IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 22, 2020, accepted August 12, 2020, date of publication August 21, 2020, date of current version September 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018663

Feature Selection for UK Disabled Students' Engagement Post Higher Education: A Machine Learning Approach for a Predictive Employment Model

DRISHTY SOBNATH[1], TOBIASZ KADUK[2], IKRAM UR REHMAN[3], (Member, IEEE), AND OLUFEMI ISIAQ[2]

[1]Department of Research, Innovation, and Enterprise, Solent University, Southampton SO14 0YN, U.K.
[2]School of Computing, Solent University, Southampton SO14 0YN, U.K.
[3]School of Computing and Engineering, University of West London, London W5 5RF, U.K.

Corresponding author: Drishty Sobnath (drishty.sobnath@solent.ac.uk)

This work was supported by the Research, Innovation, and Enterprise Department and the School of Computing, Solent University.

ABSTRACT While only 4.2 million people out of a population of 7.9 million disabled people are working, a considerable contribution is still required from universities and industries to increase employability among the disabled, in particular, by providing adequate career guidance post higher education. This study aims to identify the potential predictive features, which will improve the chances of engaging disabled school leavers in employment about 6 months after graduation. MALSEND is an analytical platform that consists of information about UK Destinations Leavers from Higher Education (DLHE) survey results from 2012 to 2017. The dataset of 270,934 student records with a known disability provides anonymised information about students' age range, year of study, disability type, results of the first degree, among others. Using both qualitative and quantitative approaches, characteristics of disabled candidates during and after school years were investigated to identify their engagement patterns. This article builds on constructing and selecting subsets of features useful to build a good predictor regarding the engagement of disabled students 6 months after graduation using the big data approach with machine learning principles. Features such as age, institution, disability type, among others were found to be essential predictors of the proposed employment model. A pilot was developed, which shows that the Decision Tree Classifier and Logistic Regression models provided the best results for predicting the Standard Occupation Classification (SOC) of a disabled school leaver in the UK with an accuracy of 96%.

INDEX TERMS Disability, feature selection, job predictors, machine learning, MALSEND, predictive model, special educational needs.

I. INTRODUCTION

Special Educational Needs and Disabilities (SEND) refers to students with requirements for education support as it is harder to learn due to a health condition or physical disability [1]. Concerns such as access to quality support or wrong career advice for disabled students were highlighted during seminar interviews carried out by steering groups [2]. Many students suffer from disabilities without a regular income or support, which eventually leads to having a negative impact on their quality of life and stability. The employment rate among the disabled population is still low even though there has been a slight increase in the last couple of years [3]. Only 4.2 million (53.2%) out of the 7.9 million disabled working population are currently in work compared to 81.4% of people without any disabilities in employment. Approximately, 3.4 million disabled people within the working-age bracket are ''economically inactive'', meaning that they are not in work and also not looking to work while about 300,000 people with disabilities are unemployed based on the government's latest figures [4].

The UK government is actively working towards reducing the gap in employment between the disabled and non-disabled workforce. The government launched the Work

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Imran Tariq.

and Health Programme (WHP) throughout England and Wales a few years go to help people with a wide range of health conditions or disabilities to enter into and stay in work using the expertise of private, public and voluntary, and community sector providers [5]. It is aiming to employ 1 million people with disabilities by 2027 [4]. Innovations in technology, and a range of other initiatives such as ''Access to Work'' grants provided to companies, have made it easier for organisations to consider more disabled employees. More students are going to university than ever before [3], so there is a need to ensure that these students secure work after studies and receive quality career guidance advice throughout their time at university to make better-informed decisions in choosing career paths.

Tertiary education providers worldwide are adopting new methods and technologies to meet the disability needs of their students. However, research shows the dropout rate for disabled students is much higher at 31.5% when compared with 12.3% for non-disabled students in the EU. There is a need for students to keep engaging with their studies throughout their degree, so as to the need for effective career guidance in completing their studies and stepping into the job market. This article builds upon the conceptual model developed in our previous study [6]. In this study, we applied Machine Learning (ML) techniques to uncover characteristic patterns among UK disabled students post higher education, mainly in terms of their engagement status within a 5-year range.

The proposed model is an analytical platform for large datasets, which aims to investigate and discover the job characteristics of disabled candidates post higher education by using a machine learning approach. Machine learning is a subset of artificial intelligence (AI) that supports processors or machines to learn from previous data to make intelligent decisions [7]. To build a good predictive model, a feature engineering process is completed to identify useful predictors. This study examines a large UK dataset for disabled students between 2012 and 2017 using suitable machine learning algorithms.

The rest of the paper is organised as follows. Section II provides the background on the need of the proposed study. Section III presents the methodology used to develop the proposed platform using approaches like Exploratory Data Analysis (EDA), Data Encoding, Dimensionality Reduction, etc., to manipulate the large dataset. The results of the selected features are then presented in Section IV after applying selected machine learning algorithms. In Section V, we described and discussed the pilot study that has been developed to evaluate the feature the selection of students' engagement post higher education. Section VI concludes the paper.

II. BACKGROUND

A. CHALLENGES OF EMPLOYMENT

A persistent employment gap for disabled people is one of the several employment inequalities people face [8]. While a slight improvement among disabled people in work has been observed in the last 4 years, a considerable contribution is still required from universities and industries to increase employability among the disabled. In particular, by providing adequate guidance on careers to achieve a level of balanced employment [9]. The Trades Union Congress report in 2019 emphasizes the Labour Force Survey, showing that only 14.8% of people with learning difficulties are in employment. Similarly, other conditions including speech impediments (20.4%), epilepsy (33.6%), mental illness (33.7%) and progressive illness e.g., cancer or HIV (45.2%), depression, bad nerves (46.4%), heart, blood pressure, circulation (48.2%) and visually impaired people (48.3%) also recorded low employability [9]. Despite the UK government's effort to encourage employers and recruitment agencies to provide more opportunities to those with a learning or physical disability, the disability employment gap remains a problem to be solved [10].

Several studies show that disabled students struggle to find jobs after graduating and perform poorly compared to their peers [11]. A few companies such as Disability Jobsite or Evenbreak assist disabled candidates in actively looking for jobs and support them through the whole process i.e., from job surf, application, interview, and the pathway to work [12], [13]. These companies work closely with potential employers who take into consideration the factor of inclusiveness. Some organisations explicitly hire people with specific disabilities. For example, autistic people have been allowed to work for Aspiritech, a software testing company in the United States [14], whose mission is to empower individuals on the autism spectrum to fulfill their potential. Similarly, other companies, including SAP, Microsoft Corporation, Ford Motor Company, DXC Technology, and Ernst and Young, even have specific employment programmes for autistic people [15]. However, there is a lack of clarity of what type of jobs disabled students are more likely to secure after graduation from a higher education institution.

To overcome the research gap, this study builds on constructing and selecting subsets of features useful to build a good predictor regarding the engagement of disabled students in employment using the big data approach with machine learning principles.

While autistic people have been employed in selected areas in the US, the common occupational fields for people with hearing impairments have also been in the medical industry. About 13.7% of hearing people are employed in the medical field, while the least common field is in extraction, with 0.6% of hearing people in this field. On the other hand, for deaf people, the most common field is manufacturing, with 13.2% of deaf people employed in this field, and the least common field is utilities, with 1.1% of deaf people working in this field [16]. However,in the UK, further research is required on large datasets to understand the trends among university graduates, their disability and their employability.

