PROJECT PRESENTATION

JobFit AI

Intelligent job-resume matching

By:- Faarid Bilal Misgar (2022BECE048)
Adnan Nazir (2022BECE051)

Interns at:- NetEdge Computing Solutions Pvt Ltd

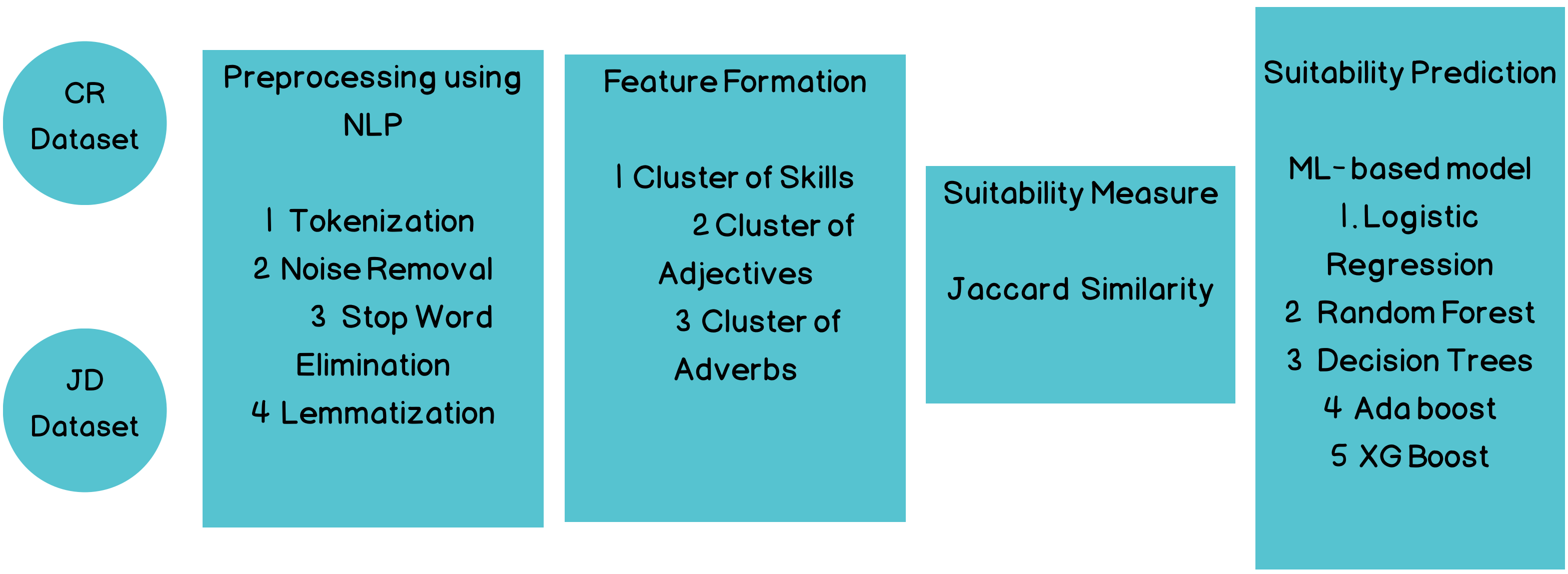
Challenges in Traditional Recruitment

- Hiring a suitable candidate for a certain job is highly demanding and requires several intense processes.
- It can be time consuming manually screening thousands of Resumes
- Many organizations face this challenge to hire a suitable candidate this way.

