Determining Job Placement

B126 Grp 8

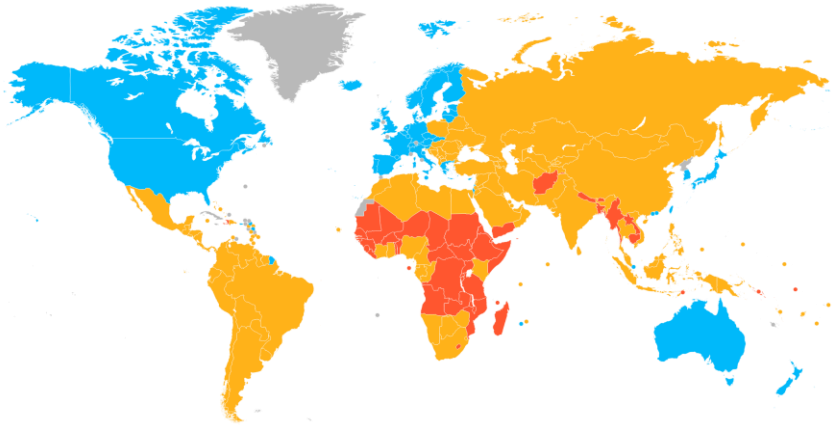
PHYO SANDAR WIN

AUGUSTINE JESURAJ SENCHIA GLADINE

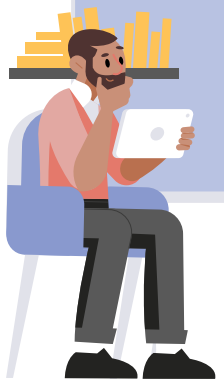
SEET TZE SHIN, CHEYENNE

GitHub Link: https://github.com/phyosandarwin/Jobmatch

Practical Motivation



Which variables are the most important in predicting someone getting a job?



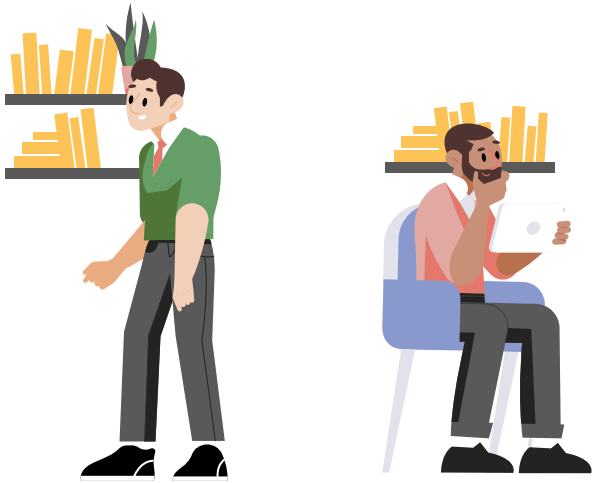
Dataset details

Based in India

Mix of numerical and categorical data

Predictor

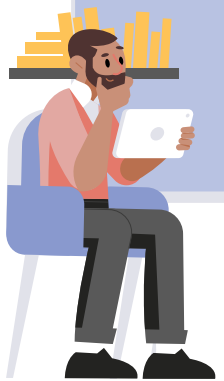
- Education history
- Work experience
- Personal information



Response

- Whether the candidate received a job offer

Predictor Variables



Categorical variables

Data type: String

01	Gender	F/M
02	SSC_board	Central/Others
03	HSC_Board	Central/Others
04	HSC_Subject	Commerce/Science/ Arts/Others
05	Undergrad_degree	Comm&Mgmt/ Sci&Tech/Others
06	Work_experience	Yes/No
07	Specialisation	Mkt&HR/Mkt&Fin

Numerical Variables

Data type: Float
Range: 0-100

01 **SSC_Percentage**

Senior secondary exams percentage (10th Grade)

02 **HSC_Percentage**

Higher secondary exams percentage (12th Grade)