B. EMPLOYABILITY PREDICTORS

A systematic review carried out by some authors [17], [18] shows that across 13 studies, a total of 7 unique predictors

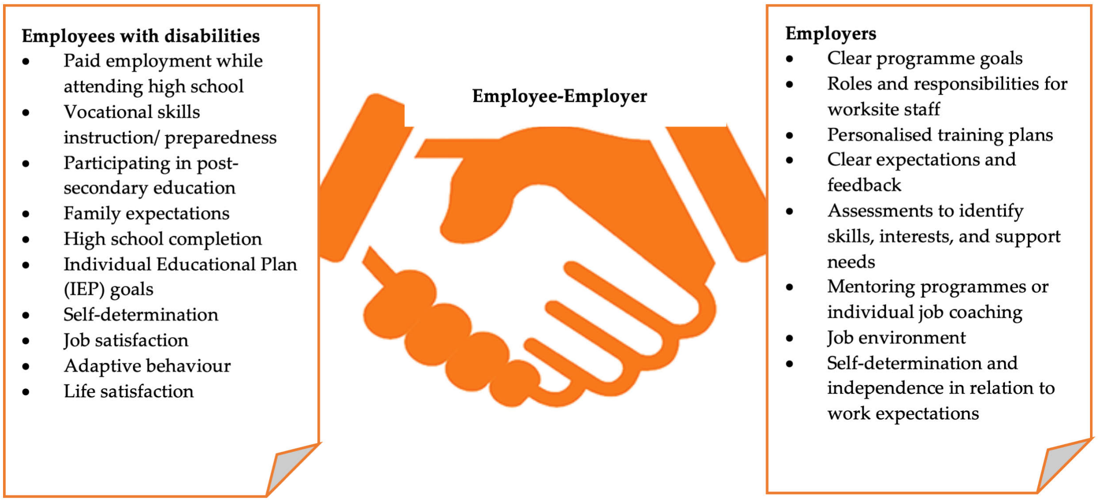

Employees with disabilities
Paid employment while attending high school
Vocational skills instruction/ preparedness
Participating in post-secondary education
Family expectations
High school completion
Individual Educational Plan (IEP) goals
Self-determination
Job satisfaction
Adaptive behaviour
Life satisfaction
Employee-Employer
Employers
Clear programme goals
Roles and responsibilities for worksite staff
Personalised training plans
Clear expectations and feedback
Assessments to identify skills, interests, and support needs
Mentoring programmes or individual job coaching
Job environment
Self-determination and independence in relation to work expectations

FIGURE 1. Employee-Employer handshake diagram to illustrate predictors of competitive employment.

of post-secondary competitive employment were identified such as paid employment while attending high school, Individual Educational Plan (IEP) goals or vocational skills [19], as shown in Figure 1. These predictors, alongside other factors, were found to help students with intellectual or developmental disabilities to secure a job after post-secondary education. Another study that looked at the relationship between work and disability shows factors important to consider by employers such as clear programme goals, roles, and responsibilities for worksite staff, personalised training plans, clear expectations, and feedback. Assessments to identify skills, interests, and support needs are also important before employing someone with a disability as well as mentoring programmes, or individual job coaching should be made accessible within the organisation. Other studies identify similar predictors such as personal experiences, vocational preparedness, job satisfaction, related environment, adaptive behaviour and life satisfaction [20], [21], which are all contributors for a person with disabilities to retain a job [22]. Other elements that appear to be key in this process for disabled workers are the feelings of self-determination and independence about to work expectations [15]. The predictors mentioned above for competitive employment are represented in Figure 1.

Recently named as the "Graduate Employer of the Year" [23], the UK Civil Service department, being one of the major employers, has a vital role to play when it comes to recruiting employees with disabilities. Being at the forefront of good practice and being a leader, the public sector can do better and provide inspiration examples to other employers in the private sector [24]. The aforementioned studies make use of general statistical analysis methods and have analysed small sample size. In this study, a large dataset of more than 270,000 student records for the past five years will be analysed using machine learning algorithms to detect any features that can be important to predict the engagement of a disabled student after graduation. Good career advice, securing a job, and contributing to the country's economy will be a real and lasting change to disabled people. This study can further build positive perceptions and promote awareness of the real capabilities of certain people with disabilities.

III. METHODOLOGY

This research utilised both qualitative and quantitative approaches to help answer the research questions: 1) whether large datasets of past disabled students' can be exploited by machine learning algorithms to provide insights on their employability? and 2) can an efficient predictive platform be built on the identified features to predict their employment type? Firstly, through the adapted PRISMA approach based on keyword search, a thorough literature review was conducted to identify some engagement predictors stated in previous studies. The big data approach with ML principles was then applied to a large dataset of student records to handle a depth of discoveries, which cannot be managed by traditional data handling methods and techniques by the previous researchers.

A. DATASET

One of the primary objectives is to identify suitable persistent data platforms that hold relevant information about historical academic background, disabilities,and related jobs within six months of graduation. A specific dataset with certain attributes was requested from the Higher Education Statistics Agency (HESA) since their platform holds data throughout the UK in a consistent format. 270,934 student records with a known disability were therefore gathered for this study.

This data consists of UK Destinations Leavers from Higher Education (DLHE) survey results, which provide anonymised information about students' age range, year of study, disability type, and results of the first degree from 2012 to 2017, among others. Ethical approval was obtained from Solent University Ethics Committee. The data otherwise is fairly distributed; for instance, academic year variables vary between 18% and 22% similarly, 40.3% of the students were male, while 59.7% were female. However, the variation in the count for age group and level DLHE appeared to be less well distributed, with 62% of the students in the age group 21-24 years old and 73% was doing a first degree. 51.7% reported they were in full-time work while 14.1% was working part-time. Others were carrying on further studies or were involved in other activities about 6 months after graduation.

TABLE 1. Dataset examples.

Field Variables	Field Description	Type of Variable	No of Categories	Unique Values	Count (Percentage)	
Academic Year (ACYEAR)	Year when the student graduated from the university	Independent	5	2012/13	49,767	18.4%
				2013/14	52,741	19.5%
				2014/15	52,906	19.5%
				2015/16	55,465	20.5%
				2016/17	60,055	22.5%
Age Group (F_XAGRPJ01)	Age group on 31 July in the reporting year.	Independent	6	17 years and under	44	.0%
				18-20 years	8187	3.0%
				21-24 years	168048	62.0%
				25-29 years	38989	14.4%
				30 years and over	55662	20.5%
				Age unknown	4	.0%
HE provider identifiers (F_INSTID)	HESA identification number for each university.	Independent	166	Examples: 0001 The Open University	UK university list can be found on HESA website	
				0002 Cranfield University		
Level of DLHE qualification (F_ZDLE501)	Illustrates the qualification level achieved by the student.	Independent	5	Other postgraduate	15594	5.8%
				First degree	197744	73.0%
				Other undergraduate	28444	10.5%
				Doctorate	3669	1.4%
				Masters	25483	9.4%
Class of first degree (F_XCL6SS01)	The undergraduate degree class that the student obtained. Applicable to first degree qualifiers only.	Independent	7	First class honours	40957	15.1%
				Upper second class honours	97085	35.8%
				Lower second class honours	41683	15.4%
				Third class honours/Pass	7911	2.9%
				Unclassified	10086	3.7%
				Classification not applicable	20	.0%
				Not applicable (not a first-degree leaver)	73192	27.0%
Mode of qualification (F_XQMODE01)	Refers to the method by which the qualification was achieved.	Independent	2	Full-time	230657	85.1%
				Part-time	40277	14.9%
Sex (F_SEXID)	This field records the sex of the student.	Independent	3	Male	109218	40.3%
				Female	161617	59.7%
				Other	99	.0%
Disability type (FXSTUDIS01)	The type of disability that a student has, based on the student's self-assessment.	Independent	10	Blind or a serious visual impairment	3404	1.3%
				Deaf or serious hearing impairment	5909	2.2%
				A physical impairment or mobility issues	8658	3.2%
				Personal care support	3	.0%
				Mental health condition	35936	13.3%
				A long-standing illness or health condition	27505	10.2%
				Two or more conditions	18051	6.7%
				Social communication/Autistic spectrum disorder	6880	2.5%
				Specific learning difficulty	140980	52.0%
				Another disability, impairment or medical condition	23608	8.7%
Highest qualification on entry (F_XQUALENT01)	The highest qualification that a student holds on entry.	Independent	10	Postgraduate (excluding PGCE	9386	3.5%
				PGCE	1330	.5%
				First degree	42236	15.6%
				Other undergraduate qualification	26282	9.7%
				Other qualification	2645	1.0%
				Level 3 qualification	180092	66.5%
				Qualifications at Level 2 and below	4706	1.7%
				No formal qualification	2732	1.0%
				Not known	1493	.6%
				Not applicable	32	.0%
Tariff band (F_TARIFF)	Tariff points based on the qualifications on the entry of the student.	Independent	13	1-79	1933	.7%
				80-119	2205	.8%
				120-179	6796	2.5%
				180-239	13381	4.9%
				240-299	27316	10.1%
				300-359	29824	11.0%
				360-419	30432	11.2%
				420-479	18837	7.0%
				480-539	10697	3.9%
				540-998	11282	4.2%
				999 and above	13	.0%
				NA	104106	38.4%
				Unknown	14112	5.2%
JACS- Joint Academic Coding System (F_XJACS01_1level)	The subject of study.	Independent	1081	(A100) Pre-clinical medicine (A200) Pre-clinical dentistry	List of unique JACS codes can be found on HESA website.	
Activity (F_XACTIV02)	Leavers self-identified most important activity within 6 months of graduation.	Target	9	Full-time work	140199	51.7%
				Part-time work	38189	14.1%
				Primarily in work and also studying	7670	2.8%
				Primarily studying and also in work	8882	3.3%

