
N	0.8	0.7	0.7	0.60	0.56	0.59	0.63	0.67	0.69
U	0.9	0.9	0.9	**0.21**	**0.24**	**0.29**	**0.02**	**0.08**	**0.08**
All	**0.7**	**0.7**	**0.7**	**0.66**	**0.67**	**0.68**	**0.71**	**0.69**	**0.71**

2) *Combined and Graphical Evaluation Measures:* Since F-measure and G-means formulated based on the combination of Precision, Recall, and Specificity, impact from insensitive evaluation measures are abolished by these combined evaluation measures. Thus both F-measure and G-means of all three classifier models give unsatisfactory results on minority class. ROC AUC gives satisfactory results on minority class since it could not capture the impact from class imbalance problem.

TABLE VI: PREDICTION PERFORMANCE ON COMBINED MEASURES (*E-EMPLOYED, N-UNEMPLOYED, U- UNDEREMPLOYED*)

Class	Accuracy			Precision			Recall		
	C4.5	**NB**	**AODE**	**C4.5**	**NB**	**AODE**	**C4.5**	**NB**	**AODE**
E	0.8	0.8	0.8	0.72	0.74	0.75	0.81	0.83	0.84
N	0.6	0.6	0.6	0.72	0.72	0.74	0.79	0.81	0.82
U	**0.1**	**0.1**	**0.1**	**0.15**	**0.27**	**0.27**	0.68	0.72	0.73
All	**0.7**	**0.7**	**0.7**	**0.71**	**0.72**	**0.73**	**0.79**	**0.81**	**0.83**

B. *Ensemble Algorithm Approach*

According to the results given by both Bagging and Boosting methods, it is evident that the original classification performance of the minority class have not improved as expected. Results are similar to the values retrieved from traditional approach. Figure 9 graphically summarize the effect of applying Bagging technique on the training dataset.

It is clear that applying Bagging technique directly on training dataset couldn't effectively increase the classification performance of the minority class significantly. Even though the F-measure and G-means have been slightly increased in C4.5 classifier and AODE classifier, that difference is not significant enough. Moreover Boosting shows similar results.

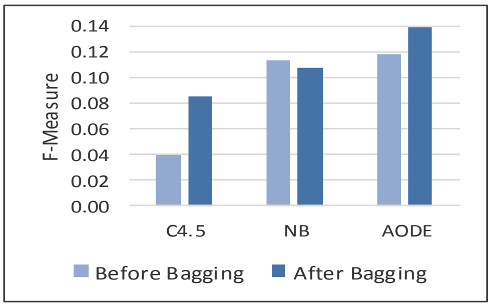

Fig 9 F-measure on minority class before and after Bagging

C. *Sampling Approach*

In oversampling, minority classes' instances will be replicated until original class ratio Emp: Unemp: UnderEmp reaches from 6: 3: 1 to 1: 1: 1. The 3 diagrams in Figure 10 show the confusion matrices for three classifier models separately. From the diagonals, it can be seen that in all 3 classifiers, even though majority class has performed really well compared to other two classes, the performance of other two classes are not significantly low compared to the majority class, indicating no symptoms of data imbalance phenomena anymore. Undersampling shows similar results.

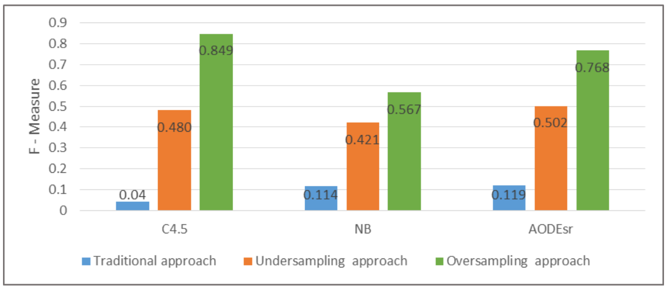

Fig 11 Performance comparison between 3 approaches (minority class)

Figure 11 attempts to summarize the F-measure results derived from traditional approach, undersampling approach and oversampling approach, with respect to classification performance of minority class. When the approaches given in this figure were compared, it's quite apparent that performance measures derived from oversampling have surpassed the performance results derived from

