BENNETT UNIVERSITY
TIMES OF INDIA GROUP

Foreign, Commonwealth
& Development Office
UK Government

N·S·D·C
National
Skill Development
Corporation
Transforming the skill landscape

Predictive analytics in skill development – Leveraging Artificial Intelligence (AI) and Machine Learning (ML) techniques

Final report

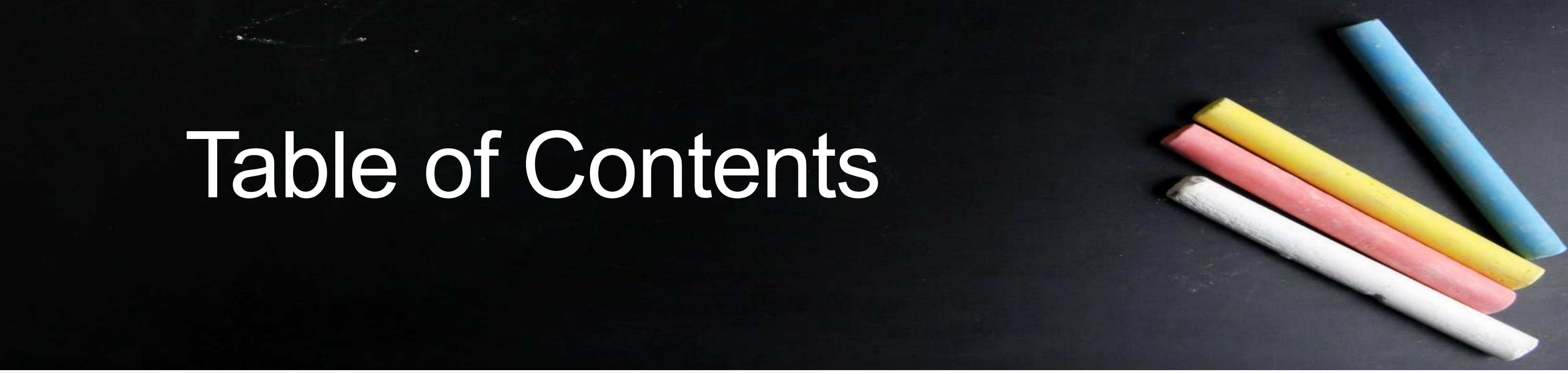


Table of Contents

List of Figures

Introduction to Bennett University

Extending the core journalistic principles of Trust, Knowledge and Public Service, Bennett, Coleman and Co. Ltd. (BCCL) established over 178 years ago, has continually undertaken initiatives for the betterment of Indian society. The group started expanding into the education sector with the launch of TimesPro - which focuses on short-term courses that enhance employability. Bennett University, a state private university in Uttar Pradesh with the aim of providing Ivy League quality of education to undergraduate and postgraduate students making them, 'life and career ready'. Bennett University located at Greater Noida, commenced its operations on 08 Aug 2016. The University fosters a proactive environment of Innovation and Entrepreneurship, while enhancing skills in all areas of higher education through the internationally acclaimed Centers of Excellence such as Centre for Innovation and Entrepreneurship (CIE) and Centre of Executive Education (CEE).

Deepak Garg
Director – LeadingIndia.AI
Head – Computer Science Engineering Department
Bennett University, UP
Email- Deepak.garg@bennett.edu.in

Vipul Kumar Mishra
Assistant Professor
Computer Science and Engineering Department
Bennett University
Email- Vipul.mishra@bennett.edu.in

Acknowledgements

The findings and recommendations of the study were derived from the insights of the data sets which has been collected across different variables which can potentially assist in creating more informed policy decisions. We would like to express our sincere thanks to all the individuals for their valuable insights and support provided for conducting this study. Special thanks and gratitude to all other individuals from Bennett University, FCDO, NSDC who supported directly or indirectly in completion of this report. NSDC also provided us inputs and feedback during the course of the project. In addition, we convey our gratitude to all those who have in some way or the other, contributed towards the successful completion of this report.

Note to Readers and Disclaimer

1. The study was commissioned by FCDO in partnership with the Bennett University. The purpose of this study is to help NSDC and other relevant stakeholders in the skilling ecosystem to get latest relevant evidence on analytics and machine learning techniques. It is not intended to be a comprehensive summary of evidence. The contents do not constitute professional advice on behalf of the UK's FCDO or Bennett University

2. The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. The report shall not be a substitute for any due diligence to be carried out by any party. No one should act on such information without appropriate professional advice after a thorough examination of the particular situation.

3. While information obtained from the public domain or external sources has not been verified for authenticity, accuracy or completeness, Bennett University have obtained information, as far as possible, from sources generally considered to be reliable. Bennett University assume no responsibility for such information.

4. In connection with the report or any part thereof, FCDO or Bennett University does not owe duty of care (whether in contract or in tort or under statute or otherwise) to any person or party to whom the Report is circulated to and FCDO or Bennett University shall not be liable to any party who uses or relies on this Report. FCDO and Bennett University thus disclaim all responsibility or liability for any costs, damages, losses, liabilities, expenses incurred by such third party arising out of or in connection with the Report or any part thereof.

5. By reading/ viewing the report, the reader of the report shall be deemed to have accepted the terms mentioned hereinabove.

6. The study was completed in July 2020 and is based on the data shared by NSDC and has not taken into account the developments subsequent to the completion of the study

Glossary

AI	Artificial Intelligence
CART	Classification and Regression Trees
FCDO	Foreign, Commonwealth and Development Office
MIS	Management Information System
ML	Machine Learning
MSDE	Ministry of Skill Development and Entrepreneurship
NSDC	National Skill Development Corporation
PMKVY	Pradhan Mantri Kaushal Vikas Yojana
R and D	Research and Development
TC	Training Centre
TP	Training Partner
UK	United Kingdom
VET	Vocational Education and Training

Executive summary

With a focused skill development policy spanning across both fresh skilling and certification of prior knowledge together with a structured vocational education and skill development set up; India has made some commendable strides. The inception and reach of the world's largest skill development scheme-Pradhan Mantri Kaushal Vikas Yojana (PMKVY) translated as the Prime Minister's initiative on Skill Development implemented under the aegis of the Ministry of Skill Development and Entrepreneurship (MSDE) and National Skill Development Corporation (NSDC) has amplified the reach and acceptance across divergent groups of stakeholders. With close to five years of its operations, PMKVY and other skilling projects anchored by NSDC have generated multiple data points about the varied stakeholders in the training value chain including the demographics and psychographics.

One of the key focus areas of NSDC is to foster private sector participation in the training and placements of the trained candidates. Given the importance of placements and industry tie ups in the skill development space, it is imperative to understand the linkages that placement draws upon the other variables of the training value chain. Through the data provided by NSDC of 1.5 lakh candidates, Bennett University created a predictive model using Artificial Intelligence (AI) and Machine Learning (ML) techniques.

The predictive model created using the tree-based methodology sought to understand some of the factors which are influencing placements. The model achieves close to 70 percent accuracy in the placement prediction. An attribute analysis of the multiple variables and their influence on placements was calculated. The analysis was imperative to understand the performance of the existing training centres, sectors, and impact of geographic location on the placement of a candidate. The placement index is not only a marker of a training centre's quality of training but also helps us understand industry demand and relevance of a job role. The analysis revealed the importance of variables such as age group, name of particular training center (TC), training partner (TP) and job role details with regard to placements.

A subsequent suggestive modelling was also created to assist training centres and potential candidates on choosing best fit sector and job role in terms of post training placements. The analysis on district, training partner, sector, and job role can help a potential candidate to select the most appropriate job role in terms of placement probability specific to educational background, age group, geographical location and training centre. The NSDC data sets pointed out that sectors like "Apparel", "Electronic and Hardware", and "Healthcare" are best performing sectors in terms of placement with "self-employed tailor", "retail sales associates", "documentation assistant" and "customer care executive" as the preferred job roles. Going forward this model can be explored to aid in data driven policy design and implementation together with assisting candidates to make informed decisions and increase industry relevance and support in the skill development domain.

