TABLE OF CONTENTS

INTRODUCTION TO PROJECT

India has a total 6,214 Engineering and Technology Institutions in which around 2.9 million students are enrolled. Every year on an average 1.5 million students get their degree in engineering, but due to lack of skill required to perform technical jobs less than 20 percent get employment in their core domain. As per a survey, 64 percent of the employers were not completely satisfied with the quality of engineering graduates' skills. They identified integrity, reliability and teamwork as the top three most important general skills, while the top three most important specific skills were entrepreneurship, communication in English and use of modern tools and technologies. The employers were relatively satisfied with the graduates' communication skills in English, but not with their reliability (Federation of Indian Chambers of Commerce and Industry (FICCI) and the World Bank. July 2007). Aspiring minds released their data publicly for analytical purposes for the year 2016 which contains information about different aspects of a potential engineering graduate looking for employment. These factors include their logical abilities and numeracy, proficiency in English, technical knowledge in different branches of engineering and programming skills; and (qualities) such as conscientiousness, agreeableness and openness to new experiences. These data were collected for a range of engineering graduates hailing from different branches such as computer science, mechanical, electrical, civil etc. We all are aware of the job scenario and salary that an engineering student in India gets just after graduating. But we all are not aware of what are those different factors that affect the salary of Indian Engineering graduates. To address and showcase this problem of high skilled unemployment among engineering graduates from institutes all across the country, we present a statistical analysis involving various machine learning methods to predict the overall employment potential of a candidate using the aforementioned factors as a basis.

PROBLEM STATEMENT

Information asymmetry between job seekers and employers is a long-standing problem. The prospective candidates generally have less knowledge about the actual treatment during the interview and only at the final stages of the interview process they are actually informed of concrete offers. Meanwhile, it is vital for the employers to correctly guess the expectations of the candidates for crafting HR strategy; too low offers could lead to high decline rate and longer vacancies in the positions whilst, offering too much could result in high personnel expenses. Thus, it will be beneficial for most of us to know the unbiased ”market price” of the job positions, so that we can reduce mismatches and unsuccessful interviews.The project is centered around predicting a graduate’s salary. The dataset provides the information for Various factors such as college grades, candidate skills, the proximity of the college to industrial hubs, the specialization one have, market conditions for specific industries determine this. On the basis of these various factors, our objective is to determine the salary of an engineering graduate in India.

The intent of Salary Prediction would be:

- ❏ Helping to see the growth in any field.
- ❏ With the help of machine learning it can easily produce a graph.
- ❏ Marketing makes it easy to estimate the salary between the x-y axis.
- ❏ Users can give any point to get the salary through the program.
- ❏ Salary of the employees can be observed to give them a particular field according to their qualifications