Role of AI in making Recruitment Feasible

- An Artificial Intelligence based system is developed to measure and predict a suitable candidate from available candidate Resume(CR)and Job description(JD).
- Three Clusters are prepared from the dataset of JD and CR as Skills, Adjectives, and Adverbs.
- The Jaccard similarity is measured between these clusters and ML-based techniques are used to predict the candidate's suitability such as Good Fit, Potential Fit or No Fit

Project Workflow



Data Source:-

Dataset was taken from Hugging Face

We divided the Main dataset into two parts

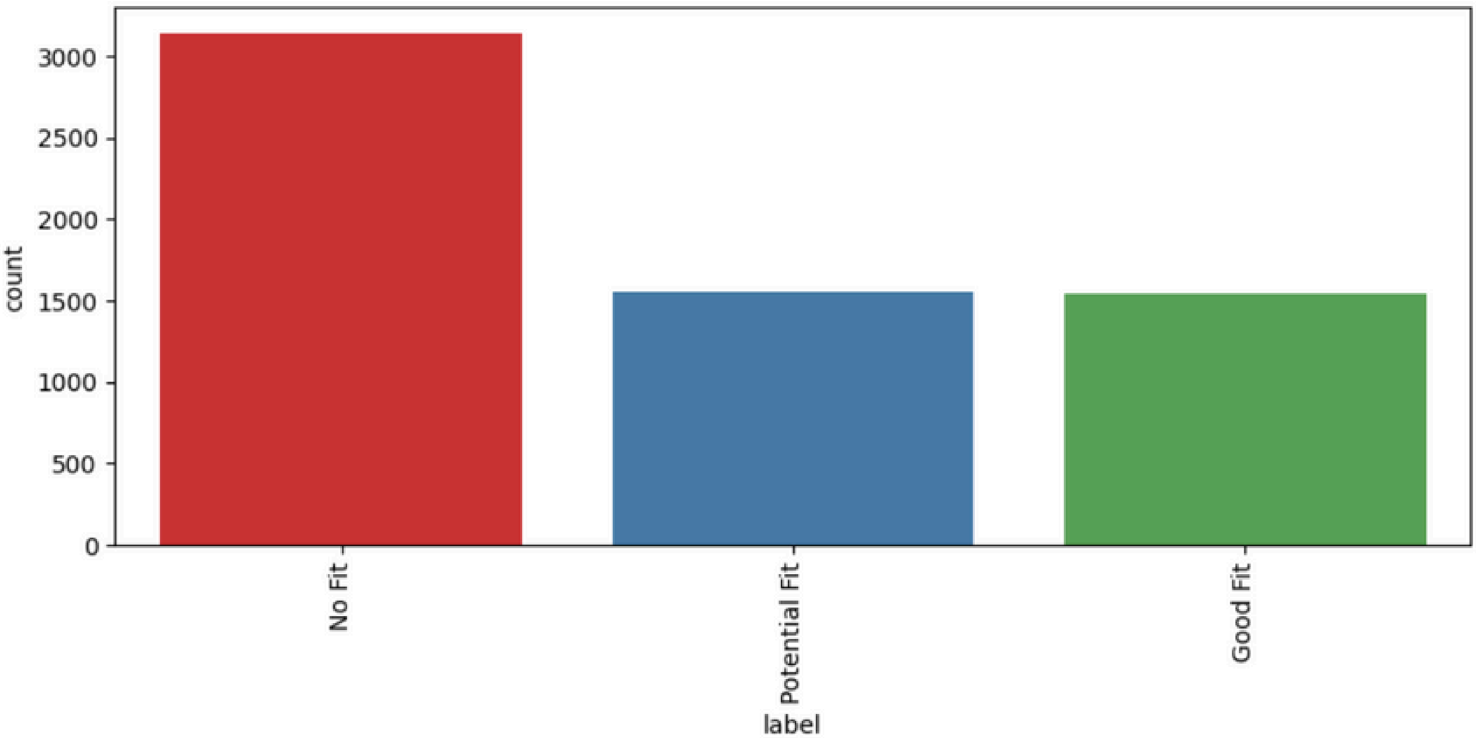
1. Train:- Shape (6241, 4)
2. Test :- Shape (1759, 4)

So, Train-Test Split was avoided.