Context setting

Over the years, the technical training and vocational education landscape in India has been evolving with the interfacing of the new age technologies and skills being introduced in the ecosystem. Cusped with a demographic dividend and with a fast-paced economy, the country has a colossal responsibility of channelizing the youth potential in economically productive ways. Hence a trained workforce is a quintessential force for fueling the industrial growth engine.

MSDE through the National Skill Development Corporation, a public-private partnership (PPP) entity has been catalyzing several short-term trainings through a plethora of schemes, the largest being the Pradhan Mantri Kaushal Vikas Yojana (PMKVY) by fostering several private entities. Launched in 2015 with an aim to skill close to one crore youth by 2020, the focus of PMKVY is now on Industry 4.0 and new age digital skills. Over the two phases of PMKVY, NSDC and other stakeholders of the VET ecosystem have amassed an enormous data of trainees and trainings.

With the proliferation of Big Data and other relevant technologies, a bevy of insights can be garnered through large data sets. Vast amounts of digital data have been captured through the training cycle especially for the short-term training. As on August 2020, close to 34 lakh candidates have received fresh skill trainings, 33 lakh candidates have enrolled for being certified for their prior experience and nearly 1.5 lakh candidates have been enrolled under special projects[1]. Data related to their demographics, job profiles and trainings have been recorded by the private training entities on the centralized portal namely the Skill Development Management System (SDMS) created by NSDC.

Going forward, the volume, velocity and variety of the data will continue to increase. Therefore, the endeavour now is to appropriately visualize and interpret the data for creating insights and opportunities in further embellishing the scheme and the larger VET ecosystem. One of the major goals of Big Data and related technologies is to create knowledge. The sources, methods and incumbent analysis from Big Data methods supplement and enrich established statistics. These analytics allow a detailed view at a granular level with space-related insights in real time, as well as predictive analysis.

Machine Learning and Artificial Intelligence methods are employed for undertaking an analysis of the multivariate data sets generated through the training lifecycle. This data has been leveraged to understand how the impact of various attributes are on the placement prospects of a potential training. Using a range of AI/ML based methodologies, a model was created for gathering insights from the data and their underlying meaning for drawing the relevant information which can potentially not only inform and strengthen the policy and budgetary prerogatives of the government but also assist in understanding the aptitude and guide the career aspirations of the VET trainees.

[1] https://pmkvyofficial.org/Dashboard.aspx

Overview of problem statement

NSDC wanted to understand the varied correlation of the different verticals of data collected during the training value chain and its impact on post training placements. The problem statement therefore was to develop a 'suggestive model' using predictive analytics. This model is envisioned to assist future vocational education and training aspirants in opting for market relevant job roles, thus enhancing their employment prospects and supplying skilled workforce mapped to industry needs.

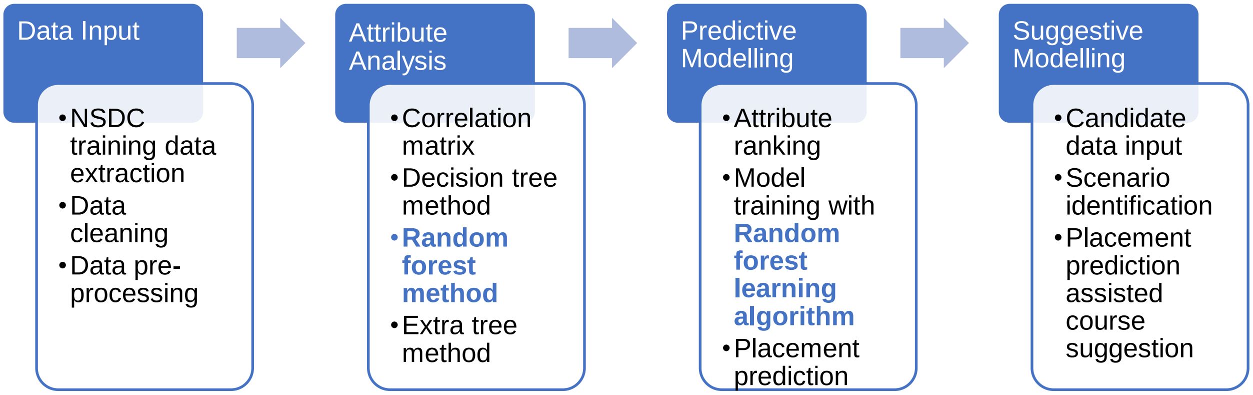
To create this model, NSDC provided information of close to 1,50,000 enrolled candidates on 40+ specific attributes including age, gender, education, location, training partner, job role, placements, results etc.

Approach and detailed methodology

Given the problem statement of finding correlation of placement with the details around training centre, job role, sector, to name a few, a three-step process was identified. A detailed overview of each of the steps have been delineated below:

1. Data structuring:

1.1. *Data cleaning*- To streamline and standardize the multivariate data set provided by NSDC, a data cleaning endeavour was performed. This process involved two parts to it; at first, the attributes serving only as identifiers such training centre (Center), candidate id (Cand ID), smart centre id (Smart Centre), training batch id (Batch ID), name of training batch (Batch Name), date of beginning the training (Batch Start Date), date of closing the training (Batch End Date), date of birth of candidates (DoB), pin code of the training centre (Pin code) have been removed. Secondly, all the attributes that could be derived from the attributes which are due for analysis and represent duplicate information like candidate’s birth year (Year of Birth), domicile details (Candidate Constituency, Sub District) id given on the job role and the sector it belongs to (Job Role ID, Sector ID), details about the id of the training partner, their centre’s constituency (Partner ID, TC constituency) have also been excluded from the further analysis.


Data Input
- NSDC training data extraction
- Data cleaning
- Data pre-processing

Attribute Analysis
- Correlation matrix
- Decision tree method
- **Random forest method**
- Extra tree method

Predictive Modelling
- Attribute ranking
- Model training with **Random forest learning algorithm**
- Placement prediction

Suggestive Modelling
- Candidate data input
- Scenario identification
- Placement prediction assisted course suggestion

Figure 1: Overview of the analysis

The tables mentioned gives an overview of the attributes considered and those which have been excluded for the analysis

Table 1: Attributes used for analysis

Attribute name	Unique entry
Gender	2
Type of disability	13
Caste category	4
Religion	9
Candidate state	35
Candidate district	687
Job role	240
Sector name	35
Partner name	2415
TC name	7813
Total candidates in batch	22
Training status	2

Table 2: Attributes not used for analysis

Attribute name	Unique entry
Centre cand id	150000
Date of birth	9420
Year of birth	58
Education attained	12
Candidate subdistrict	11791
Candidate constituency	538
Pin code:	14913
Job role id	254
Sector id	35
Partner id	2432
Tc state	35
Tc district	675

Attribute name	Unique entry
Drop out reason	8
Result	4
Grade	5
Certified	2
Age group	7
Education level	6
Placed	

Attribute name	Unique entry
TC constituency	533
Smart centre batch id	86091
Batch name	86091
Batch start date	1
Batch end date	1
Certificate generate date	34816
Salary per month	2334
Cand id	68814
Employer type	4
Age	62
Inc group	2
Emp_status	2

1.2. Data pre-processing: Post the data cleaning exercise, pre-processing of the data was completed. During the previous step it was observed that all the attributes considered for the analysis excluding ‘Total Candidates in Batch’ are categorical in nature with data being in the text format. This step is a pre-requisite for conducting the analysis, for it converts the categorical data in text format to categorical data in numerical format. This was done by assigning a unique number to every unique entry of an attribute which is an essential step for any predictive analysis. For example, the attribute ‘Caste Category’ has four unique data values “Gen, OBC, SC and ST”. All the values of Gen are replaced with 1 and similarly OBC with 2, SC with 3, ST with 4. After this the entire data in the attribute ‘Caste Category’ would contain the values 1, 2, 3, 4 which are categorical data in numerical format. Similar conversion technique was applied to all the other attributes which were available in the text format.