DATA WAREHOUSING AND MINING TOOLS USED

1. Google Collab.
2. Orange
3. MySQL

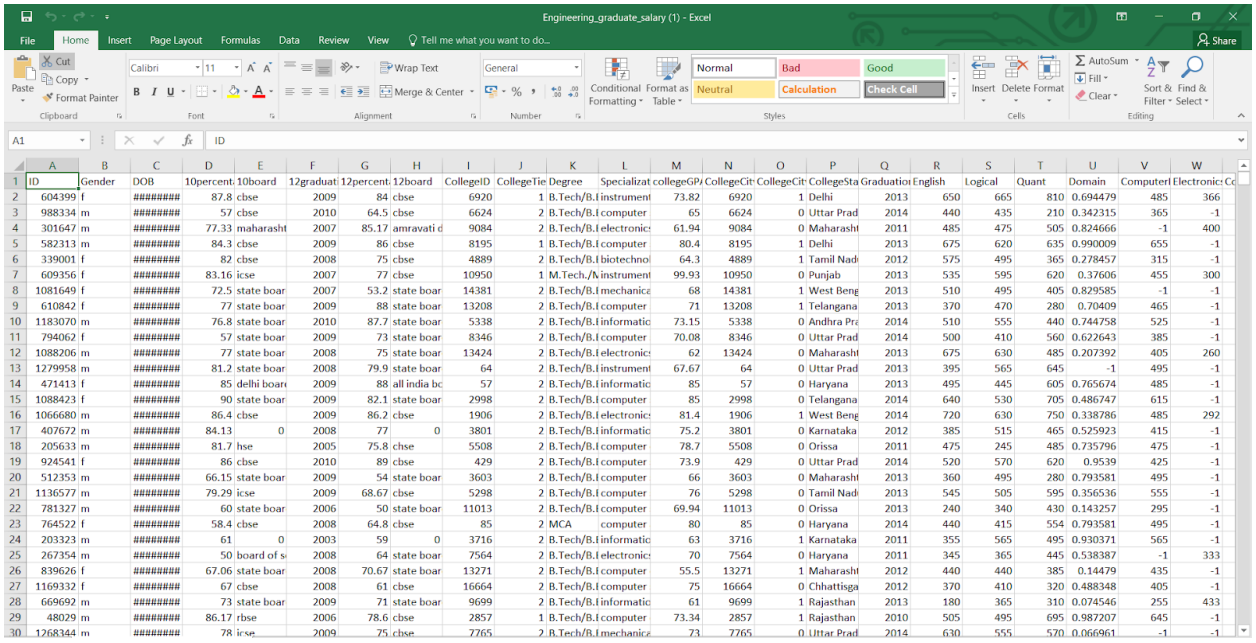
DATASET USED

DATASET USED

The dataset contains various columns:

- ID: A unique ID to identify a candidate
- Salary: Annual CTC offered to the candidate (in INR)
- Gender: Candidate's gender
- DOB: Date of birth of the candidate
- 10 percentage: Overall marks obtained in grade 10 examinations
- 10 board: The school board whose curriculum the candidate followed in grade 10
- 12 graduation: Year of graduation - senior year high school
- 12 percentage: Overall marks obtained in grade 12 examinations
- 12 board: The school board whose curriculum the candidate followed
- College ID: Unique ID identifying the university/college which the candidate attended for her/his undergraduate
- CollegeTier: Each college has been annotated as 1 or 2. The annotations have been computed from the average AMCAT scores obtained by the students in the college/university. Colleges with an average score above a threshold are tagged as 1 and others as 2.
- CollegeState: Name of the state in which the college is located
- Graduation Year: Year of graduation (Bachelor's degree)
- English: Scores in AMCAT English section
- Logical: Score in AMCAT Logical ability section
- Quant: Score in AMCAT's Quantitative ability section
- Domain: Scores in AMCAT's domain module
- Computer Programming: Score in AMCAT's Computer programming section
- Electronics And Semicon: Score in AMCAT's Electronics & Semiconductor Engineering section
- Computer Science: Score in AMCAT's Computer Science section
- Mechanical Engg: Score in AMCAT's Mechanical Engineering section
- Electrical Engg: Score in AMCAT's Electrical Engineering section
- Telecom Engg: Score in AMCAT's Telecommunication Engineering section
- Civil Engg: Score in AMCAT's Civil Engineering section
- conscientiousness: Scores in one of the sections of AMCAT's personality test
- agreeableness: Scores in one of the sections of AMCAT's personality test
- extraversion: Scores in one of the sections of AMCAT's personality test
- neuroticism: Scores in one of the sections of AMCAT's personality test
- Openness to experience: Scores in one of the sections of AMCAT's personality test

Dataset Snapshot:



Engineering_graduate_salary (1) - Excel

ID	Gender	DOB	10percent	10board	12graduat	12percent	12board	CollegeID	CollegeTie	Degree	Specializat	collegeGP	CollegeCit	CollegeCit	CollegeSta	Graduatio	English	Logical	Quant	Domain	Computerl	Electronics
604399	f	########	87.8	cbse	2009	84	cbse	6920	1	B.Tech/B.E	instrument	73.82	6920	1	Delhi	2013	650	665	810	0.694479	485	366
988334	m	########	57	cbse	2010	64.5	cbse	6624	2	B.Tech/B.E	computer	65	6624	0	Uttar Prad	2014	440	435	210	0.342315	365	-1
301647	m	########	77.33	maharasht	2007	85.17	amravati d	9084	2	B.Tech/B.E	electronics	61.94	9084	0	Maharasht	2011	485	475	505	0.824666	-1	400
582313	m	########	84.3	cbse	2009	86	cbse	8195	1	B.Tech/B.E	computer	80.4	8195	1	Delhi	2013	675	620	635	0.990009	655	-1
339001	f	########	82	cbse	2008	75	cbse	4889	2	B.Tech/B.E	biotechnol	64.3	4889	1	Tamil Nad	2012	575	495	365	0.278457	315	-1
609356	f	########	83.16	icse	2007	77	cbse	10950	1	M.Tech./M	instrument	99.93	10950	0	Punjab	2013	535	595	620	0.37606	455	300
1081649	f	########	72.5	state boar	2007	53.2	state boar	14381	2	B.Tech/B.E	mechanica	68	14381	1	West Beng	2013	510	495	405	0.829585	-1	-1
610842	f	########	77	state boar	2009	88	state boar	13208	2	B.Tech/B.E	computer	71	13208	1	Telangana	2013	370	470	280	0.70409	465	-1
1183070	m	########	76.8	state boar	2010	87.7	state boar	5338	2	B.Tech/B.E	informatic	73.15	5338	0	Andhra Pra	2014	510	555	440	0.744758	525	-1
794062	f	########	57	state boar	2009	73	state boar	8346	2	B.Tech/B.E	computer	70.08	8346	0	Uttar Prad	2014	500	410	560	0.622643	385	-1
1088206	m	########	77	state boar	2008	75	state boar	13424	2	B.Tech/B.E	electronics	62	13424	0	Maharasht	2013	675	630	485	0.207392	405	260
1279958	m	########	81.2	state boar	2008	79.9	state boar	64	2	B.Tech/B.E	instrument	67.67	64	0	Uttar Prad	2013	395	565	645	-1	495	-1
471413	f	########	85	delhi boar	2009	88	all india bc	57	2	B.Tech/B.E	informatic	85	57	0	Haryana	2013	495	445	605	0.765674	485	-1
1088423	f	########	90	state boar	2009	82.1	state boar	2998	2	B.Tech/B.E	computer	85	2998	0	Telangana	2014	640	530	705	0.486747	615	-1
1066680	m	########	86.4	cbse	2009	86.2	cbse	1906	2	B.Tech/B.E	electronics	81.4	1906	1	West Beng	2014	720	630	750	0.338786	485	292
407672	m	########	84.13	0	2008	77	0	3801	2	B.Tech/B.E	informatic	75.2	3801	0	Karnataka	2012	385	515	465	0.525923	415	-1
205633	m	########	81.7	hse	2005	75.8	cbse	5508	2	B.Tech/B.E	computer	78.7	5508	0	Orissa	2011	475	245	485	0.735796	475	-1
924541	f	########	86	cbse	2010	89	cbse	429	2	B.Tech/B.E	computer	73.9	429	0	Uttar Prad	2014	520	570	620	0.9539	425	-1
512353	m	########	66.15	state boar	2009	54	state boar	3603	2	B.Tech/B.E	computer	66	3603	0	Maharasht	2013	360	495	280	0.793581	495	-1
1136577	m	########	79.29	icse	2009	68.67	cbse	5298	2	B.Tech/B.E	computer	76	5298	0	Tamil Nad	2013	545	505	595	0.356536	555	-1
781327	m	########	60	state boar	2006	50	state boar	11013	2	B.Tech/B.E	computer	69.94	11013	0	Orissa	2013	240	340	430	0.143257	295	-1
764522	f	########	58.4	cbse	2008	64.8	cbse	85	2	MCA	computer	80	85	0	Haryana	2014	440	415	554	0.793581	495	-1
203323	m	########	61	0	2003	59	0	3716	2	B.Tech/B.E	informatic	63	3716	1	Karnataka	2011	355	565	495	0.930371	565	-1
267354	m	########	50	board of s	2008	64	state boar	7564	2	B.Tech/B.E	electronics	70	7564	0	Haryana	2011	345	365	445	0.538387	-1	333
839626	f	########	67.06	state boar	2008	70.67	state boar	13271	2	B.Tech/B.E	computer	55.5	13271	1	Maharasht	2012	440	440	385	0.14479	435	-1
1169332	f	########	67	cbse	2008	61	cbse	16664	2	B.Tech/B.E	computer	75	16664	0	Chhattisga	2012	370	410	320	0.488348	405	-1
669692	m	########	73	state boar	2009	71	state boar	9699	2	B.Tech/B.E	informatic	61	9699	1	Rajasthan	2013	180	365	310	0.074546	255	433
48029	m	########	86.17	rbse	2006	78.6	cbse	2857	1	B.Tech/B.E	computer	73.34	2857	1	Rajasthan	2010	505	495	695	0.987207	645	-1
1268344	m	########	78	icse	2009	75	cbse	7765	2	B.Tech/B.E	mechanica	73	7765	0	Uttar Prad	2014	630	555	570	0.066961	-1	-1