Raw DataSet

	resume_text	job_description_text	label
0	SummaryHighly motivated Sales Associate with e...	Net2Source Inc. is an award-winning total work...	No Fit
1	Professional SummaryCurrently working with Cat...	At Salas OBrien we tell our clients that were ...	No Fit
2	SummaryI started my construction career in Jun...	Schweitzer Engineering Laboratories (SEL) Infr...	No Fit
3	SummaryCertified Electrical Foremanwith thirte...	Mizick Miller & Company, Inc. is looking for a...	No Fit
4	SummaryWith extensive experience in business/r...	Life at Capgemini\nCapgemini supports all aspe...	No Fit

Target Label



Handling Imbalanced Data

SMOTE (Synthetic Minority Over-sampling Technique) was applied on Train Set to handle imbalances in the raw dataset

After applying SMOTE the balanced dataset was formed
Train:- Shape (9429, 4)

Text preprocessing(Using NLP)

Example Job Description (JD) and Resume (CR) Text:

- **Job Description(JD):**
- "We are looking for a skilled Data Scientist with experience in Python, Machine Learning, and NLP. Candidates should have 3+ years of experience and a Master's degree in Computer Science."

- **Candidate Resume(CR):**
- "Experienced Data Scientist with expertise in Python, ML, and Deep Learning. Holds an MSc in CS with 4 years of industry experience."

Text preprocessing(Using NLP)

1. Tokenization:-

Splitting text into words, punctuation, numbers.

Output (JD):
['We', 'are', 'looking', 'for', 'a', 'skilled', 'Data', 'Scientist', 'with', 'experience', 'in', 'Python', ',', 'Machine', 'Learning', ',', 'and', 'NLP', '.', 'Candidates', 'should', 'have', '3', '+', 'years', 'of', 'experience', 'and', 'a', 'Master's', 'degree', 'in', 'Computer', 'Science', '.']

Output (CR):
['Experienced', 'Data', 'Scientist', 'with', 'expertise', 'in', 'Python', ',', 'ML', ',', 'and', 'Deep', 'Learning', '.', 'Holds', 'an', 'MSc', 'in', 'CS', 'with', '4', 'years', 'of', 'industry', 'experience', '.']

2. Stopword Removal

Removes common words like "are", "with", "and", etc.

Output (JD)

['skilled', 'Data', 'Scientist', 'experience', 'Python', 'Machine', 'Learning', 'NLP', '3', 'years', 'experience', 'Master's', 'degree', 'Computer', 'Science']

Output (CR):

['Experienced', 'Data', 'Scientist', 'expertise', 'Python', 'ML', 'Deep', 'Learning', 'MSc', 'CS', '4', 'years', 'industry', 'experience']

3. Lemmatization

Converts words to their base form (reducing variations like "years" → "year").

Output (JD):

['skill', 'Data', 'Scientist', 'experience', 'Python', 'Machine', 'Learn', 'NLP', '3', 'year', 'experience', 'Master', 'degree', 'Computer', 'Science']

Output (CR):

['experience', 'Data', 'Scientist', 'expert', 'Python', 'ML', 'Deep', 'Learn', 'MSc', 'CS', '4', 'year', 'industry', 'experience']

4. Part-of-Speech (POS) Tagging

Identifies nouns, verbs, adjectives, etc.

Output (JD):

- Scientist -> NOUN
- Machine -> NOUN
- Deep -> ADJ
- Learning -> VERB
- years -> NOUN

Output (CR):

- Data -> NOUN
- Scientist -> NOUN
- Python -> NOUN
- Deep -> ADJ

5. Named Entity Recognition (NER)

Detects names, skills, degree etc.