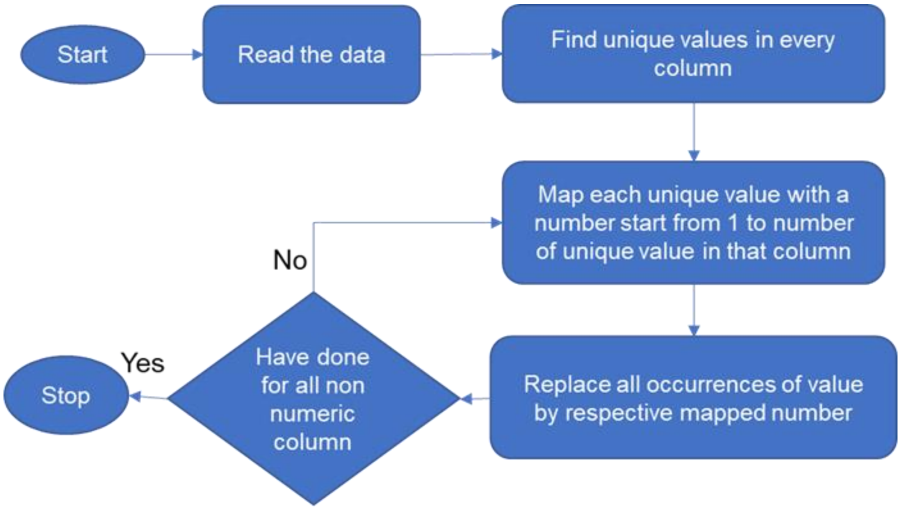

Start
Read the data
Find unique values in every column
Map each unique value with a number start from 1 to number of unique value in that column
No
Yes
Stop
Have done for all non numeric column
Replace all occurrences of value by respective mapped number

Figure 2: Pre-processing of non-numeric (text) data

Data processing allowed one to not only separate out attributes which are not required but also ensured that the non-numeric details can be given a numeric value essential for analysis. Categories such as caste which had numerical values such as 'Gen', 'OBC' where given specific numbers such as '1' and '2'

2. Attribute analysis:

After the step entailing the pre-processing of the data, an 'attribute analysis' was conducted to understand the impact of selected attributes on the placement of the candidates considering two scenarios. The first scenario looked at when a candidate is enrolled in a training centre, and the second one entailed a scenario wherein training is completed, and the assessment results of the trained candidates are available.

In the preliminary stage of the attribute analysis, a correlation matrix was created to examine the linear relation among the various attributes.

2.1. <u>Correlation matrix</u>

A correlation matrix values range from [-1, 1]. In Pearson correlation, an absolute value of 1 indicates a perfect linear relationship between the variables. A value close to 0 indicates no linear relationship between the variables. Higher absolute value indicates strong correlation and lower value represents a weak relation. In general, absolute correlation value should be greater than 0.3 to consider it as a strong relation. The sign of value represents the type of proportionality. Positive value indicates a directly correlation and a negative value indicates inverse correlation. In direct correlation, value of one variable increases with the other and in inverse correlation, value of the variable decreases with increase in the other.

From the figures 3 and 4 mentioned below, a clear inference is drawn that most of the attributes have a weak relation with the other attributes. In fact, the attribute 'placement' does not have any strong linear relation with any other attributes, with all the values being less than 0.3.

2.2. <u>Tree Methodology</u>

Based on the clear intuition from correlation matrix, there is almost no linear relationship between placement with the other attributes, a tree-based model was adopted to explore the relation between placement and other attributes. Given the problem statement and scenarios at hand, tree-based models are considered more appropriate and perform better than other models on categorical data. The three tree-based models considered for this analysis are Decision Tree, Random Forest and Extra Tree Classifier. The explanations of these models have been given in the Annexures.

The attribute analysis and predictive modelling was based on the above approaches. After applying all the three models on the data, it was observed that Random Forest Model was best suited in providing results with a relatively higher accuracy than the other two methods. In view of the same, the attribute relevance analysis in the further sections is based on Random Forest Model.

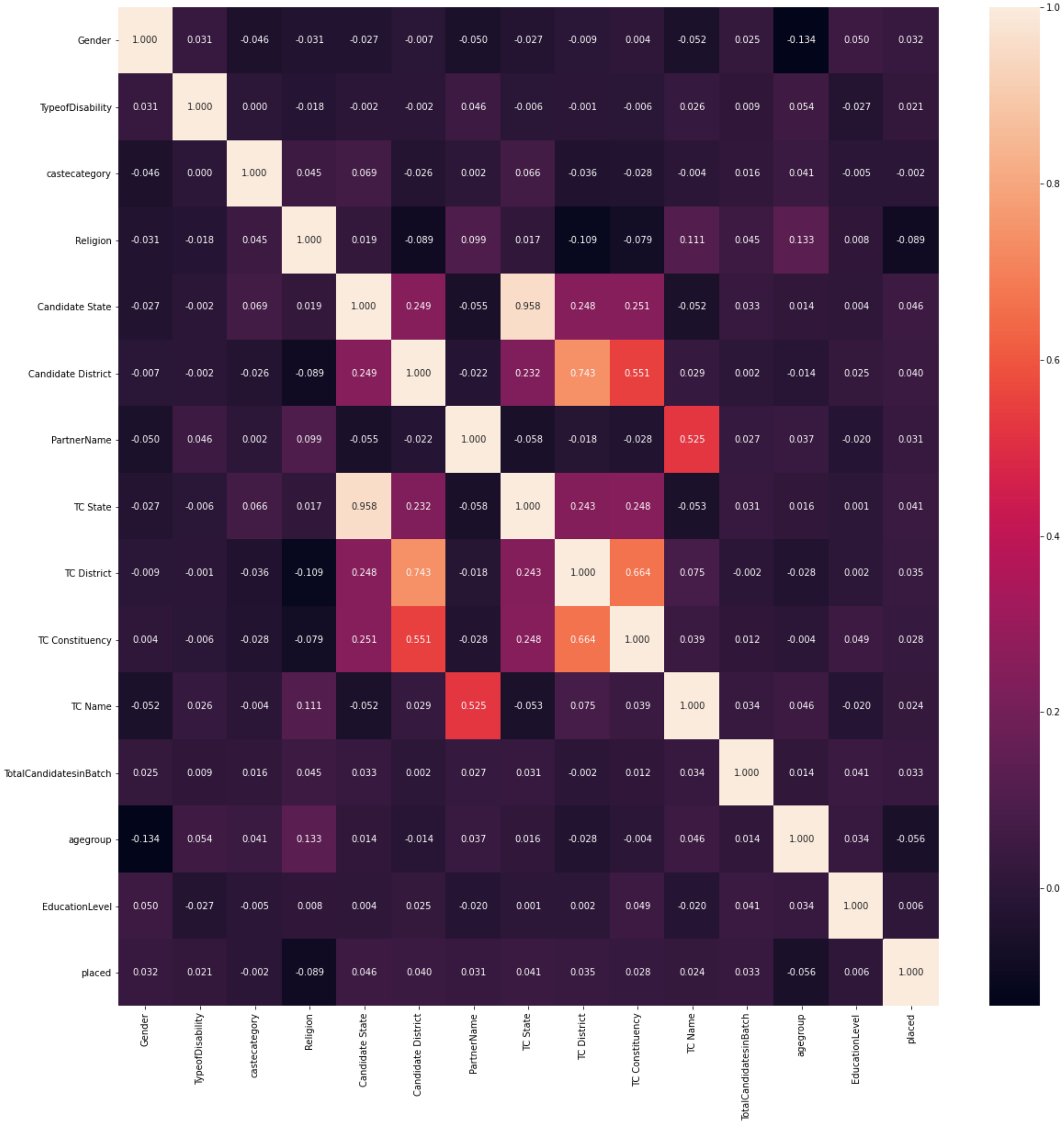


Figure 3: Correlation matrix after enrollment in course

The correlation matrix shows the linear relation between attributes. In 95 percent cases, the 'state' of the training center is the same as that of the candidate. The linear relation of 'placement' has a very weak relation with other attributes. Another key finding of this matrix shows placement has a negative relation with age group meaning that higher the age of the candidates, lesser are the chances of placement.