B. EXPLORATORY DATA ANALYSIS (EDA)

Data exploration is the preliminary investigation of the dataset to gain a better understanding of the students' data. To make optimum use of the available information, it is imperative that we learn the characteristics of the provided variables through summary statistics and visualisation techniques, as shown in Table 1 and Figure 3, respectively. Using these techniques (as shown in Figure 2), it was important to look for correlations, trends, and outliers that could have affected our analysis.

C. DATA CLEANSING

1) DEALING WITH MISSING VALUES

Usually, it is impractical to have a perfect dataset in the real-world hence, resulting in a negative performance of machine learning models. The dataset received, however, has already been partially cleaned and structured. Missing values were appropriately substituted in the pre-processing phase with unique values rather than following a non-parametric approach. For example, values including ''Not applicable'', ''Unknown'' and blank cells were replaced with unique values so that they do not affect our findings.

2) HANDLING HIGH CARDINALITY AND IMBALANCED DATA

We considered the dataset to be mainly of categorical hence, the biggest challenge of this project. Also, some of the provided variables are imbalanced with the appearance of more specific classes in our observations and some with high cardinalities such as HE Provider (n = 166), JACS (Joint Academic Coding System) code (n = 1081) or Industrial Classification (n = 89). High-cardinality nominal attributes can pose an issue for inclusion in machine learning predictive models, and therefore, be reduced before processing. JACS code, a way of classifying academic subjects and modules by the UK higher education institutions, consists of a 4-digit number such as N810, which represents the course ''Travel management''. However, HESA also has a 2-digit and subject classification; therefore we could easily further reduce this. In the above example, N810 was converted to N8, which is in the category of ''Hospitality, leisure, sport, tourism & transport'' and resulted in fewer subject areas (n = 20). HE Provider and Industrial Classification values could not be further reduced. While the handling of skewed data varies from techniques such as Log Transform, Square Root Transform, Box-Cox Transform [25] or SMOTE-NC (Synthetic Minority Over-sampling TEchnique-Nominal Continuous) [26], decision trees algorithms often perform well on imbalanced datasets [27], and therefore, have been used on the dataset.

D. DIMENSIONALITY REDUCTION

In order to achieve the second objective of this work, which is to investigate and discover the characteristics of disabled candidates during and after school years, the dataset was divided into a subset of 11 independent variables and 3 target

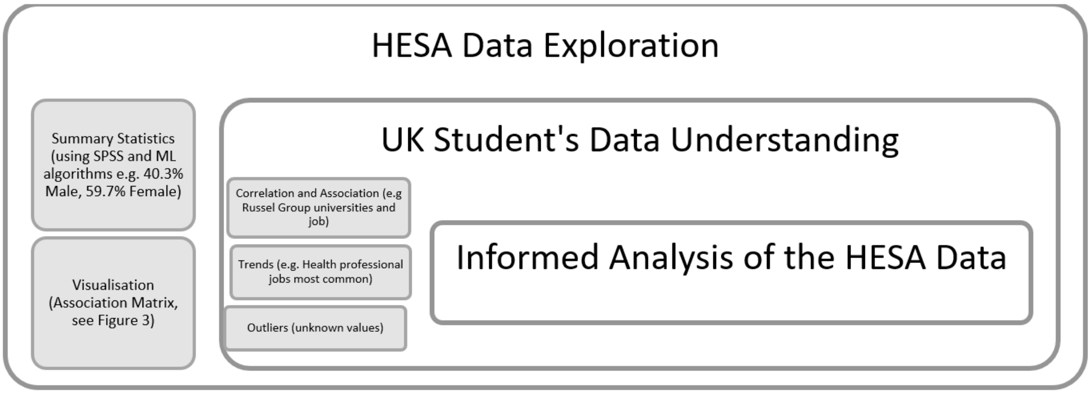
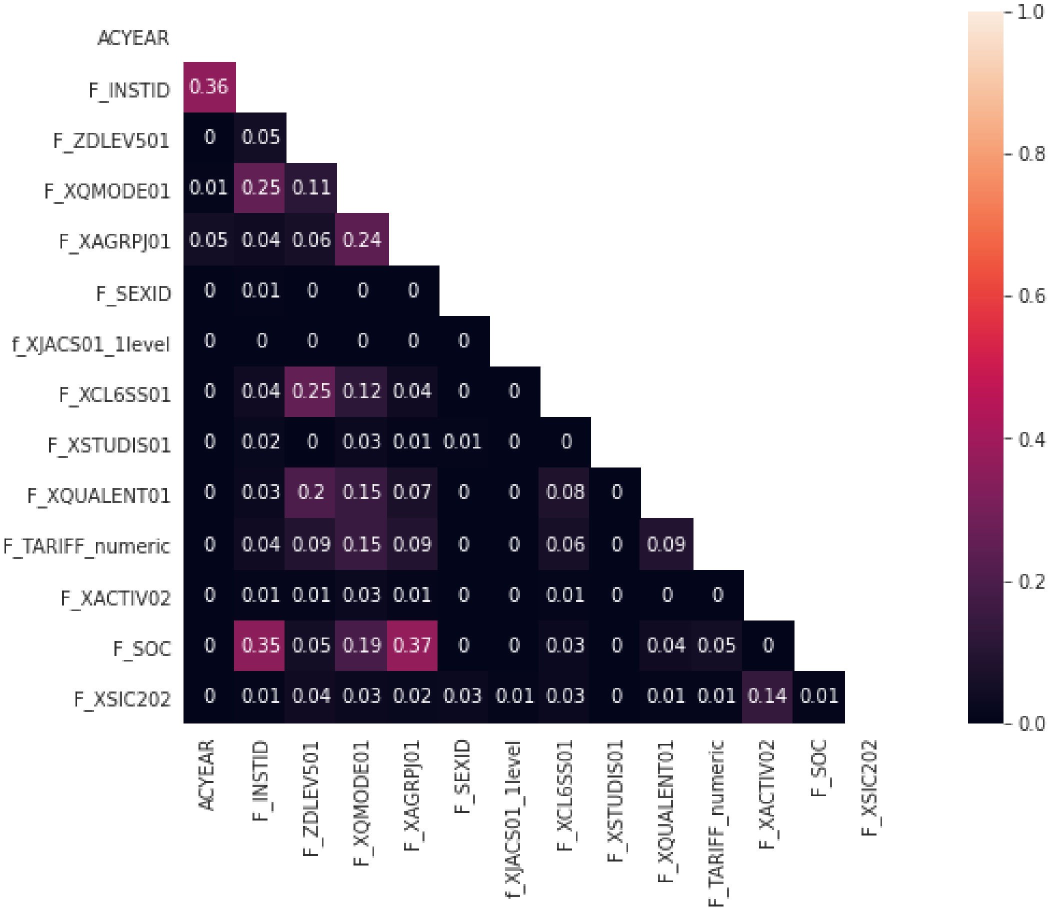

HESA Data Exploration
Summary Statistics (using SPSS and ML algorithms e.g. 40.3% Male, 59.7% Female)
Visualisation (Association Matrix, see Figure 3)
UK Student's Data Understanding
Correlation and Association (e.g Russel Group universities and job)
Trends (e.g. Health professional jobs most common)
Outliers (unknown values)
Informed Analysis of the HESA Data

FIGURE 2. Exploratory data analysis.