Output (JD):

- Python -> SKILL
- Machine Learning -> SKILL
- NLP -> SKILL
- 3+ years -> EXPERIENCE

Output (CR):

- Python -> SKILL
- ML -> SKILL
- Deep Learning -> SKILL
- 4 years -> EXPERIENCE

Feature Extraction (Forming Clusters)

1. Skill Cluster for both CR and JD

For Pattern based Named Entity Recognition (NER) we used Entity Ruler and Spacy Library to Extract skills from dataset

2. Adjective Cluster for both CR and JD

We used Part-of-Speech Tagging (POS) for identifying adjectives in text.

if the token's POS is “ADJ” then append

3. Adverb Cluster for both CR and JD

We used Part-of-Speech Tagging (POS) for identifying adverb in text.

if the token's POS is “ADV” then append

Dataset with Extracted Features

- We then used these three clusters on CR's and JD's dataset to create new features using apply function of Pandas Library

label	pre_resume	pre_jd	resume_skills	jd_skills	resume_adj	resume_adv	jd_adj	jd_adv
No Fit	summary7 + year experience bi developer prove ...	key responsibility create intricate wiring net...	[testing, analytics, query optimization, data ...	[component, interaction, manufacturing enginee...	[valuable, active, ambitious, new, internal, s...	[primarily, also, most, high]	[seamless, vital, high, detailed, strong, mech...	[seamless, vital, high, detailed, strong, mech...
No Fit	professional backgroundanalyst verse data anal...	personal development good growth explore new s...	[testing, business, crystal, server, data anal...	[software, testing, business, engineering, des...	[external, registration, high, managerial, new...	[weekly, effectively, as, well, daily]	[innovative, appropriate, intelligent, federal...	[innovative, appropriate, intelligent, federal...
No Fit	executive profilededicated professional accomp...	location tampa fl exp 7 10 yrs spoc tushar ksh...	[business, play, accounting, compliance, resea...	[javascript, component, business, certificatio...	[prestigious, sure, high, new, mutable, profes...	[extensively, successfully, solely, independen...	[dental, innovative, fide, federal, competitiv...	[dental, innovative, fide, federal, competitiv...
No Fit	summarytyee highlightsmicrosoft excel word out...	primary location melbourne florida v soft cons...	[business, box, analytics, documentation, mark...	[testing, software engineering, engineering, c...	[valuable, afterschool, weekly, various, good,...	[daily]	[strategic, accurate, architectural, specific,...	[strategic, accurate, architectural, specific,...
No Fit	summaryit certify engineer astqb certified qa...	at oregon specialty group accounting & payroll...	[testing, engineering, library, software, crys...	[software, business, accounting, compliance, d...	[unique, exploratory, high, detailed, specific...	[more, accurately, effectively, as, together, ...	[big, high, accurate, critical, appropriate, i...	[big, high, accurate, critical, appropriate, i...