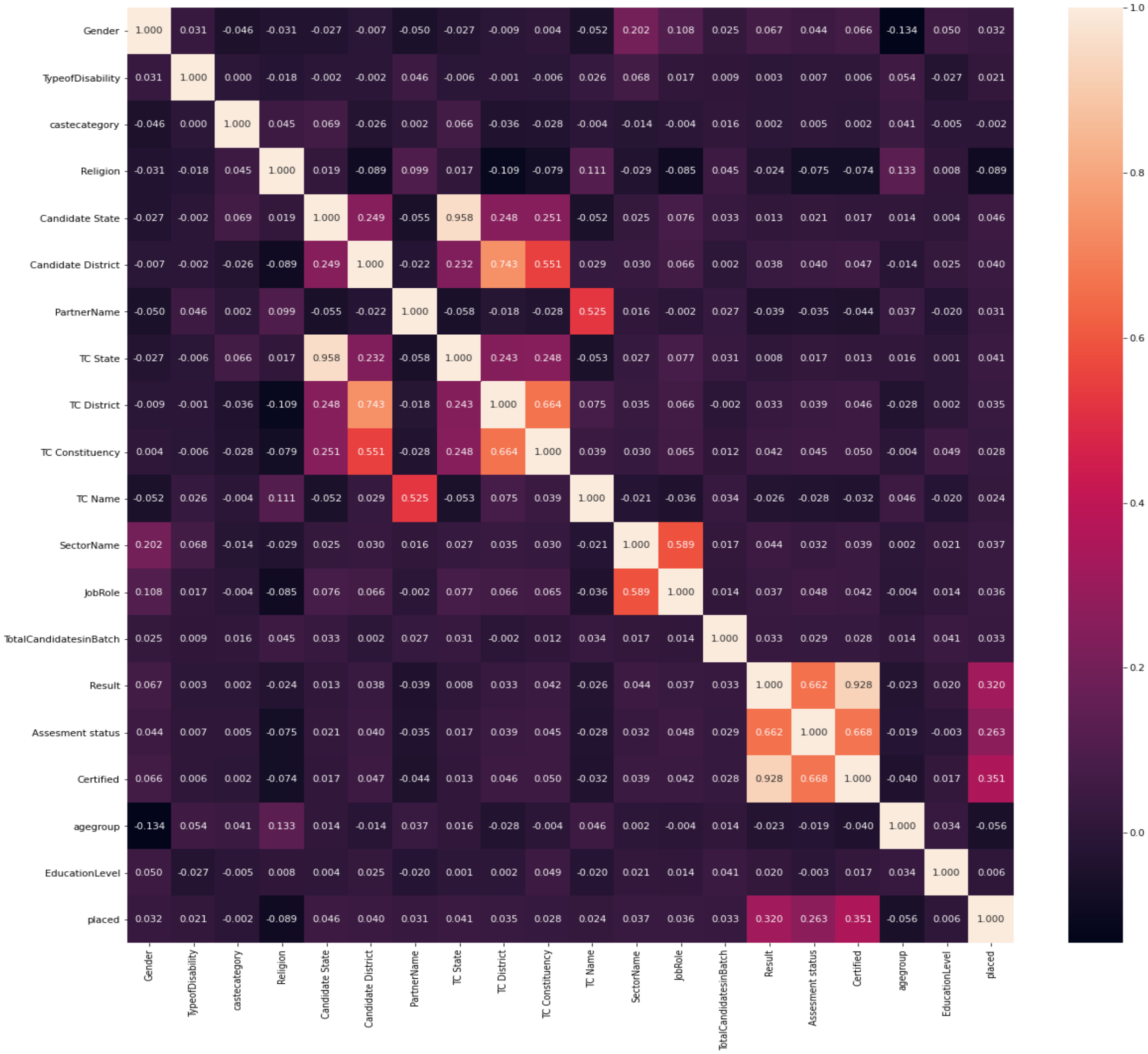


Figure 4: Correlation matrix after declaration of assessment results

In this correlation matrix, it is clearly visible that placement has a high relation with assessment results and candidate certification status. Moreover, another key finding shows that disability of a candidate has very low impact on placement.

Attribute analysis of the variables showed that an analysis of the data post enrolment in a course shows that placement has a negative relation with age group meaning that higher the age of the candidates, lesser are the chances of placement.
Moreover an analysis of the training data of candidates who have undergone assessments showed that placement has a high relation with assessment results and candidate certification status. Moreover, another key finding shows that disability of a candidate has very low impact on placement.

3. Attribute Relevance Analysis

Attribute Ranking: The importance of the attributes and their associated rankings with respect to placement are presented in the tables 3 and 4 mentioned below. It is evident from the tables that the attributes such as the age of the candidates, the name and of the training centre and training partner the district where the former is situated, together with the job role in which the candidate is undertaking training (Age Group, TC Name, Partner Name, Job Role and District) have maximum impact on placement and whereas the type of disability has least impact on placement.

As mentioned in the above cases, attribute ranking was conducted considering two scenarios, to understand the impact of various attributes on the placement of a candidate. These scenarios are at first when a candidate is enrolled in a training center, secondly post completion of training when the assessment results are available.

Table 3: Attribute ranking after enrolment

Attribute Name	Attribute Importance	Attribute Ranking
Age group	0.093482	1
TC Name	0.093174	2
Partner Name	0.089485	3
Job Role	0.074157	4
Candidate District	0.072621	5
Certified	0.066479	6
TC Constituency	0.061563	7
TC District	0.060808	8
Caste category	0.058643	9
Sector Name	0.054181	10
Education Level	0.049821	11
Result	0.043399	12
Religion	0.040059	13
Candidate State	0.031376	14
TC State	0.030837	15
Gender	0.030048	16
Total Candidates in Batch	0.027482	17
Assessment status	0.020848	18
Type of Disability	0.001537	19

Table 4: Attribute ranking after course completion

Attribute Name	Attribute Importance	Attribute Ranking
Age group	0.122636	1
TC Name	0.101461	2
Partner Name	0.097093	3
Job Role	0.082297	4
Candidate District	0.081303	5
Caste category	0.075664	6
TC Constituency	0.066421	7
TC District	0.065638	8
Education Level	0.062593	9
Sector Name	0.059727	10
Religion	0.048541	11
Gender	0.036983	12
Candidate State	0.033473	13
TC State	0.03244	14
Total Candidates in Batch	0.031912	15
Type of Disability	0.001821	16

In the attribute relevance analysis for both the scenarios namely at the time of enrolment and next at the end of the training when the assessment results, showed that attributes such as 'age', 'training centre name' were seen to be of primary importance whereas variables such as type of disability had no consequential impact on the placement potential of the candidates

4. Predictive Modelling

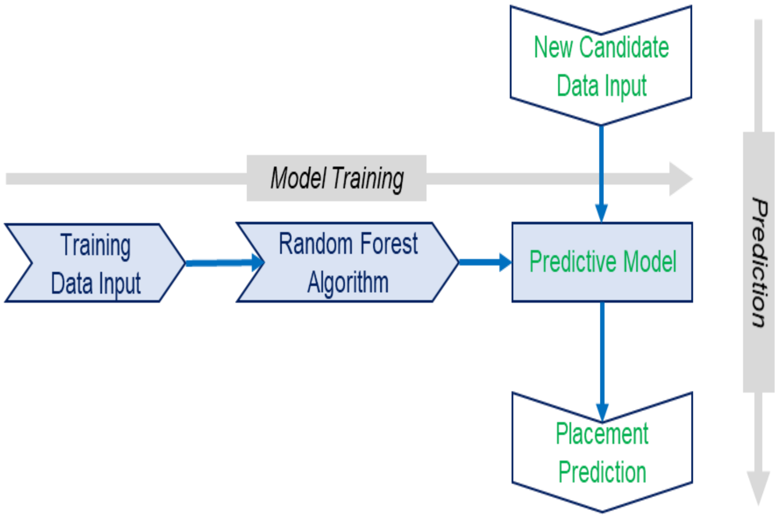
4.1. *Model Training*

Based on the attribute ranking shown in above mentioned tables (Table 3 and 4) a predictive modelling of placement was done. This modelling allowed us to explore the possibilities of placement of a skilled candidate. Going forward this analysis can help the training centres (TCs) to direct the potential candidates to specific sectors and job roles which are most suitable for them, especially in terms of placement.