	ACYEAR	F_INSTID	F_ZDLEV501	F_XQMODE01	F_XAGRPJ01	F_SEXID	f_XJACS01_1level	F_XCL6SS01	F_XSTUDIS01	F_XQUALENT01	F_TARIFF_numeric	F_XACTIV02	F_SOC	F_XSIC202
ACYEAR														
F_INSTID	0.36													
F_ZDLEV501	0	0.05												
F_XQMODE01	0.01	0.25	0.11											
F_XAGRPJ01	0.05	0.04	0.06	0.24										
F_SEXID	0	0.01	0	0	0									
f_XJACS01_1level	0	0	0	0	0	0								
F_XCL6SS01	0	0.04	0.25	0.12	0.04	0	0							
F_XSTUDIS01	0	0.02	0	0.03	0.01	0.01	0	0						
F_XQUALENT01	0	0.03	0.2	0.15	0.07	0	0	0.08	0					
F_TARIFF_numeric	0	0.04	0.09	0.15	0.09	0	0	0.06	0	0.09				
F_XACTIV02	0	0.01	0.01	0.03	0.01	0	0	0.01	0	0	0			
F_SOC	0	0.35	0.05	0.19	0.37	0	0	0.03	0	0.04	0.05	0		
F_XSIC202	0	0.01	0.04	0.03	0.02	0.03	0.01	0.03	0	0.01	0.01	0.14	0.01	

FIGURE 3. The Association matrix using the Cramer's V method.

variables. The target variables were mainly Activity (employed, unemployed, studying), Standard Industrial Classification and Standard Occupational Classification. Dimensionality reduction, the process of reducing the number of random variables under consideration by obtaining a minimum number of parameters, was applied on the dataset. The "Unique Identifier" column was dropped for data analysis as it was only a 12-digit unique number that helped to identify each record. An Association Matrix, using the Cramer's V method, was then created to visually identify relationships among the variables [28], since most of the data obtained was classified as categorical data. The association matrix shows that there is some association among the variables in the dataset, such as Standard Occupational Classification (SOC)

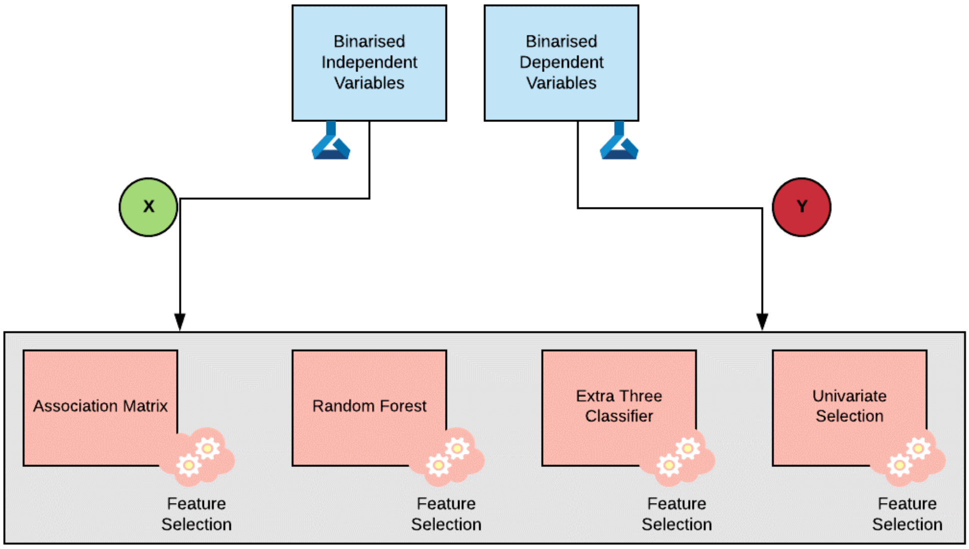


FIGURE 4. **Process of feature selection.**

and Age Group (Cramer's V = 0.37) or Institution and SOC (Cramer's V = 0.35). Since Cramer's V test's values are less than 0.5, this only shows a low association and further analysis needs to be carried out. Also, it is important to note that since there is no high association among the variables, none of the variables were dropped for future analysis.

E. DATA ENCODING

Tariff and Age variables, as seen in Table 1, were the only two numerical variables. However, the distance between two points, for example (Tariff Band 1-79 and 80-119), was not standardized or equal, leading to inconsistency in the interval range. Therefore, these two variables were also treated as categorical for future analysis. There are many different types of encoding, including Classical, Bayesian, Contrast, and more [29]. The two most popular techniques for categorical data are Label encoding and One Hot encoding [30]. One hot encoded method resulted in higher granularity with 261 independent columns and 119 dependent columns for the next phase of the analysis.

F. FEATURE SELECTION

This project aims to investigate the suitability of identified features for the development of a predictive model in terms of job selection for a disabled student. Variable features are trained using machine-learning models as irrelevant features in the data can decrease the model performance and accuracy. Three distinct feature selection techniques, namely 1) Random Forest, 2) Extra Tree Classifier, and 3) Univariate Selection, were adopted after creating the association matrix to identify impactful features of both target and independent variables. Subsequently, we selected the top 20 features present in at least two of the adopted feature selection models. The feature selection process is schematically represented, as shown in Figure 4.

1) FEATURE SELECTION ALGORITHMS

a: RANDOM FOREST

In our proposed study, random forest algorithms were first used on the dataset to identify features that could provide insights into the engagement of UK disabled students about 6 months after graduation. Random forest algorithms incorporate feature selection and interactions while they are efficient and provide high prediction accuracy [31]. The first selection showed that Age and Institution are two important variables to predict the engagement of a disabled student. An example of the feature selection using a random forest method is shown in Figure 5 before and after encoding the data. To improve the granularity of the variables, the same random forest algorithm was performed on one-hot encoded data to see what universities and age range were found to be important. Due to its advantages of being able to deal with small sample size, high dimensional features and complex data structures, it is a popular choice for many research projects.

b: EXTRA TREE CLASSIFIER

The main difference between Random Forest and Extra Tree Classifier lies in the fact that, instead of computing the locally optimal feature/split combination (for the random forest), for each feature under consideration, a random value is selected for the split (for the extra trees) [32]. This leads to more diversified trees and fewer splitters to evaluate when training an extremely random forest. To conclude on a set of selected features, this second method was applied on one-hot encoded data to provide more insights on the dataset. With the three main algorithms used, it was, therefore, more reliable to select features common in at least two of them. Figure 6 illustrates JACS code W, L, C, B, N, which represented subjects such as Creative Arts & Design, Social Studies, Biological Sciences,