Suitability Measurement

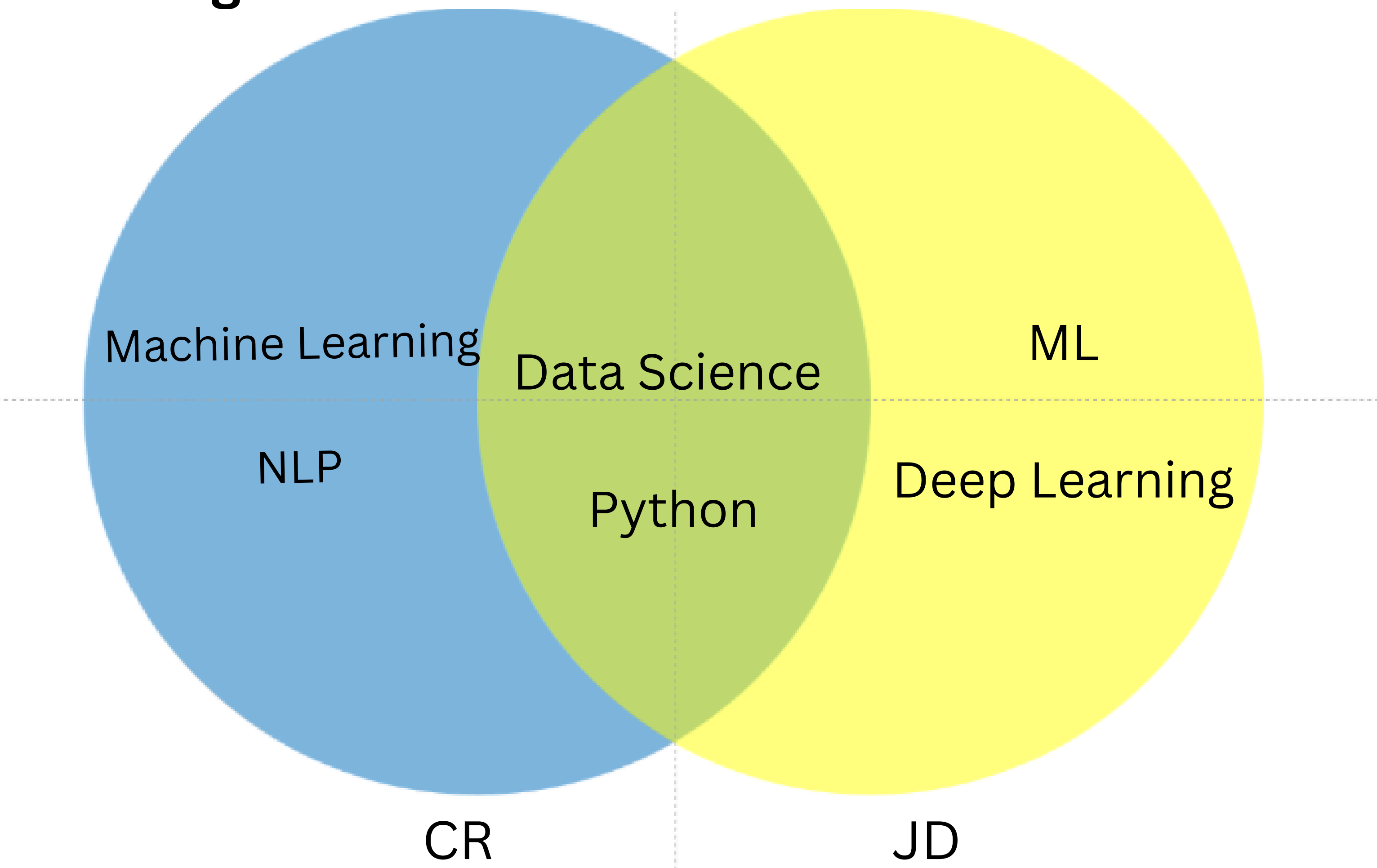
Jaccard Similarity

The Jaccard similarity of clusters is the ratio of number of common words to total words in those clusters

$$J(A, B) = A \cap B / A \cup B$$

Here:- A-> CR
B-> JD

Working:-



2. Identify Common and Unique Skills

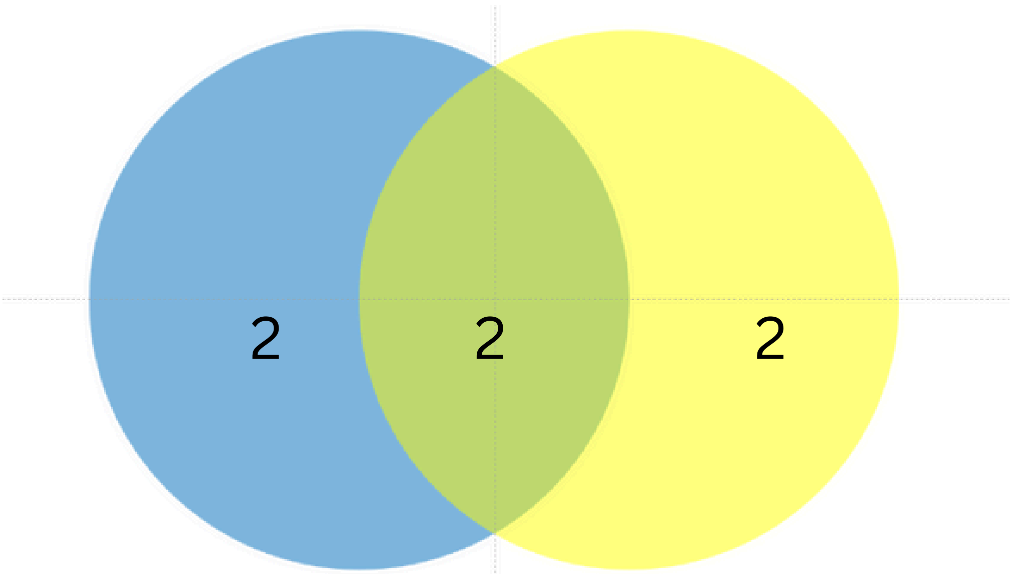
Intersection (Common Skills between JD & CR):
[Data Science, Python]