ALGORITHMS IMPLEMENTED

1. **Linear Regression:** Linear regression is a model that assumes a linear relationship between the input variables (x) and the single output variable (y) i.e. y can be calculated from a linear combination of the input variables (x).

2. **Support Vector Machine (SVM):** Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data.

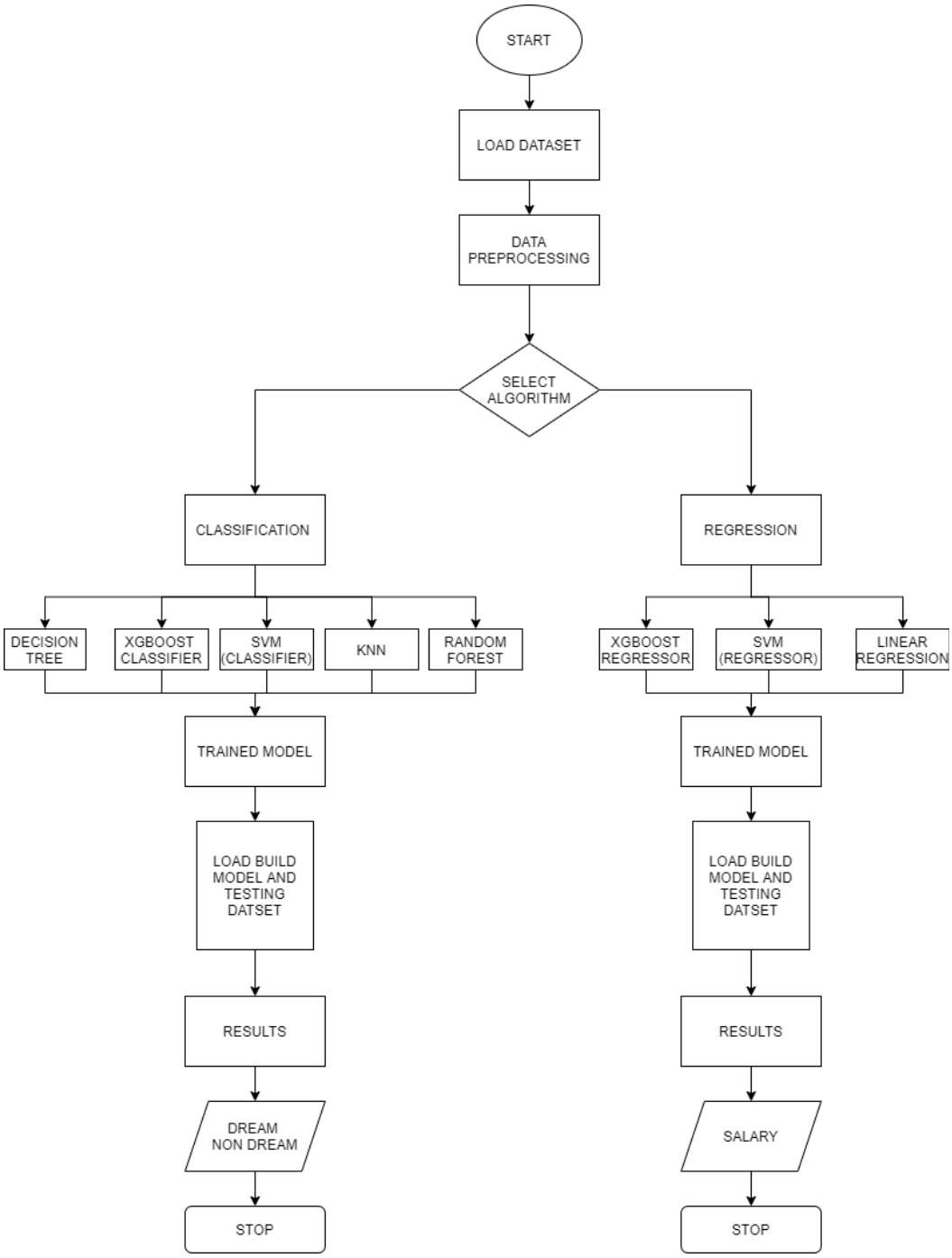
3. **XGBoost regression:** XGBoost regression is eXtreme Gradient Boosting an algorithm used for structured or tabular data. The library is laser focused on computational speed and model performance.