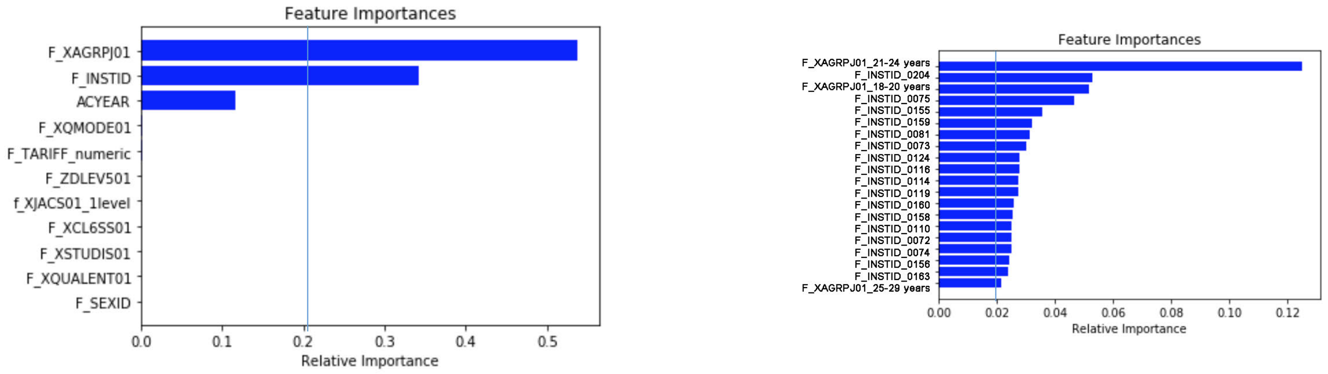
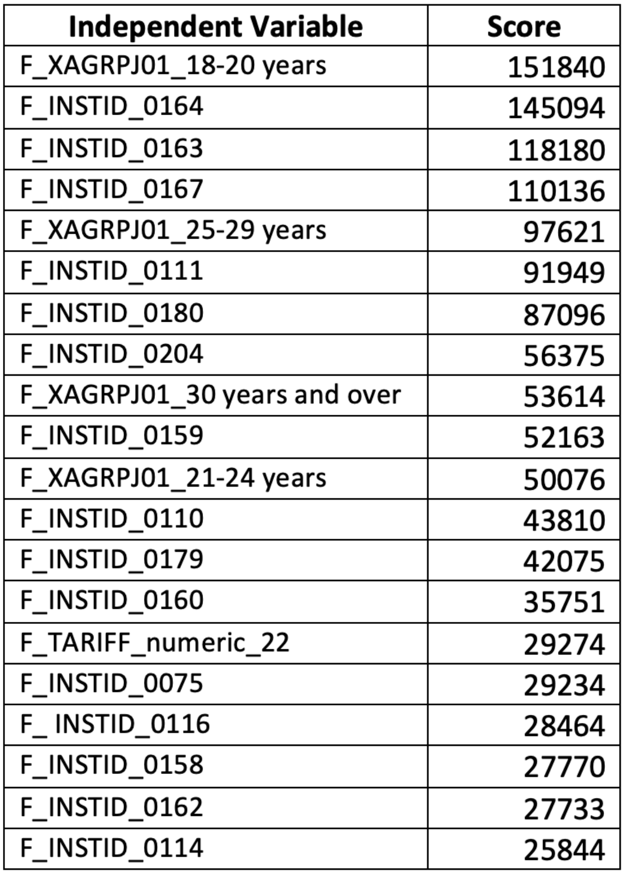


FIGURE 5. Random Forest feature selection before and after encoding with target variable "Standard Occupational Classification" (Job).

TABLE 2. Univariate Selection (Top 20 features).



Independent Variable	Score
F_XAGRPJ01_18-20 years	151840
F_INSTID_0164	145094
F_INSTID_0163	118180
F_INSTID_0167	110136
F_XAGRPJ01_25-29 years	97621
F_INSTID_0111	91949
F_INSTID_0180	87096
F_INSTID_0204	56375
F_XAGRPJ01_30 years and over	53614
F_INSTID_0159	52163
F_XAGRPJ01_21-24 years	50076
F_INSTID_0110	43810
F_INSTID_0179	42075
F_INSTID_0160	35751
F_TARIFF_numeric_22	29274
F_INSTID_0075	29234
F_ INSTID_0116	28464
F_INSTID_0158	27770
F_INSTID_0162	27733
F_INSTID_0114	25844

Subjects Allied to Medicine and Business and Administrative Studies came up as essential features.

c: UNIVARIATE SELECTION

Since the dataset consisted of categorical variables, the univariate method was appropriate to see the strength of the relationship among them. The top 20 features with the highest scores were included during the selection process,and important features were mainly universities and the age variables. These are listed in the Table 2.

According to many authors, univariate selection for feature selection can improve the accuracy of classification models [33], [34]. Univariate feature selection works by examining the effects of a single variable, such as Tariff Band or Class of Degree, on a set of data. Each feature to the target variable is compared to see whether there is any statistically significant relationship between them. It uses the chi-squared test, which belongs to the family of univariate analysis, i.e., those tests that evaluate the possible effect of one variable, the independent variable, upon an outcome, dependent variables). After performing the feature selection process to identify the engagement factors of disabled students, the features identified in at least two algorithms were listed, and the results are discussed in the next section.

IV. RESULTS

This article builds on constructing and selecting subsets of features useful to build a good predictor regarding the engagement of disabled students 6 months after graduation. Following the adopted algorithms, features selected included age, HE institution, level of DLHE qualification, class of the first degree, disability type, highest qualification on entry, and JACS code. These features were selected based on a threshold of relative importance (20%) and top 20 features found to be common in at least 2 algorithms from the selection process.

A. SELECTED FEATURES

The selected features illustrated in Table 3, are further discussed in the subsequent sections.

1) AGE

The age variable appears a significant predictor with the age range 18-20 years, 21-24 years, 25-29 years to be important features. However, the feature selection algorithms used did not find the age ranges "17 and under" or "30 & over" as important factors even though this accounted for over 20% of the total population that falls under these age groups.

2) HE INSTITUTION

Thirteen universities were highlighted as important features from the independent variables. We noted that ten of these

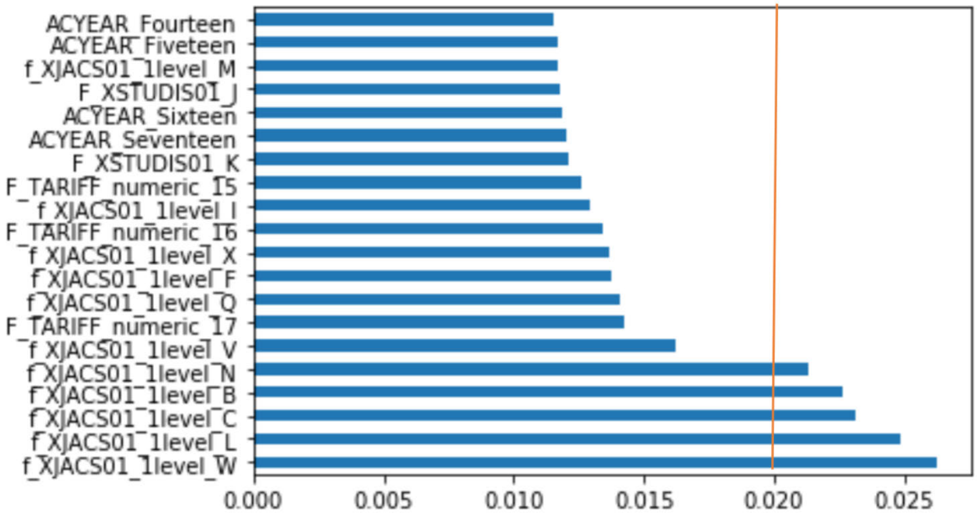


FIGURE 6. Extra-tree classifier algorithms on the activity (employed, unemployed, still studying, etc) on UK disabled graduate data.

TABLE 3. Selected Features from ML algorithms.

Independent Variables	Important Feature	Features Selected	% of each selected feature
1. Academic Year	No	-	N/A
2. Age Group	Yes	18-20 years 21-24 years 25-29 years	3% 62% 14.4%
3. HE Institution	Yes	Russel Group Universities Institution Other Universities	10.2% 5.1%
4. Level of DLHE qualification	Yes	Other Undergraduate Other Postgraduate	10.5% 5.8%
5. Class of first degree	Yes	Unclassified (pass/certificate of attendance)	3.7%
6. Mode of qualification	No	-	N/A
7. Sex	No	-	N/A
8. Disability Type	Yes	Specific learning difficulty	52%
9. Highest qualification on entry	Yes	First degree	15.6%
10. Tariff band	No	-	N/A
11. JACS Code	Yes	X (Education)	9.1%

selected universities were Russel Group Universities representing about 77% of the selected universities. Although most UK universities carry out similar activities for managing disabled candidates, the three other universities share tightly similar activities to the selected Russel Group universities. For example, the three non-Russel Group universities have a high number of disabled students who join their courses, have residential accommodation adapted for disabled students, and have many accessibility features such as assistive technologies, and are given priority when allocating residential campus rooms. Some of these universities have been adapted for students having hearing impairment issues, for example, rooms with a visual fire alarm and socket for vibrating pad alarm are available. Furthermore, additional adaptations may be made on an individual basis, subject to resources. One of the selected universities is the UK's largest providers of health and social care courses, teacher training, and sport and physical activity courses. These universities aim to support students through the transition period from further to higher education to raise disabled learners' aspirations, giving them the confidence to apply for higher education.