Union (Total Unique Skills from JD & CR):
[Data Science, Python, Machine Learning, NLP, ML, Deep Learning]

3. Calculate Jaccard Similarity

Jaccard Similarity Formula: $J(A,B) = A \cap B \,/\, A \cup B$

therefore,
$J(A, B) = 2/6 = 0.33$ (33%)



$J(A, B) = 2/6 = 0.33$ (33%)

This Jaccard Similarity is applied on these features

J(resume_skills , jd_skills) = jaccard_skills
J(resume_adjectives, jd_adjectives) = jaccard_adjectives
J(resume_adverbs, jd_adverbs) = jaccard_adverbs

	jaccard_skills	jaccard_adj	jaccard_adv	label
0	0.041667	0.004204	0.000000	No Fit
1	0.029762	0.002467	0.000000	No Fit
2	0.012500	0.001923	0.002222	No Fit
3	0.000000	0.002262	0.000000	No Fit
4	0.011765	0.003395	0.002778	No Fit

Model Selection

The suitability prediction is carried out using AI-based classifiers namely:-

1 Logistic Regression

2 Random Forest

3 Decision Tree

4 XG Boost

5 AdaBoost

Model Classification

Multi-Class Classification is performed on these clusters and are categorized into three classes:-

1 Good Fit

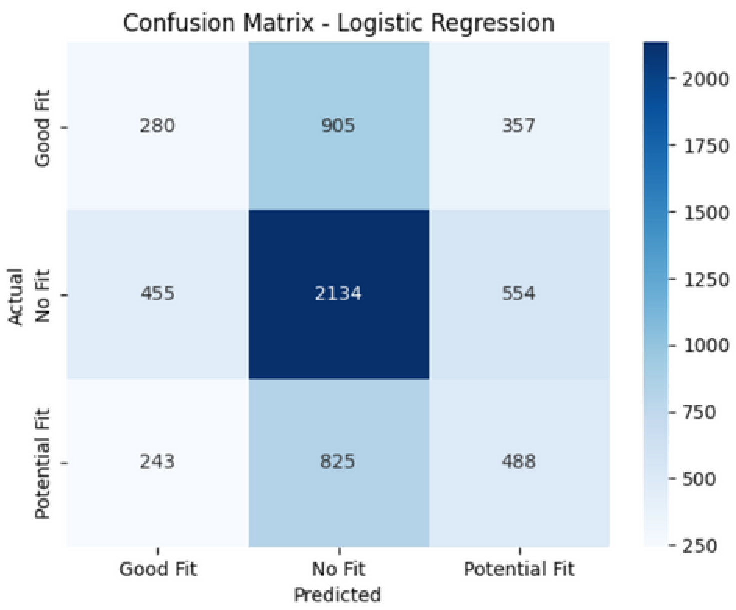
2 Potential Fit

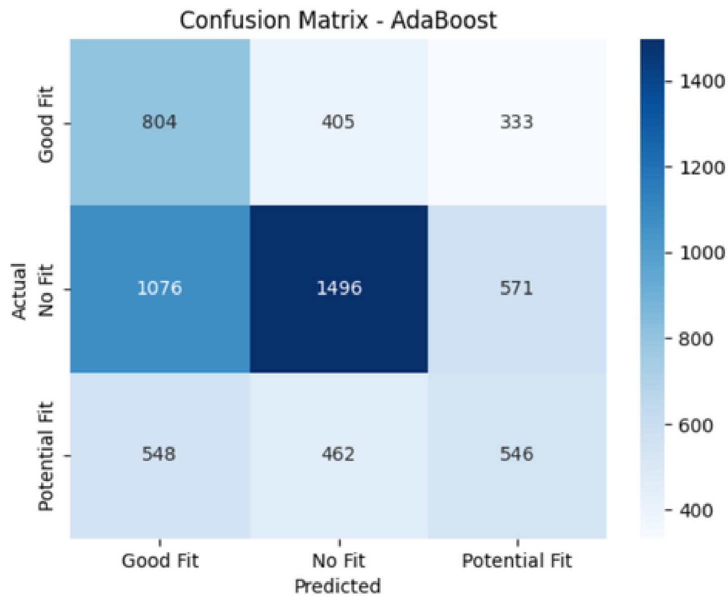
3 No Fit

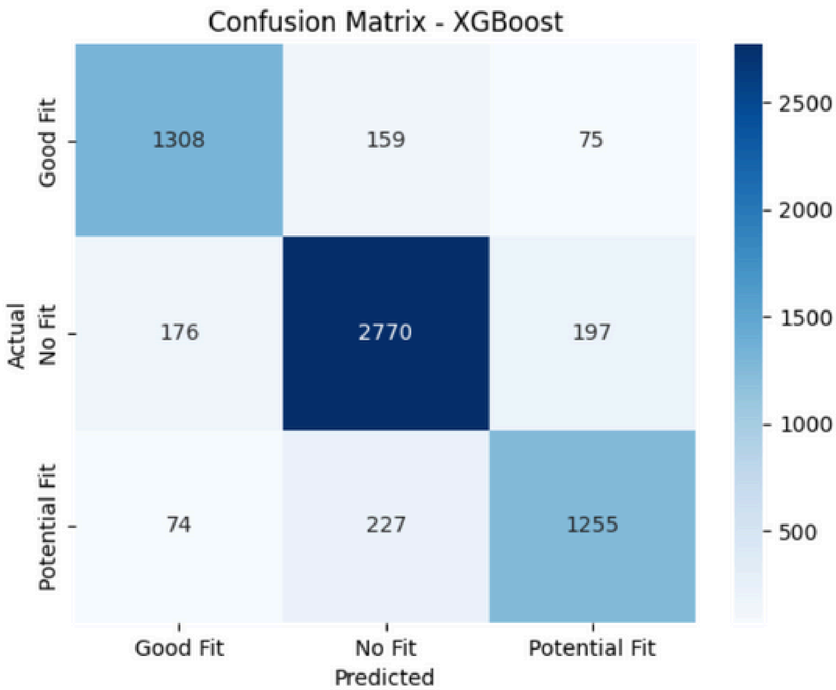
Hyperparameters were tuned on 5 folds of Cross-Validation using GridSearchCV for all ML models