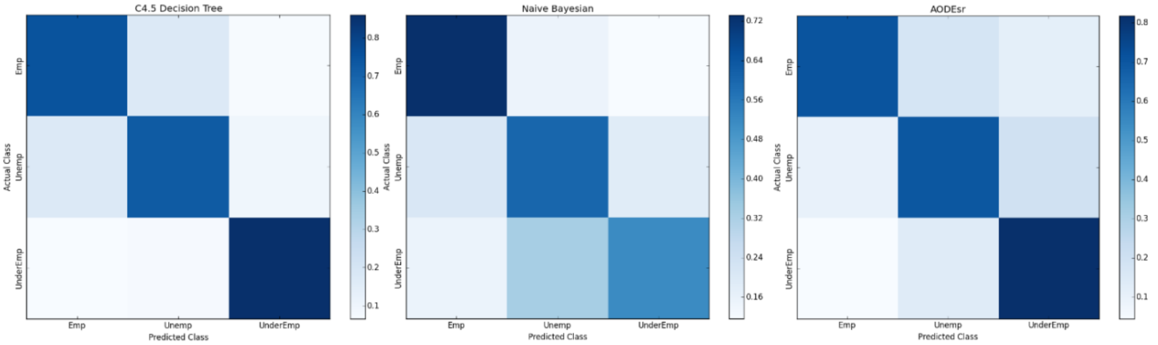

Fig 10(a),(b),(c) : Confusion Matrices for C4.5, NB and AODE (Oversampled) undersampling in minority classes.

D. Hybrid Approach

From sampling approaches it is already shown that oversampling technique has surpassed the results of undersampling. Hence in this hybrid approach, oversampling was used as the sampling technique to apply on Ensemble approaches.

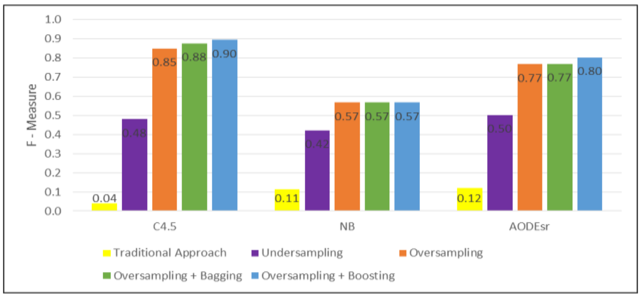

Fig 12 F-measure comparison on minority class in all 5 approaches

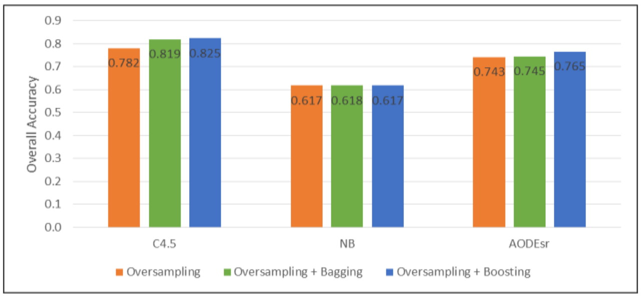

Fig 13 Overall accuracies among 3 approaches related to oversampling

When the Figure 12 is analyzed, it's quite apparent that the potential hybrid approach hasn't significantly augmented the results of oversampling. Although it has surpassed the evaluation measures initially acquired through the traditional approach significantly, Bagging or Boosting haven't add a significant benefit to original oversampling results for minority classes. Hence it can be concluded that the improvement on this minority class classification embedded in this hybrid approach merely contributed from the oversampling setting.

From Figure 13, it can be seen that both hybrid approaches (i.e. Oversampling+Bagging and Oversampling+Boosting) have slightly improved the overall accuracy of sole oversampling approach in both C4.5 classifier and AODE classifier, while for Naïve Bayesian classifier the accuracy have not been changed. The possible reasoning behind this might be that stable classifiers like NB do not respond well for bagging and boosting.

TABLE VII: PREDICTION PERFORMANCE ON INDIVIDUAL MEASURES(OVERALL)

Class	Bagging			Boosting			RF
	C4.5	NB	AODE	C4.5	NB	AODE	
Accuracy	0.82	0.62	0.75	0.83	0.62	0.77	**0.85**
AUC	0.94	0.80	0.90	0.94	0.75	0.88	**0.96**
G-means	0.86	0.71	0.81	0.87	0.71	0.82	**0.89**
F-measure	0.75	0.58	0.70	0.76	0.58	0.70	**0.79**

As previously shown, hybrid approach has recorded the best performance out of all four approaches and from Table VII, it can be seen that, RF ensemble method has shown the highest classification performance when considering all four measures under hybrid approach. Thus it can be concluded that for predicting the employment status using this dataset, RF with hybrid approach is the best technique. The order of other classifiers using the accuracy measure under the hybrid approach is C4.5+Boosting, C4.5+Bagging, AODE+Boosting, AODE+Bagging and lowest performance from NB.