For the predictive modelling the linear Regression, Artificial Neural Network, Decision Tree, Random Forest, and Extra Tree methods were used. Post explorations of the various models, the Random Forest method worked most efficiently. This method allowed a 74.4 percent accuracy for analyzing the scenario when assessment results of the candidates was available, and a 69 percent accuracy was recorded when the assessment results were not available (before enrolment in any course). In other words, one can say that at the time of enrolment a candidate's future placement can be predicted with 0.69 probability and post training and assessment completion the same can be predicted with ~0.75 probability (confidence).

For predictive modelling, the trained model was created using given data along with "Random Forest learning algorithm" is shown in figure 5.

4.2. *Prediction of placement*



Post the creation and training of the model, one has to provide the new data as input and the model will be able to foretell the placement possibilities of the candidates. This model can be leveraged at the counselling stage where possibility of placement of the potential candidate can be generated with the help of candidate profile data and the training centre data such as : Age Group, Candidate District, Caste Category, Education Level, Religion, Gender, Candidate State, Type of Disability, TC Name, Partner Name, TC Constituency, TC District, TC State, Total Candidates in Batch, Job Role, Sector Name.

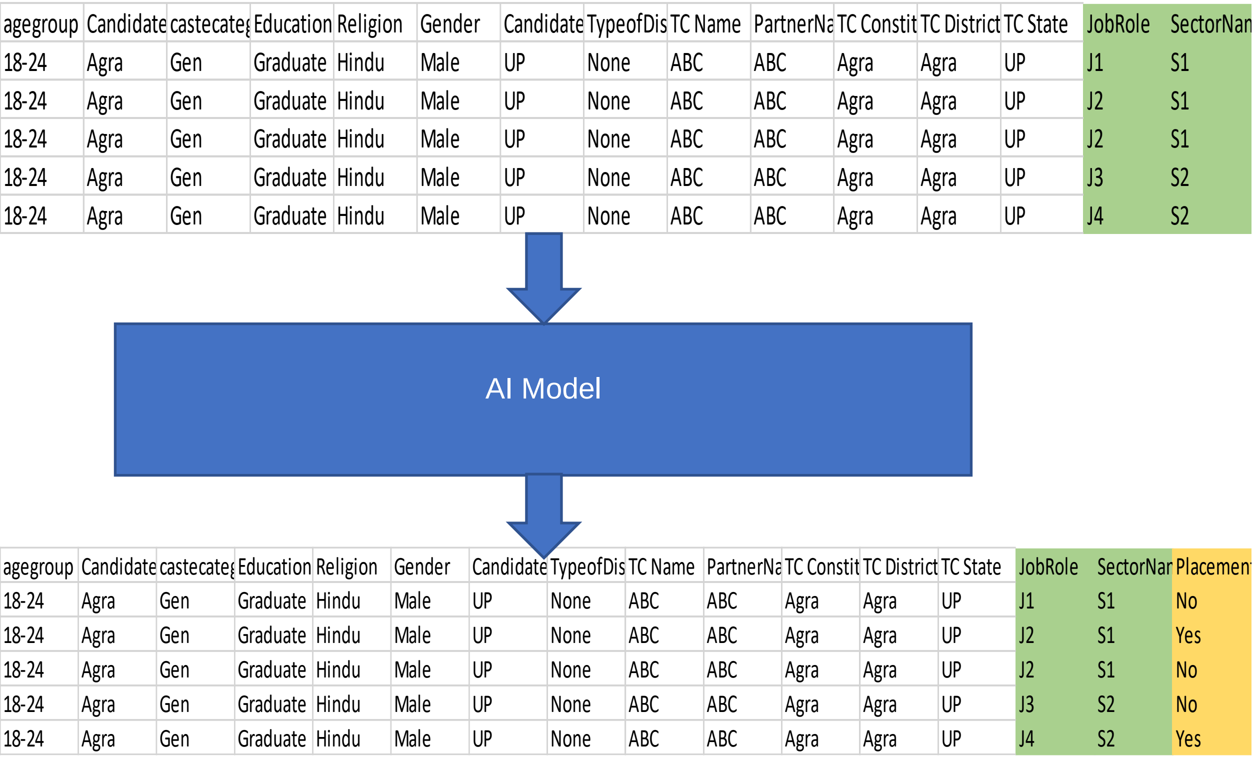
Figure 5: Predictive Modelling of placement

Using the Random Forest Method, the predictive model created allowed one to gauge the placement possibilities of the candidates with 69 percent accuracy at the time of the enrolments whereas when the assessment results were available the same could be predicted with a 75 percent probability or confidence level. The use of such model can be leveraged at the mobilization and counselling stages to help the candidates to navigate towards industry driven skills

Inferences from the analysis

Suggestive Modelling

The trained model which is described in figure 5 can also be used as suggestive modelling. For an example if a candidate in Agra went to a training centre and requested suggestion for a placement-oriented job role, the model can generate the placement probabilities of the candidate basis some generic data as mentioned previously. Hence this model can aid candidates wishing to pursue vocational education and skill development to select industry relevant courses. Similarly, if a candidate lives halfway in between Agra and Mathura, it provides him/her the option to undertake training in any of the district. Using the created model one can check the details of all training centers, job roles being taught in both the districts and exactly suggest candidates the right mix of job role, district and training centre for maximizing their placement potential. The flow diagram of the example is given in figure 6:



agegroup	Candidate	castecateg	Education	Religion	Gender	Candidate	TypeofDis	TC Name	PartnerNa	TC Constit	TC District	TC State	JobRole	SectorNan
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J1	S1
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J2	S1
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J2	S1
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J3	S2
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J4	S2

agegroup	Candidate	castecateg	Education	Religion	Gender	Candidate	TypeofDis	TC Name	PartnerNa	TC Constit	TC District	TC State	JobRole	SectorNar	Placemen
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J1	S1	No
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J2	S1	Yes
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J2	S1	No
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J3	S2	No
18-24	Agra	Gen	Graduate	Hindu	Male	UP	None	ABC	ABC	Agra	Agra	UP	J4	S2	Yes

Figure 6: An execution example of suggestive modelling

As per the figure mentioned above, the model predicted that the candidate in question should choose job role J2 and sector S1 or job role J4 and sector S2 as these entail the highest probability of placement. Similarly, the predictions can be based on district or training partner or training centre name etc.

Example of predictive analysis

This section elucidates an example to show the workings of the predictive modelling. The sample input data used for demonstration is given in Table 5. Once the data is inserted in the predictive model, then it will predict the placement (as shown in Table 6) of the candidates based on their personal data and training center profile as shown in Figure 7.