Confusion Matrices:-

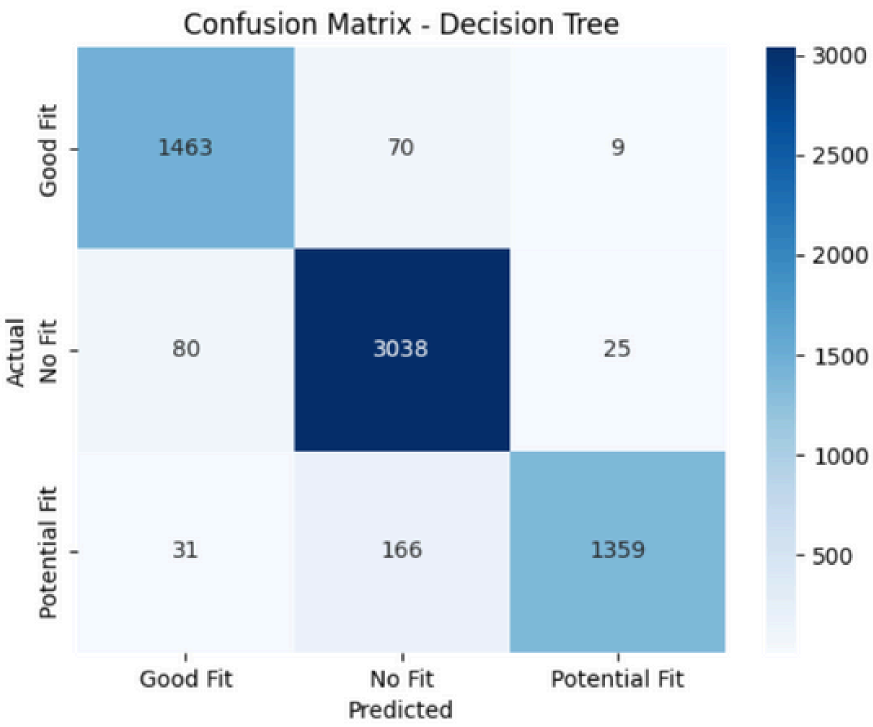
It's used to evaluate the performance of models by comparing actual vs. predicted values.

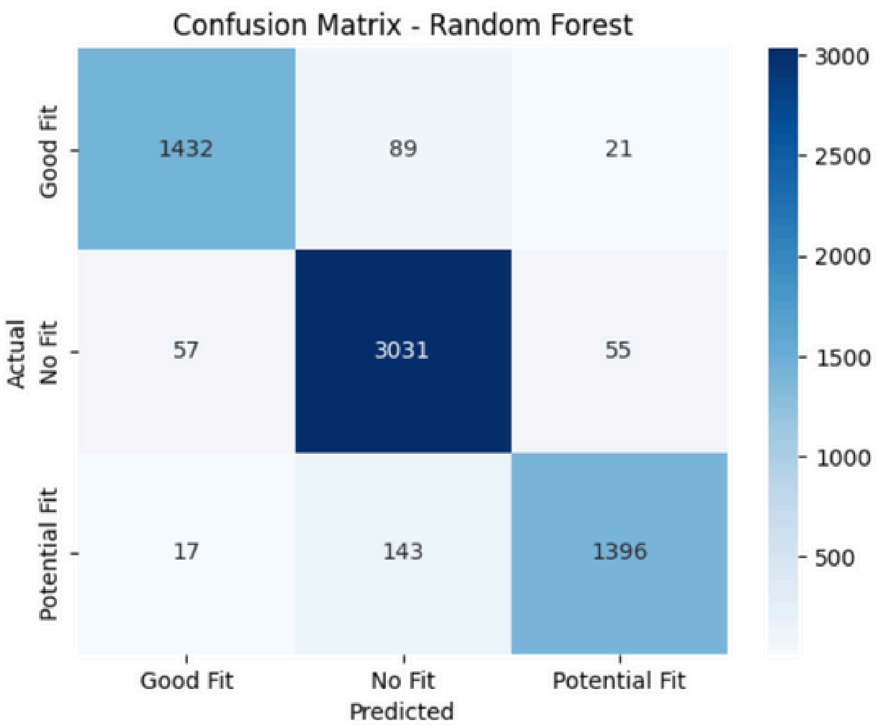






- Best Performance was observed in Decision Tree and Random Forest





Model Evaluation

	Accuracy	Precision	Recall	f1-score
1. Logistic Regression	0.4650	0.29	0.18	0.22
2. **Random Forest**	**0.9388**	**0.95**	**0.93**	**0.94**
3. **Decision Tree**	**0.9390**	**0.93**	**0.95**	**0.94**
4. AdaBoost	0.4560	0.33	0.52	0.41
5 XG Boost	0.8545	0.85	0.84	0.85

Conclusion

Best Models:-

- Random Forest & Decision Tree (High accuracy, precision, recall, and F1-score).

Reason:-

- Tree-based models are best suited for small datasets with non-linear relationships.
- Tree-based models handle class imbalance well which were perfected using SMOTE

THANK YOU