VI. DISCUSSIONS

In first, third and fourth modules (employment status prediction, job field prediction, job relevance prediction), at the beginning (i.e. in our traditional approach), any of the classifier algorithms did not produce very good results with respect to the minority class due to the class-imbalance problem. Thus we experimented few mechanisms to improve this classification performance of minority class while also increasing the overall classification performance. For all these 3 modules, the order of classification performance with respect to different approaches is shown below in decreasing order.

1. Hybrid approach (Oversampling + Ensemble learning)
2. Oversampling
3. Undersampling
4. Ensemble learning
5. Traditional approach

Even though in these three modules, hybrid approach has shown the highest performance, hybrid approach hasn't significantly augmented the results of oversampling. Hence we can conclude that the improvement on this minority class classification embedded in this hybrid approach merely contributed from the oversampling setting and furthermore ensembling has not added a significant benefit to original oversampling results.

When considering the second module (job salary prediction), the results are bit different. Even though there was a class imbalance in this dataset as well, this has not affected the classification performance of minority class. Hence the imbalance data has not been a problem to the classification of minority class. Thus we did not carry out sampling approach and hybrid approach. But we carried out ensemble approach only to improve the overall classification performance, and we showed that applying ensemble method has slightly increased the performance of the traditional approach.

When we consider the traditional approach of three modules which had the class imbalance problem, in all three modules, C4.5 decision tree classifier has shown the poorest performance on minority class while AODE has shown the highest performance in minority class as well as in overall classification as well.

Furthermore when we compare the results under ensemble approach, in all three modules, ensemble approach has not been able to mitigate the class-imbalance problem. But this ensemble method has significantly increase the minority class performance in C4.5 classifier. Yet this increase has not been significant enough to rectify class-imbalance problem. Moreover in some of these 3 modules, ensemble approach has slightly decreased the minority class performance of AODE and NB classifiers.

Both undersampling approach and oversampling approach has significantly enriched the classification performance of minority class in all these 3 modules. But the difference is while enhancing the minority class performance, undersampling has decreased the majority class performance significantly while oversampling has been managed to increase or keep the traditional majority class performance as it is. Thus it can be concluded that in all three modules, classification performance of oversampling has surpassed the results of undersampling. It's also observed that after applying any of these sampling techniques C4.5 classifier has achieved the highest percentage uplift on their evaluation measures with respect to minority class.

As mentioned before applying hybrid technique in all these 3 modules has slightly increased the performance of minority class when compared to oversampling technique in C4.5 classifier. Yet for AODE and especially for NB classifier applying bagging or boosting has not changed the classification performance at all in 2 of 3 modules. Moreover it is perceived that in C4.5 and AODE classifiers, minority class performance has been increased more in oversampling with boosting technique compared to oversampling with bagging technique. But when we consider the overall performance in NB and AODE classifiers oversampling with boosting has decreased the classification performance than the original oversampling approach. The most probable reason for these unchanged or decreased performance of NB and AODE classifiers after applying an ensembling method is ensemble methods usually works best at unstable classifiers (like decision trees). For stable classifiers like NB and AODE ensemble methods will not do a much.

When we consider the RandomForest method, which is a decision tree based ensemble method, the hybrid approach has been able to significantly increase the performance of in all these three modules.

VII. Conclusions and Future Work

Table VIII gives the summary of best models selected under each of the four modules with the different combined and graphical evaluation measures. A significant result which can be seen from the following table is in all 4 modules, RandomForest based approaches have been selected as the best model and apart from job salary prediction module, best models selected in all the modules have been able to take AUC values greater than 90% indicating that all of them are 'Excellent' experiments according to AUC interpretation.

From this research we have proved that applying a machine learning approach to predict employability is a plausible option, given that the constraints embedded in employability data (like class imbalance) is properly handled.

Another major objective of this research was identifying the important factors relevant to each of four modules and Table 2 shows the factors which are relevant to these modules separately, according to its importance.