4. **Decision Trees:** Decision Trees follow Sum of Product (SOP) representation. The Sum of product (SOP) is also known as Disjunctive Normal Form. It is a decision support tool that uses a tree-like model of decisions and their possible consequences.

5. **K-nearest neighbors:** KNN is a non parametric supervised algorithm used in classification and regression. All work is done at query time. The KNN algorithm assumes that similar things exist in close proximity.

6. **Random Forest:** The random forest is a classification algorithm consisting of many decision trees. It uses bagging and features randomness when building each individual tree.

FLOW DIAGRAM



IMPLEMENTATION DETAILS

COLAB CODE

We are using an engineering salary prediction dataset from kaggle for our data house project. Dataset contains 34 columns and 2998 entries. Data Cleaning is an important part of any data analysis and classification problem. We must remove unnecessary features from our dataset. In this dataset we drop irrelevant columns like student ID, college ID, graduation year etc. Entire implementation details are given in the google collab link. After cleaning we preprocess our dataset by dividing it into Training and Testing data. Due to preprocessing accuracy of algorithm increases. Also we convert all alphanumeric input vectors in one hot vector representation so that it is easy for algorithms to process the data and find patterns in the dataset.

For training and classification we are using various Machine Learning Algorithms like KNN, Linear Regression, Support vector Machine, XGBoost,Decision Trees and Random Forest. We are also comparing accuracy, R2 score and F1 scores of each model to see which model is giving us better results. We build models and pass testing and training data. Based on salary we convert it into two classes. If salary is greater than 6,00,000 we assign label 1 (Dream salary) and for less than 6,00,000 we assign label 0 (Non-Dream salary). Sklearn library provides a great variety of pre-built models. Sklearn also has a prebuilt accuracy calculator. In our project we are using prebuilt sklearn models for prediction and classification of our data points. Also they are very easy to use and implement.