3) DISABILITY

In addition to features of the institutional variable, highly relevant features were highlighted from the disability variable. Specific learning difficulty (a cluster of disabilities such as dyslexia, dyspraxia, and ADHD) is highlighted as an essential feature, which can play a significant role in a predictive model. Accordingly, the HESA Disability records the type of disability or disabilities a student has, based on the student's own self-assessment upon enrollment. Code 51, which is "A specific learning difficulty such as dyslexia, dyspraxia or ADHD" (HESA,2020), was selected as compared to other disabilities during the feature engineering process by the machine learning algorithms.

4) LEVEL OF DLHE QUALIFICATION

The list from the dataset contained different levels of DLHE qualification, including Other Postgraduate, First degree, Other Undergraduate, Masters, and Doctorate. It is surprising to note that Other Postgraduate (5.8%) and Other undergraduates (10.5%) were highlighted as important features within the qualification variable even though only a small

TABLE 4. Class of Degree awarded (Results Classification).

	Gender	First class honours	Upper second class honours	Lower second class honours	Third class honours/pass	Unclassified	Classification N/A	N/A (not a first degree leaver)
	Male	15.3%	34.8%	16.4%	3.3%	3.7%	0.0%	26.6%
	Female	15.0%	36.6%	14.7%	2.7%	3.7%	0.0%	27.3%
	Other	25.3%	39.4%	8.1%	3.0%	2.0%	-	22.2%
Total		15.1%	35.8%	15.4%	2.9%	3.7%	0.0%	27.0%

TABLE 5. The 10 most and least common jobs secured by UK disabled HE leavers 2012-2017 by gender.

Job Type	Frequency	Total %	Male%	Female%
Most Common Jobs				
Health professionals	25236	9.3	21.9	78.1
Business and public service associate professionals	24259	9.0	41.9	58.1
Teaching and educational professionals	22297	8.2	27.5	72.5
Business, media and public service professionals	17304	6.4	43.9	56.1
Culture, media and sports occupations	16387	6.0	46.2	53.8
Science, research, engineering and technology professionals	13953	5.1	70.3	29.7
Sales occupations	12480	4.6	37.4	62.6
Caring personal service occupations	11663	4.3	18.6	81.4
Elementary administration and service occupations	9420	3.5	44.9	55.1
Administrative occupations	8961	3.3	32.4	67.6
Least Common Jobs				
Secretarial and related occupations	2401	0.9	19.2	80.8
Leisure, travel and related personal service occupations	2019	0.7	38.3	61.7
Textiles, printing and other skilled trades	1527	0.6	48.4	51.6
Protective service occupations	958	0.4	55.2	44.8
Elementary trades and related occupations:	461	0.2	77.9	22.1
Transport and mobile machine drivers and operatives	432	0.2	84.0	16.0
Process, plant and machine operatives	406	0.1	59.9	40.1
Skilled metal, electrical and electronic trades	398	0.1	88.4	11.6
Skilled agricultural and related trades	355	0.1	74.1	25.9
Skilled construction and building trades	350	0.1	89.4	25.9

percentage of the population had that qualification level. While other postgraduate category consists of postgraduate diplomas, certificates and professional qualifications such as Postgraduate Certificate in Education (PGCE), Diploma in Teaching and non-formal postgraduate qualifications. Whereas, other undergraduate category includes all undergraduate courses with the exclusion of bachelor's degrees such as foundation degrees, diplomas in higher education, the Higher National Diploma (HND) which could be interesting predictors of engagement of disabled students.

5) CLASS OF FIRST DEGREE

According to HESA, the class (First Class, Upper Second Class, Lower Second Class, etc.) of the award is given by higher education providers to UK students at the completion of their studies. After analysing more than 270,000 records from disabled students from 2012-2017, the algorithms found out that a significant predictor of a student being active and in full-time employment is not necessarily to be among those who receive a ''First Class'' honors' degree. However, ''Unclassified'' came up as a vital employability predictor. Both undergraduate and postgraduate degrees have a category ''unclassified'', and even though only 3.7% (Table 4) of the disabled students had an unclassified degree, the selection process included this as a significant predictor for an engaged student.

Unclassified undergraduate awards are those that operate on a simple pass/fail basis, for example, CertHE (Certificate of Higher Education), DipHE (Diploma of Higher Education), PGCE (Postgraduate Certificate of Education) or MClinRes (Master of Clinical Research).

6) HIGHEST QUALIFICATION ON ENTRY

''First degree'' as a highest qualification on entry was found to be an important predictor of the standard occupational classification of students with a disability. A 'first degree' is more commonly known as a bachelor's degree. Officially this includes first degrees (including eligibility to register to practice with a health or social care or veterinary statutory regulatory body), first degrees with Qualified Teacher Status (QTS)/registration with a General Teaching Council (GTC), postgraduate bachelor's degree at level H, enhanced first degrees (including those leading towards obtaining eligibility to register to practice with a health or social care or veterinary statutory regulatory body), first degrees obtained concurrently with a diploma, and intercalated first degrees.

7) JACS CODE

The JACS code that came up from the feature selection analysis was grouped under the ''X'' category from the dataset. The X category, which is Education, can be subdivided as follows 1) Broadly-based programmes within education, 2) Training teachers, 3) Research & study skills in education, 4) Academic studies in education and 5) Others in education, according to HESA classification. These were

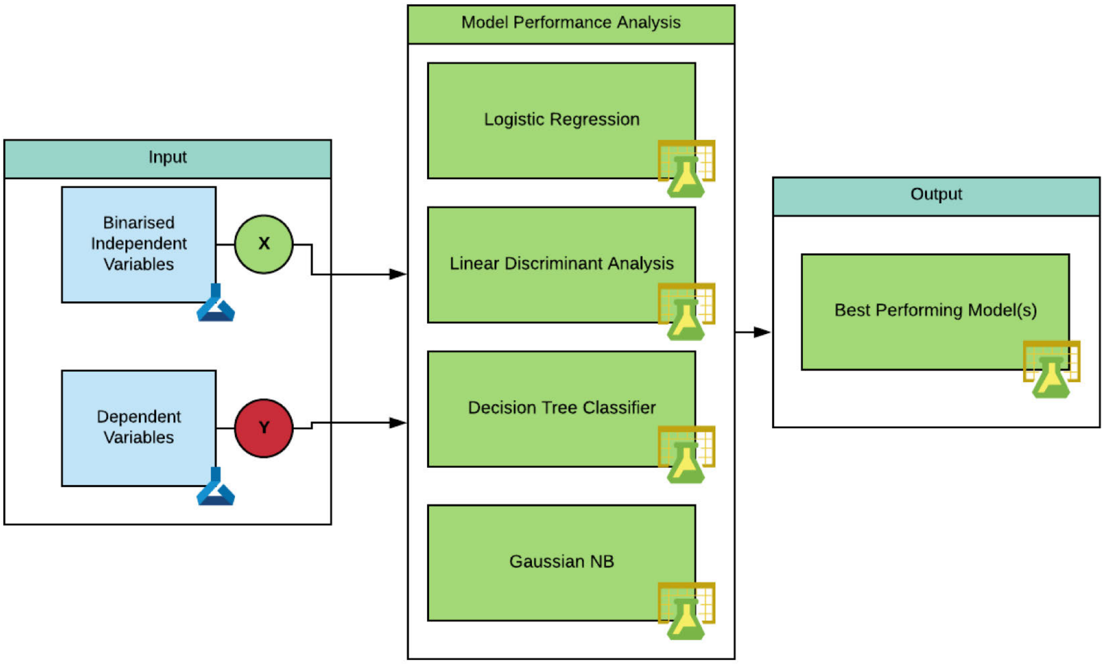


FIGURE 7. Process for the model performance analysis.

picked by the ML algorithms, which shows that it can play an important role as a predictor of a UK disabled student's engagement.