Table 5: Input data sample

Sr. No	1	2	3	4	5
Gender	Female	Female	Female	Female	Female
Type of Disability	None	None	None	None	None
Caste Category	OBC	Gen	OBC	Gen	OBC
Religion	Hinduism	Hinduism	Hinduism	Islam	Islam
Candidate State	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh
Candidate District	Aligarh	Aligarh	Aligarh	Aligarh	Aligarh
Age Group	36-40	19-21	26-30	22-25	31-35
Education Level	10th Std and below	10th Std and below	11th-12th Std	10th Std and below	11th-12th Std
Partner Name	Chanakya Education and Charitable Trust	Chanakya Education and Charitable Trust	Sunaina Samriddhi Foundation	Sunaina Samriddhi Foundation	Sunaina Samriddhi Foundation
Tc State	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh
Tc District	Aligarh	Aligarh	Aligarh	Aligarh	Aligarh
Tc Constituency	Aligarh	Aligarh	Aligarh	Aligarh	Aligarh
Tc Name	Samajik Vikas Seva Samiti	Samajik Vikas Seva Samiti	Manav Samman Seva Samiti	Manav Samman Seva Samiti	Manav Samman Seva Samiti
Sector Name	Apparel	Apparel	Apparel	Apparel	Apparel
Job Role	Self Employed Tailor	Self Employed Tailor	Hand Embroiderer	Hand Embroiderer	Self Employed Tailor
Total Candidates in Batch	30	30	30	30	30

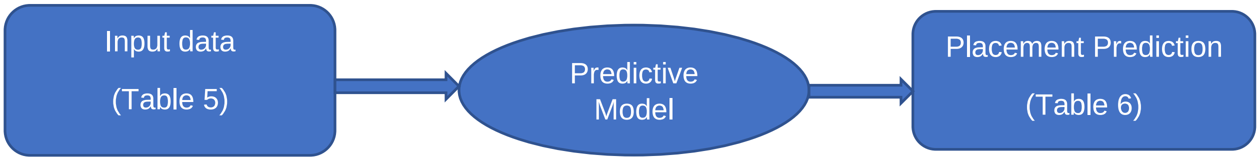


Figure 7: An execution example of predictive modelling

Table 6: Placement prediction output

Sr. No	1	2	3	4	5
Gender	Female	Female	Female	Female	Female
Type of Disability	None	None	None	None	None
Caste category	OBC	Gen	OBC	Gen	OBC
Religion	Hinduism	Hinduism	Hinduism	Islam	Islam
Candidate State	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh
Candidate District	Aligarh	Aligarh	Aligarh	Aligarh	Aligarh
Age group	36-40	19-21	26-30	22-25	31-35
Education Level	10th std and below	10th std and below	11th-12th std	10th std and below	11th-12th std
Partner Name	Chanakya Education and Charitable Trust	Chanakya Education and Charitable Trust	Sunaina Samriddhi Foundation	Sunaina Samriddhi Foundation	Sunaina Samriddhi Foundation
TC State	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh	Uttar Pradesh
TC District	Aligarh	Aligarh	Aligarh	Aligarh	Aligarh
TC Constituency	Aligarh	Aligarh	Aligarh	Aligarh	Aligarh
TC Name	Samajik Vikas Seva Samiti	Samajik Vikas Seva Samiti	Manav Samman Seva Samiti	Manav Samman Seva Samiti	Manav Samman Seva Samiti
Sector Name	Apparel	Apparel	Apparel	Apparel	Apparel
Job Role	Self Employed Tailor	Self Employed Tailor	Hand Embroiderer	Hand Embroiderer	Self Employed Tailor
Total Candidates in Batch	30	30	30	30	30
Predicted placement	**No**	**No**	**No**	**Yes**	**Yes**

The suggestive modelling allows us to understand to get a sectoral overview on a district level. This suggestive model can be leveraged for performance evaluation of job roles, districts and individual training partners in terms of their peers and related stakeholders to get a holistic understanding of the landscape

Risks and possible mitigation

The model was created based on the data shared by NSDC wherein 1.5 lakh random data points was shared out of a very large data set. Therefore, the reported model performance and confidence on it is based on the shared data sets. A training of the created model on the entire available data set would create more accurate results.

Moreover, the prediction from the model is a possibility based on the previous records and currently the model does not capture very uncertain events as well as the aptitude of a potential candidate. A more in-depth analysis would be required of ascertaining the best fitted job roles wherein their interests are documented as a variable for the model which is currently not there.

Recommendations

The AI/ML predictive model as created can be leveraged to garner deeper insights into the skill development ecosystem and the linkages with the placement structures. Based on the data sets shared by NSDC while a lot of details could be understood, to ensure greater feasibility and enhanced use of the model, the following recommendations can be looked into:

1. **Incorporation of salary details**: Given intrinsic linkage of placements with the skilling landscape, if the salary details are well captured by the Government agencies and the private training partners then the model would also be able to predict the salary scales and range of the job roles when interlinked with the district and other relevant details of the training value chain. While in schemes such as PMKVY, the salary details are captured, it is important that it is made mandatory and even updated during the placement tracking interval.
2. **Integration of additional data sets in the MIS or data templates:** When variables such as population of the district, industrial growth indices, literacy index etc. are enmeshed into the model, a more holistic understanding can be developed. Given that the skill development ecosystem and its related placements works in consonance with the larger industrial and demographic variables of the country, for a deeper policy level engagement on the model, these variables can be included in the model. Additional templates and data sets can be incorporated in the data sets and collected from the secondary sources or government officials of the district for a holistic understanding and seamless predictive model. Given that the current model allows us a close to 70 percent accuracy, a larger data set with more variables can potentially enhance the accuracy levels
3. **Frequency of running the model:** While the model is based on certain data sets which are captured during a specified period of the training value chain- namely at the time of enrolment and at the time of announcement of results, it is imperative that the model runs at regular intervals. These intervals could be pre-defined i.e. every three months to be able to add a dynamism to the model while also being able to take account of the constantly evolving economic and social prerogatives.
4. **Capacity building of skills stakeholder:** Given the primacy of data in decision making, it would be essential that skills stakeholders are able to interpret and draw insights from such analysis. To facilitate the process, continuous capacity building sessions can be curated on such tools and the interpretation of results.

Conclusion

This report documented how the AI/ML methods can be leveraged in getting a deep dive into the nuances of data generated during the training lifecycle. The analytics has the potential to map skills by occupation, identify discrepancies in skills and potentialities of becoming obsolete, predictive analysis of demand for new occupations and skills – in quasi real time.

The insights from this exercise showcased the relevance of age groups, training centre name, training partner name and districts as the four key attributes in defining the probabilities of placements. The predictive model which has been created using the available data sets has the potential of capitalizing on key insights of the training value chain. Some of the ways in which the model can be leveraged are the following:

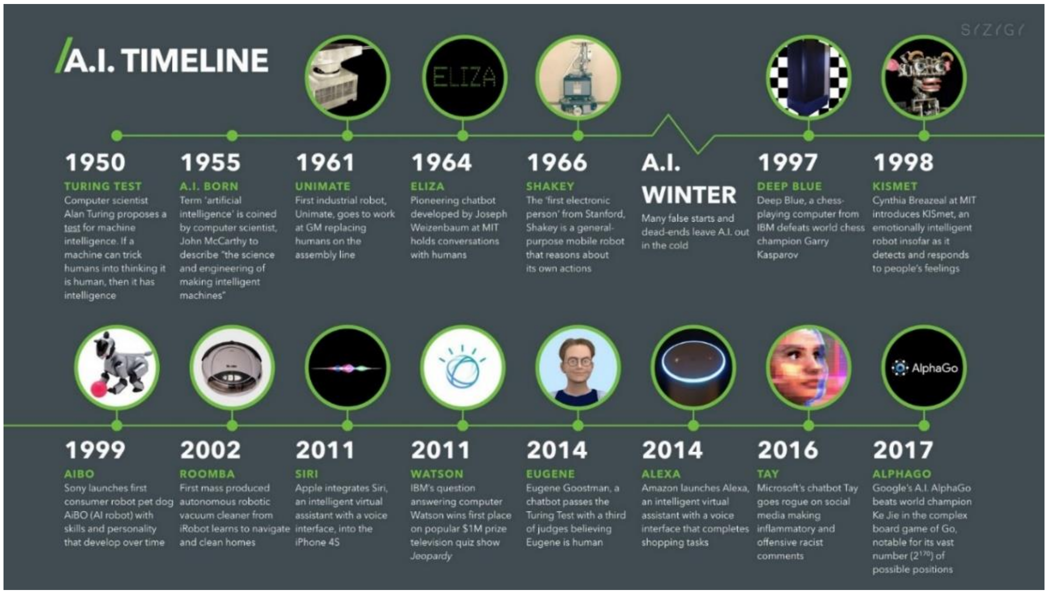
- **Career guidance support for candidates**: Given that the model can analyse the probabilities of placement linkages for a job role, it can be deployed with empaneled training centres or on the NSDC website for potential candidates to assess the industrial relevance of the training. Hence a direct linkage can be bridged on improving the employment prospects and informed decision making among the potential VET candidates.
- **Course offerings:** The model through its analysis is able to understand the market relevance of skills. Since it is imperative that for any placement linked course, its industrial worth needs to be adjudged, NSDC can leverage the model to study the placement scenario at regular interval of times and create the necessary amendments in their course offerings.
- **Policy making:** The agility of the model to map labour market intelligence would allow government officials to take more data driven decisions. Given the veracity of the results of mathematical models, the insights provided would be beneficial in analyzing micro trends which cumulatively can be affect the scheme design and implementation. Moreover, budgetary allocations for trainings would be informed by the labour market intelligence allowing an emphasis of funding support to courses with higher placement probabilities and thereby have more trained candidates to satiate the industry demands. As mentioned in the current model, variables such as age and training centre name have very strong correlation with placements, hence when scheme designs and funding is decided, it would be appropriate to take account of these variables in aligning the decisions.
- **Creation of market relevant business models:** Given the stake of the private entities in skill development, an access to these insights would be able to help training providers in investing in infrastructure for skills which are market driven and would allow them to place candidates post the training. Hence better-informed investment opportunities can be created not just for the national but also for the international entities which would want to enter the Indian VET domain in varied capacities.

Data driven insights would not just help training providers but also assessment agencies as they would be able to develop personnel and tools in the employment driven sectors and also ensure continuous review of the assessment tools to be up to date with the industry practices**.** Leveraging analytics combines a range of voluminous data sets, digital transformation, and specific computing architecture. While these techniques and the related data sources continue to evolve, their relevance in the field of skill development would further enrich the understanding of the ecosystem of not just the policy makers but also of the large number of stakeholders- both private and public in developing their operational and working models in skills which are relevant and which would be a persistent asset for the human resources to fuel the rise of a growing economy.

Annexure

1. Introduction Artificial Intelligence: Past Present and Future

Artificial intelligence (AI) is not new. The term was coined in 1956 by John McCarthy, a computer science professor from Stanford University who organized an academic conference on the topic at Dartmouth College in the summer of that year. The field of AI has gone through a series of boom-bust cycles since then, characterized by technological breakthroughs that stirred activity and excitement about the topic. As you can see in figure 8, today we in the era of an 'AI'. Artificial intelligence can be defined as human intelligence exhibited by machines; systems that approximate, mimic, replicate, automate, and eventually improve on human thinking. Throughout the past half-century a few key components of AI were established as essential: the ability to perceive, understand, learn, problem solve, and reason.



1950 TURING TEST
Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence

1955 A.I. BORN
Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"

1961 UNIMATE
First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line

1964 ELIZA
Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans

1966 SHAKEY
The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions

A.I. WINTER
Many false starts and dead-ends leave A.I. out in the cold

1997 DEEP BLUE
Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov

1998 KISMET
Cynthia Breazeal at MIT introduces KISmet, an emotionally intelligent robot insofar as it detects and responds to people's feelings

1999 AIBO
Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time

2002 ROOMBA
First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes

2011 SIRI
Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S

2011 WATSON
IBM's question answering computer Watson wins first place on popular $1M prize television quiz show *Jeopardy*

2014 EUGENE
Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human

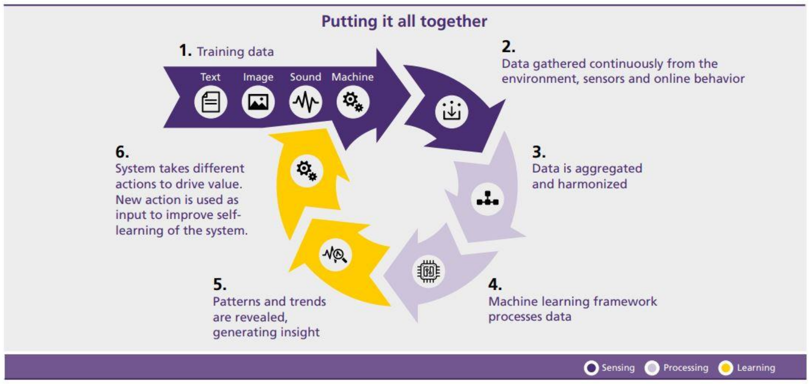
2014 ALEXA
Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks

2016 TAY
Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments

2017 ALPHAGO
Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

Figure 8: AI timeline Source: Digital wellbeing, Paul Marsden

Despite the oversimplification that tends to define AI in the popular press, AI is not one single, unified technology. AI is actually a set of interrelated technology components that can be used in a wide variety of combinations depending on the problem it addresses. Generally, AI technology consists of sensing components, processing components, and learning components as shown in figure 9.



1. Training data

2. Data gathered continuously from the environment, sensors and online behavior

3. Data is aggregated and harmonized

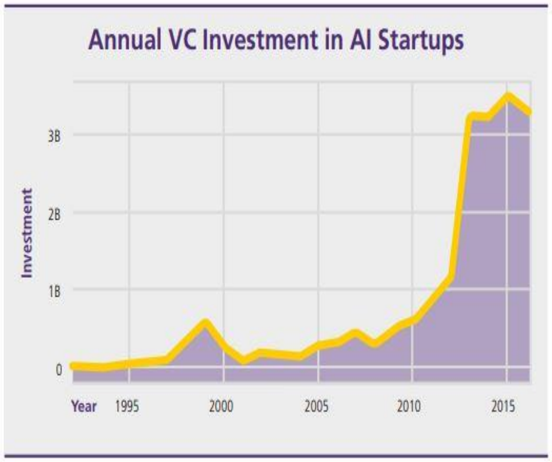
4. Machine learning framework processes data

5. Patterns and trends are revealed, generating insight

6. System takes different actions to drive value. New action is used as input to improve self-learning of the system.

Figure 9: An AI learning cycle

2. Investment and Funding for AI:



Other contributing factors to the recent surge in progress and interest in AI are the precipitous spikes in venture capital investment in AI startups and corporate funding for AI R and D and acquisitions. In 2017 alone, a group of 100 AI startups raised $11.7 billion in aggregated funding across 367 deals, contributing to a six-fold increase in investment since 2000 as depicted in figure 10. Among technology corporations, Baidu and Google specifically spent between $20-$30 billion on AI in 2016, where 90% was allocated for Rand D and deployment, and 10% for acquisitions. Finally, 9,043 U.S. patents were issued to IBM in 2017, more than 3,300 of which were related to AI or cloud technologies.

Figure 10: Investment in in US start-ups developing AI systems

The NSDC has an extensive network of skilling partners across the country. The partners receive funding and payments based on skilling targets they achieve during a point in time. One key challenge that NSDC faces is to analyse and utilize the existing data and make new plans for funding support to geographic area and sectors which are good performer in terms of placement. For this they are required to develop a system that can predict the placement. NSDC has over 20 lakhs trained individuals and over 538 training partners. To top it all, there are more than 37 Sector Skill Councils and over 10373 training centres scattered across the length and breadth of India. To analyse such enormous data and develop a model that can predict the placement of a candidate is a complex task. In addition, a suggestive model, that can guide to a new candidate for choosing a sector and job role so that the placement probability will be more after completing the training.