RESULTS/SCREENSHOTS

RMSE Value of different Regression algorithms

```
results = pd.DataFrame({
 'Model': [
  'Support Vector Machine(regressor',
 'XGBoost Regressor','linear regression'
 ],
 'Score': [rmse_svr,
 rmse_xgbr, rmse_lr]})
result_df = results.sort_values(by='Score', ascending=True)
result_df = result_df.set_index('Score')
result_df.head(9)
```

Score	Model
0.19	Support Vector Machine(regressor
0.27	XGBoost Regressor
0.41	linear regression

Mean absolute error of different Regression algorithms

```
results = pd.DataFrame({
 'Model': [
  'Support Vector Machine(regressor)',
 'XGBoost Regressor','linear regression'
 ],
 'Score': [mae_svr,
 mae_xgbr, mae_lr]})
result_df = results.sort_values(by='Score', ascending=True)
result_df = result_df.set_index('Score')
result_df.head(9)
```

Score	Model
0.14	Support Vector Machine(regressor)
0.21	XGBoost Regressor
0.35	linear regression

R2 Score of different Regression algorithms

```
results = pd.DataFrame({
 'Model': [
  'Support Vector Machine(regressor)',
 'XGBoost Regressor','linear regression'
 ],
 'Score': [svr_r2,
 xgb_r2, lin_r2]})
result_df = results.sort_values(by='Score', ascending=True)
result_df = result_df.set_index('Score')
result_df.head(9)
```

Score	Model
0.32	linear regression
0.72	XGBoost Regressor
0.85	Support Vector Machine(regressor)

Accuracy of different classification algorithms

```
results = pd.DataFrame({
 'Model': [
 'Random Forest','KNN', 'svc',
 'xgb','decision tree'
 ],
 'Score': [acc_rf,
 acc_knn, acc_svmc, acc_xgbc,
 acc_dtc]})
result_df = results.sort_values(by='Score', ascending=False)
result_df = result_df.set_index('Score')
result_df.head(9)
```

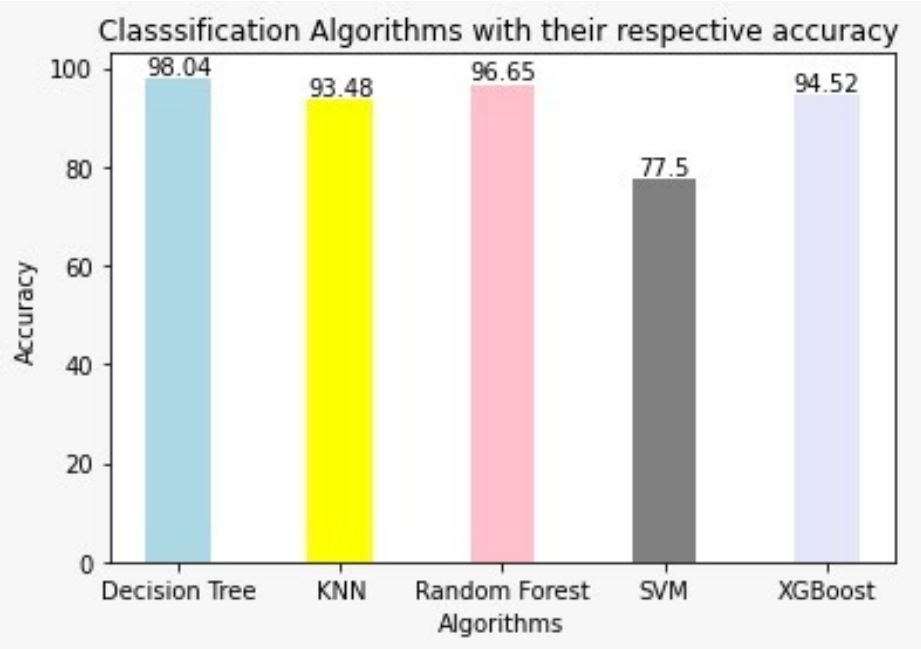
Score	Model
98.04	decision tree
96.65	Random Forest
94.52	xgb
93.48	KNN
77.50	svc