B. JOB TYPES

The data was analysed in terms of most common and least common jobs secured by UK DLHE leavers about 6 months after graduation. According to the dataset, most common jobs were that of healthcare, business and public service professionals as well as teaching and educational professionals. Whereas, the least common jobs were in the skilled agricultural, construction and building trades. The different jobs secured by UK disabled HE leavers 2012-2017 by gender is shown in Table 5.

V. DISCUSSION

A. MODELLING WITH SELECTED FEATURES

In this section, we describe a pilot study that has been developed to evaluate the feature selection of students' engagement post higher education. For this study, only features found to be common in at least two out of the three methods, as explained earlier, were considered for the classification model. Different machine learning tests such as logistic regression, linear discriminant analysis, decision tree classifier etc. had to be performed to choose which one will be best suited for the selected data. The process for the model performance analysis is shown in Figure 7.

These models were applied to selected features to get an insight into ML algorithms' predictive capability. Further research needs to be done to ensure there was no overfitting of the data, and a new dataset needs to be tested. The next section discusses our study findings.

TABLE 6. Performance of different ML methods with SOC as target variable.

ML Model SOC	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.96	0.96	0.96	0.96
Linear Discriminant Analysis	0.95	0.93	0.93	0.92
Decision Tree Classifer	0.96	0.96	0.96	0.96
Gaussian NB	0.95	0.93	0.94	0.93

B. CLASSIFICATION

Different ML methods were used to compare the results in terms of accuracy of the model to predict the 1) Activity, 2) Standard Industrial Classification (Industry) and, 3) Standard Occupational Classification SOC (Job) of a DLHE leaver about 6 months after graduation. Results show that the current datasets with the selected features used by ML algorithms such as Logistic Regression, Linear Discriminant Analysis Decision Tree and, Gaussian NB performed the least for the Activity and Industry as the target variable. However, the Decision Tree Classifier and Logistic Regression models provided the best results for predicting the Standard Occupation Classification (SOC) of a disabled school leaver in the UK with an accuracy of 96%, as shown in Table 6.

VI. CONCLUSION AND FUTURE WORK

This study identifies and discusses features selected from the dataset of 270,934 student records that could be used to build a predictive model for classifying the Standard Occupation Classification of UK disabled students and their engagement about 6 months after graduation. This data consists of UK Destinations Leavers from Higher Education (DLHE) survey results, which provide anonymised information about students' age range, year of study, disability type, and results of the first degree from 2012 to 2017, among others. To the authors' knowledge, no similar studies have explored feature selection for UK disabled students' engagement post higher education. Features such as age, institution, disability type, among others, were found to be important predictors. 10 out of 13 (77%) universities selected through the feature engineering process are from the Russel Group. It was also interesting to see that the ''Unclassified'' class of first degree, which operates on a pass/fail basis for courses such as PGCE or MClinRes, was picked up by ML algorithms during the feature selection process. The feature selection algorithms also selected the specific learning difficulty such as dyslexia, dyspraxia, or ADHD as important features for a predictive model. The selected features were then further quickly tested on four different ML methods to compare the results in terms of accuracy. Results show that the current datasets with the selected features used by ML algorithms such as Logistic Regression, Linear Discriminant Analysis Decision Tree, and Gaussian NB performed the least for the Activity and Industry as target variables. However, the Decision Tree Classifier and Logistic Regression models provided the best results for predicting the Standard Occupation Classification (SOC) of a disabled school leaver in the UK with an accuracy of 96%. Further research needs to be carried out on new datasets using neural networks and deep learning to improve the model and ensure that there is no overfitting of the data.

REFERENCES

[1] H. Kryszewska, ''Teaching students with special needs in inclusive classrooms special educational needs, *Oxford Univ. Press, ELT J.*, vol. 71, no. 4, pp. 525–528, Oct. 2017, doi: 10.1093/elt/ccx042.

[2] Disability Rights UK. (2017). *Careers Guidance and Advice for Disabled Young People*. Accessed: Aug. 23, 2018. [Online]. Available: https://www.disabilityrightsuk.org/sites/default/files/pdf/CareersGuidanceAndAdviceForDisabledYoungPeople.pdf

[3] B. Baumberg, M. Jones, and V. Wass, ''Disability prevalence and disability-related employment gaps in the UK 1998–2012: Different trends in different surveys?'' *Social Sci. Med.*, vol. 141, pp. 72–81, Sep. 2015.

[4] A. Powell. (2020). *People With Disabilities in Employment*. Accessed: Jan. 27, 2020. [Online]. Available: https://file:///Users/sobnathd/Downloads/CBP-7540(2).pdf

[5] Department for Work and Pensions. (2020). *Employing Disabled People and People With Health Conditions*. Accessed: Jun. 24, 2020. [Online]. Available: https://www.gov.uk/government/publications/employing-disabled-people-and-people-with-health-conditions/employing-disabled-people-and-people-with-health-conditions

[6] D. Sobnath *et al.*, ''Using machine learning advances to unravel patterns in subject areas and performances of university students with special educational needs and disabilities (MALSEND): A conceptual approach,'' in *Proc. 4th Int. Congr. Inf. Commun. Technol.* Singapore: Springer, 2020.

[7] M. I. Jordan and T. M. Mitchell, ''Machine learning: Trends, perspectives, and prospects,'' *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015.

[8] S. Ramachandra, G. S Murthy, B. Shamanna, K. Allagh, H. Pant, and N. John, ''Factors influencing employment and employability for persons with disability: Insights from a city in South India,'' *Indian J. Occupational Environ. Med.*, vol. 21, no. 1, p. 36, 2017.

[9] Trades Union Congress. (2019). Accessed: Jan. 27, 2020. *Disability Employment and Pay Gaps*. [Online]. Available: https://www.tuc.org.uk/sites/default/files/2019-11/Disability_gaps_2019.pdf

[10] The Guardian. (2017). *The Truth Behind Rising Disabled Employment: Cuts, Death and Zero-Hour Contracts*. Accessed: Feb. 3, 2020. [Online]. Available: https://www.theguardian.com/careers/2017/feb/23/disability-employment-gap-sanctions-cuts-and-death-after-fit-to-work-tests

[11] British Association for Supported Employment. (2015). *How To Support Young People With Special Educational Needs and Disabilities Into Work: A Short Guide for Schools, Colleges and Career Advisors*. Accessed: Mar. 12, 2019. [Online]. Available: https://www.plotr.co.uk/

[12] Disability Jobsite. (2020). *Disability Jobsite-Supporting People With Disability*. Accessed: Jan. 27, 2020. [Online]. Available: https://www.disabilityjobsite.co.uk/

[13] Evenbreak. (2020). *Evenbreak-Talent First*. Accessed: Jan, 27, 2020. [Online]. Available: https://www.evenbreak.co.uk/en

[14] Aspiritech. (2020). *About us*. Accessed: Jun. 25, 2020. [Online]. Available: https://www.aspiritech.org/about

[15] R. L. Flower, D. Hedley, J. R. Spoor, and C. Dissanayake, ''An alternative pathway to employment for autistic job-seekers: A case study of a training and assessment program targeted to autistic job candidates,'' *J. Vocational Edu. Training*, vol. 71, no. 3, pp. 407–428, Jul. 2019.