3. Algorithms and Methodology

3.1. Decision Tree Classifier:

Decision Tree is a type of supervised learning algorithm (having a predefined target variable) that is mostly used in classification problems. It works for both categorical and continuous, input and output variables. In this technique, the population or sample is split into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

Consider an example; There is a sample of 30 students with three variables gender (boy / girl), class (ix / x) and height (5 to 6 ft). 15 of students play cricket in leisure time. If a model is to be created for predicting who will play cricket during their leisure time, students playing cricket in their leisure time needs to be segregated based on highly significant input variable among all three as shown in figure 11.

This is where Decision Tree helps, wherein it will segregate the students based on all values of three variable and identify the variable which creates the most apt homogeneous sets of students (which are heterogeneous to each other). For this example, as shown in the image below, it can be observed that the variable 'gender' is able to identify best homogeneous sets compared to the other two variables.

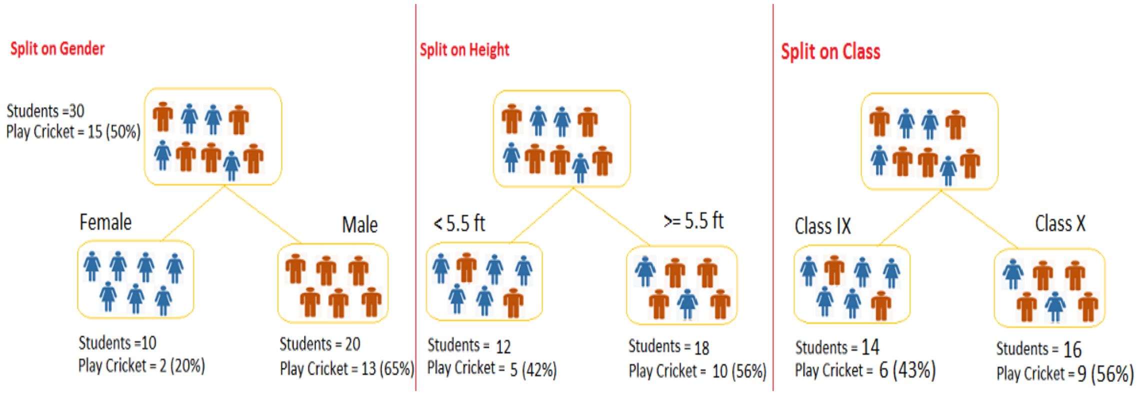


Figure 11: Example of decision tree

As mentioned above, decision tree identifies the most significant variable and its value lies in the method being able to estimate most appropriate homogeneous sets of population. For identifying the most significant variable and its value, Decision Tree employs various algorithms including Categorical Variable Decision Tree, which has categorical target variable (which is the interest of this analysis).

For the problem statement identified by NSDC, Categorical Variable Decision Tree was used. A Decision Tree which has a categorical target variable is referred to as a Categorical Variable Decision Tree. For understanding the placement scenario, the target variable was “candidate will be placed or not” i.e. YES or NO.

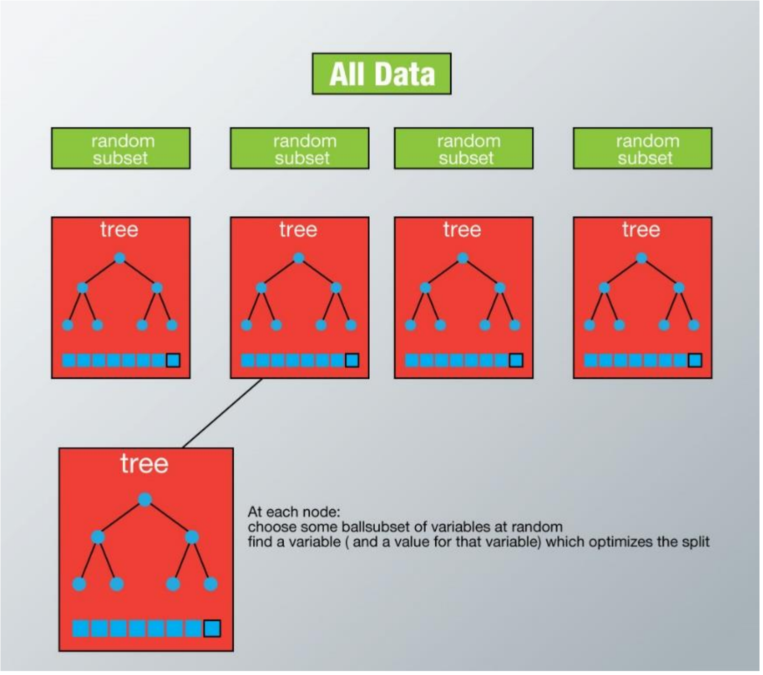
3.2. Random Forest method:

Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model. In this method, multiple trees are grown as opposed to a single tree in CART model[2]. To classify a new object based on attributes, each tree gives a classification and the tree “votes” for that class. The Forest chooses the classification having the most votes (over all the trees in the Forest) and in case of regression, it takes the average of outputs by different trees.

It works in the following manner wherein each tree is planted and grown as follows:

[2] The CART or Classification and Regression Trees algorithm is structured as a sequence of questions, the answers to which determine what the next question if any should be. The result of these questions is a tree-like structure where the ends are terminal nodes at which point there are no more questions. The main elements of CART are:

- Rules for splitting data at a node based on the value of one variable;
- Stopping rules for deciding when a branch is terminal and can be split no more;
- Finally, a prediction for the target variable in each terminal node.


All Data
random subset
random subset
random subset
random subset
tree
tree
tree
tree
tree
At each node:
choose some ballsubset of variables at random
find a variable (and a value for that variable) which optimizes the split

Let us assume the number of cases in the training set is N. Then, sample of these N cases is taken at random but with replacement. This sample will be the training set for growing the tree.

If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest.

Each tree is grown to the largest possible extent and there is no pruning. Prediction of new data is done by aggregating the predictions of the n trees (i.e., majority votes for classification)

Figure 12: Example of Random forest

3.3. Extra Tree Classifier:

Extra Trees Classifier is similar to Random Forest but with 2 key differences.

Considering a scenario wherein multiple decision trees are being built in the process, which would entail the requirement for multiple datasets. A best practice is not to train the decision trees on the complete dataset only on fraction of data (around 80 percent) for each tree. In a Random Forest, we draw observations with replacement wherein we can have repetition of observations in a random forest. While in an Extra Tree Classifier, observations are drawn without replacement, so there will not be any repetition of observations unlike in Random Forest model. The difference lies is the process of converting a non-homogeneous parent node into 2 homogeneous child nodes (best possible cases). In Random Forest, it selects the best split to convert the parent into the two most homogeneous child nodes. In an Extra Tree Classifier, it selects a random split to divide the parent node into two random child nodes. Summary of algorithms is mentioned in the below table.

Table 7: Summary of tree-based algorithms

	Decision Tree	Random Forest	Extra Trees
Number of trees	1	many	many
Number of features considered for split at each decision node	All Feature	Random subset of feature	Random subset of feature
Bootstrapping (drawing sample without replacement)	Not applied	No	Yes/No
How split is made	Best Split	Best Split	Random Split

References

[1] Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660-674.

[2] Liaw, A., and Wiener, M. (2002). Classification and regression by Random Forest. R news, 2(3), 18-22.

[3] Parvin, H., MirnabiBaboli, M., and Alinejad-Rokny, H. (2015). Proposing a classifier ensemble framework based on classifier selection and decision tree. Engineering Applications of Artificial Intelligence, 37, 34-42.

[4] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42.

[5] Decision Tree vs. Random Forest – Which Algorithm Should you Use: : https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

[6] An Intuitive Explanation of Random Forest and Extra Trees Classifiers: https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b

[7] https://nsdcindia.org/about-us

Thank you