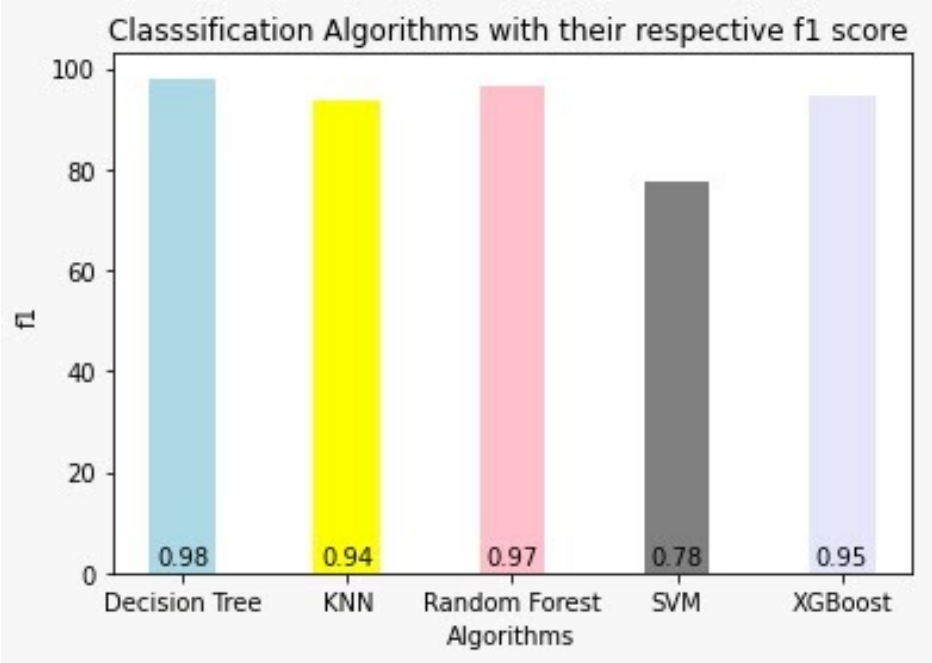
F1 Score of different classification algorithms

```
results = pd.DataFrame({
 'Model': [
 'Random Forest','KNN', 'svc',
 'xgb','decision tree'
 ],
 'Score': [f1_rf,
 f1_knn, f1_svc, f1_xgb,
 f1_dt]})
result_df = results.sort_values(by='Score', ascending=False)
result_df = result_df.set_index('Score')
result_df.head(9)
```

Score	Model
0.98	decision tree
0.97	Random Forest
0.95	xgb
0.94	KNN
0.78	svc

COMPARISON OF EVALUATION MEASURES





Classification:

Algorithms Implemented	F1 Score	Accuracy
Decision Tree	0.98	98.04
KNN	0.94	93.48
Random Forest	0.97	96.65
SVM	0.78	77.50
XGBoost	0.95	94.52

Regression:

Algorithms Implemented	RMSE Score	R2 Score	Mean Absolute Error
Linear Regression	0.41	0.32	0.35
XGBoost	0.27	0.72	0.21
SVM	0.19	0.85	0.14

CONCLUSION

Prediction of graduate placements is a key problem in the analysis of employability. Data-driven studies on empirical data have the potential to address this challenge. Aspiring Minds dataset is one of the most suitable dataset towards this objective as it appropriately describes the attributes of a freshly graduate engineering student. Thus, we have used it for prediction of graduate placements and salary, as well as for building machine learning models for analysis of employability. We highlighted the inherent imbalance in these data that render studies solely relying on performance metrics such as degree, irrelevant. Such a methodology of data-driven approach can also serve as a basis for future studies towards prediction of salary and placements, and identification of key features linked to employability of candidates. The objective of this study was to predict the overall employability of a new engineering graduate. While we have used the inherent features in the Aspiring Minds dataset, one may use other features as well for the same data-driven strategy, such as any previous work experience etc.

REFERENCES

1. DATASET: https://www.kaggle.com/manishkc
2. MACHINE LEARNING ALGORITHMS: https://www.geeksforgeeks.org/machine-learning/
3. https://towardsdatascience.com/
4.