[16] C. L. Garberoglio, S. Cawthon, and M. Bond. (2016). *Deaf People and Employment in the us*. Accessed: Feb. 4, 2020. [Online]. Available: https://significantcommunity.com/wp-content/uploads/2018/12/Deaf-Employment-Report_final.pdf

[17] J. D. Southward and K. Kyzar, ''Predictors of competitive employment for students with intellectual and/or developmental disabilities,'' *Edu. Training Autism Develop. Disabilities*, vol. 52, no. 1, pp. 26–37, 2017.

[18] M. Magrin, E. Marini, and M. Nicolotti, ''Employability of disabled graduates: Resources for a sustainable employment,'' *Sustainability*, vol. 11, no. 6, p. 1542, Mar. 2019.

[19] J. Hanson, G. Codina, and S. Neary. (2017). *Transition Programmes for Young Adults With Send*. Accessed: Feb. 4, 2020. [Online]. Available: https://www.careersandenterprise.co.uk/sites/default/files/uploaded/careers-enterprise-what-works-report-transition-prog.pdf.

[20] A. M. Shahiri, W. Husain, and N. A. Rashid, ''A review on predicting Student's performance using data mining techniques,'' *Procedia Comput. Sci.*, vol. 72, pp. 414–422, Dec. 2015.

[21] Y. Park, D. G. Seo, J. Park, E. Bettini, and J. Smith, ''Predictors of job satisfaction among individuals with disabilities: An analysis of South Korea's national survey of employment for the disabled,'' *Res. Develop. Disabilities*, vols. 53–54, pp. 198–212, Jun. 2016.

[22] M. Heyman, J. E. Stokes, and G. N. Siperstein, ''Not all jobs are the same: Predictors of job quality for adults with intellectual disabilities,'' *J. Vocational Rehabil.*, vol. 44, no. 3, pp. 299–306, 2016.

[23] (2019). *Civil Service Named Top Graduate Employer-Civil Service*. Accessed: Feb. 3, 2020. [Online]. Available: https://civilservice.blog.gov.uk/2019/10/22/civil-service-fast-stream-named-top-graduate-employer/

[24] Department for Work Pensions. (2017). *Improving Lives the Future of Work, Health and Disability*. Accessed: Jan. 27, 2020. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/663399/improving-lives-the-future-of-work-health-and-disability.PDF

[25] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Newton, MA, USA: O'Reilly Media, Inc, 2018.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ''SMOTE: Synthetic minority over-sampling technique,'' *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[27] T. Hoens *et al.*, ''Building decision trees for the multi-class imbalance problem,'' in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Berlin, Germany: Springer, 2012.

[28] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, ''Threshold-based feature selection techniques for high-dimensional bioinformatics data,'' *Netw. Model. Anal. Health Informat. Bioinf.*, vol. 1, nos. 1–2, pp. 47–61, Jun. 2012.

[29] J. Brownlee, ''Why one-hot encode data in machine learning,'' *Mach. Learn. Mastery*, Jul. 2017. Accessed: Sep. 1, 2020. [Online]. Available: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

[30] H. Abbas, F. Garberson, E. Glover, and D. P. Wall, "Machine learning approach for early detection of autism by combining questionnaire and home video screening," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 8, pp. 1000–1007, Aug. 2018.

[31] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning*. Boston, MA, USA: Springer, 2012, pp. 307–323.

[32] A. Sharaff and H. Gupta, "Extra-tree classifier with metaheuristics approach for email classification," in *Advances in Computer Communication and Computational Sciences*. Singapore: Springer, 2019, pp. 189–197.

[33] A.-C. Haury, P. Gestraud, and J.-P. Vert, "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures," *PLoS ONE*, vol. 6, no. 12, Dec. 2011, Art. no. e28210.

[34] U. G. Okeke, D. Akdemir, I. Rabbi, P. Kulakow, and J.-L. Jannink, "Accuracies of univariate and multivariate genomic prediction models in African cassava," *Genet. Selection Evol.*, vol. 49, no. 1, p. 88, Dec. 2017.

DRISHTY SOBNATH received the Ph.D. degree in software engineering from Kingston University, London. She is currently a Lecturer in data science and artificial intelligence with Solent University, U.K. She has worked on the WELCOME EU funded project (2013–2017), which aimed at providing an integrated care platform to support COPD patients suffering from different comorbidities in five EU countries as part of her Ph.D. thesis. Her work included implementation and testing of the IoT solutions, wearables and mobile application development to improve lifestyle, and self-management of the chronic condition. She has also developed an Android application (CANAdvice) to help cancer patients undergoing chemotherapy treatment to self-manage their side effects and receive real-time advice from the platform. Her research interests include machine learning, the Internet of Things (IoT), user experience and usability testing methods, m-health, behavioral change management, and decision support systems in the health sector. Her research interests also include the design and evaluation of m-health systems for patients with chronic conditions. She is a member of the Technical Program Committee for IEEE Global Communications (Globecom 2018) and the British Computer Society. She has presented her work at several local and international conferences.

TOBIASZ KADUK received the master's degree in computer engineering from Solent University, U.K. He is currently pursuing the Ph.D. degree with the University of Osnabruck, Germany. He was a Research Assistant with Solent University. His research interests include machine learning and data analysis.

IKRAM UR REHMAN (Member, IEEE) received the Ph.D. degree in mobile healthcare (m-health) from the Faculty of Science, Engineering, and Computing, Kingston University, London. His thesis title was Medical QoS and Medical QoE for Ultrasound Video Streaming Over 4G and Beyond Small Cell Networks. He is currently the Course Director of information technology for managing business with the University of West London. His expertise lies in the investigation and development of 4G and beyond network technologies, such as LTE/LTE-advanced, small cells, 5G, the Internet of Things, and device-to-device communication. In addition, the exploitation of artificial intelligence and machine learning algorithms to assess the influence of QoS on QoE for VoIP and video traffics. His research work provides integration between emerging wireless communications and computing technologies for multi-disciplinary applications (e.g., telemedicine). He has been involved in a number of m-health and tele-health projects in the areas of medical video streaming, chronic disease management, decision support systems, behavioral change management, social robotics, tele-health and tele-care, and evaluation of tele-health services in U.K. healthcare environment. He has built a strong academic and research portfolio through publishing state-of-the-art research outcomes in high-quality IEEE conferences, Elsevier Journals, and Springer book. He also serves as a Technical Programme Committee Member of several publications, and an Invited Speaker in a number of conferences, workshops, and seminars.

OLUFEMI ISIAQ received the Ph.D. degree in artificial intelligence and semantic web modeling from Nottingham Trent University. He is currently a Senior Lecturer with the School of Media Arts and Technology, involves conducting up-to-date research and delivering of state of the art and research-informed lectures and seminars across a variety of computing courses. As a Senior Fellow of the Higher Education Academy, HEA, he has a proven record of helping students have enjoyable experience throughout their learning journey. Prior his lecturing role, he had acquired a vast industrial experience as a Software Engineer in multinational organization, such as Eon, U.K., and Siemens Enterprise and Communication Ltd., where he had worked on various clients' projects, including software systems development for companies such as Rolls Royce and Bosch among others. The thirst of seeking, and dissemination of knowledge particularly, new knowledge led him into academics. Subsequently, he took up a teaching role with the Nottingham Trent International College, where he had helped numerous students on university progression especially, students of international backgrounds. Proceeding as a Lecturer/Senior Lecturer with Nottingham Trent University provided him ample opportunities to teach, mentor, create curriculum, and pedagogical activities for students and supervised research projects among other responsibilities. He is currently a Senior Lecturer with Solent University, where he carries out academic responsibilities, including design and delivery of research-informed teaching, supervision and mentoring of doctoral and postdoctoral candidates on research projects, and knowledge exchange activities amongst other things. His research interest in the application of artificial intelligence tools and techniques stood on a tripod of education, health and wellbeing, and business. His research interests include decision support systems, technology application for mental health protocols, educational support and intervention techniques, and business processes modeling and improvement among others. He is a Chartered Engineer, a member of the British Computer Society, a Senior Fellow of HEA, and a Reviewer of the ELITE-Standards Fellowship among other honors.