TABLE VIII: Summary of best models for 4 modules

Module	**Employment Status Prediction**	**Job Salary Prediction**	**Job Field Prediction**	**Job Relevance Prediction**
Approach	Hybrid	Ensemble	Hybrid	Hybrid
Classifier	RF with oversampling	C4.5 after Bagging	RF with oversampling	RF with oversampling
ROC AUC	95.7%	80.6%	95.9%	98.2%
Accuracy	85.0%	58.4%	78.6%	93.0%
F-measure	78.7%	58.0%	78.7%	89.3%
G-Means	88.7%	69.7%	87.0%	93.9%
Kappa statistic	77.0%	42.0%	75.0%	88.0%
AUC interpretation	Excellent experiment	Very Good experiment	Excellent experiment	Excellent experiment

The tangible benefits derived from these four types of prediction models can be revealed by implementing these selected best four models in a web system so that current undergraduates can use this system to predict their future related to employability and enhance their skills before they complete the degree, until this system predicts their desired employment status, job field and job salary.

Even though we have been able to achieve very good results on first, third and fourth modules, classification performance is comparatively less in second module (i.e. salary prediction). Thus, it is better to consider other machine learning algorithms such as Neural Network methods, SVM, CART, Bayesian networks, etc. for the prediction of job salary in order to see whether a different algorithm could increase this performance.

Cost sensitivity learning was not carried out when trying to overcome the class imbalance problem since we did not know the costs of misclassification in each class at the learning time. However as a future work, if cost of misclassification for each class can be defined, cost sensitivity learning and cost curve may do even more better classification than the class imbalance mitigation techniques we applied. Furthermore in model evaluations we only compared the AUC of ROC curves only. However we explained that ROC AUC is not a measure which is sensitive to class imbalance problem. Even though we had G-means and F-measure which are sensitive to class imbalance phenomena, as a future work it is suggested to compare the results using AUC of PR (Precision-Recall) curve as well since it is one of the best measures to evaluate imbalanced data.

Acknowledgment

We would like to thank Dr. Asoka Ramanayake and the team at HETC for providing their graduand employment census 2014 dataset for this research work.

References

[1] A. G. W. Nanayakkara. *Employment and Unemployment in Sri Lanka: Trends, Issues, and Options*. Department of Census and Statistics, 2004.

[2] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. "The elements of statistical learning: data mining, inference and prediction." *The Mathematical Intelligencer*, vol. 27, pp. 83-85, 2005.

[3] B. Jantawan, and C. Tsai. “The Application of Data Mining to Build Classification Model for Predicting Graduate Employment.” *International Journal of Computer Science and Information Security*, vol. 11, pp.1-8, Oct. 2013.

[4] B. Jantawan, and C. Tsai. “A Classification Model on Graduate Employability Using Bayesian Approaches : A Comparison.” *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 6, pp. 4584–4588, 2014.

[5] M. A. Sapaat, A. Mustapha, J. Ahmad, and K. Chamili. “A Data Mining Approach to Construct Graduates Employability Model in Malaysia.” *International Journal on New Computer Architectures and Their Applications*, vol. 1, pp. 1111–1124, 2011.

[6] R. S. J. D. Baker, and K. Yacef. “The State of Educational Data Mining in 2009 : A Review and Future Visions.” *Journal of Educational Data Mining*, vol. 1, pp. 3-16, 2009.

[7] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical machine learning tools and techniques*, 2005.

[8] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali. “A comparative study of decision tree ID3 and C4.5.” *International Journal of Advanced Computer Science and Applications*, vol. 4, pp. 13–19, 2014.

[9] L. Breiman. “Bagging Predictors.” *International Conference on Machine Learning*, vol. 24, pp. 123–140, 1996.

[10] Y. Freund, R. E. Schapire, and M. Hill. “Experiments with a New Boosting Algorithm,” in *International Conference on Machine Learning*, pp. 148–156, 1996.

[11] L. Breiman. “Random Forests.” *International Conference on Machine Learning*, vol. 45, pp. 5–32, 2001.

[12] N. V. Chawla. “Data Mining for Imbalanced Datasets: An Overview.” *Data Mining and Knowledge Discovery Handbook*, pp. 853–867, 2005.

[13] R. Longadge, S S. Dongre, and L. Malik. “Class Imbalance Problem in Data Mining : Review” *International Journal of Computer Science and Network (IJCSN)*, vol. 2, 2013.

[14] C. F. Aliferis, A. Statnikov, and I. Tsamardinos. “Challenges in the Analysis of Mass-Throughput Data : A Technical Commentary from the Statistical Machine Learning Perspective.” *Cancer Informatics,*, vol. 2, 2006.