03 **Degree_percentage**

Percentage of marks in undergrad degree

04 **EMP_Test_Percentage**

Aptitude test percentage

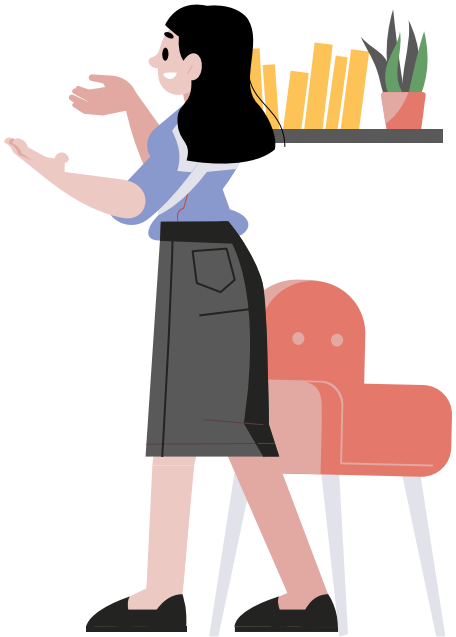
05 **MBA_Percent**

Percentage of marks in MBA degree

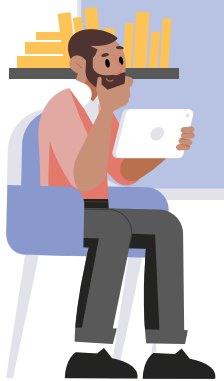
Response Variable

Status

- Categorical variable
- String
- Placed/Not placed



Cleaning Data



Cleaning Data - Overview

	gender	ssc_percentage	ssc_board	hsc_percentage	hsc_board	hsc_subject	degree_percentage	undergrad_degree	work_experience	emp_test_percentage	specialisation	mba_
0	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	
1	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	
2	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	
3	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	
4	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	
...	...	...	...	...	...	...	...	...	...	...	...	
210	M	80.60	Others	82.00	Others	Commerce	77.60	Comm&Mgmt	No	91.0	Mkt&Fin	
211	M	58.00	Others	60.00	Others	Science	72.00	Sci&Tech	No	74.0	Mkt&Fin	
212	M	67.00	Others	67.00	Others	Commerce	73.00	Comm&Mgmt	Yes	59.0	Mkt&Fin	
213	F	74.00	Others	66.00	Others	Commerce	58.00	Comm&Mgmt	No	70.0	Mkt&HR	
214	M	62.00	Central	58.00	Others	Science	53.00	Comm&Mgmt	No	89.0	Mkt&HR	

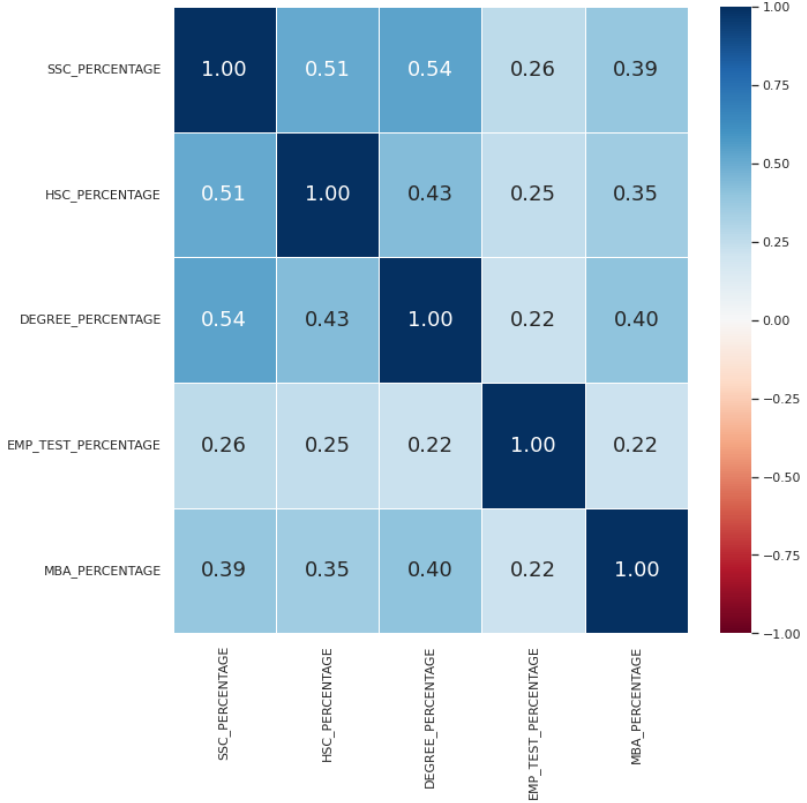
215 rows × 13 columns

```
Number of duplicate records :  0
```

Cleaning Data – One Hot Encoding

GENDER_F	GENDER_M	SSC_BOARD_Central	SSC_BOARD_Others	HSC_BOARD_Central	...	HSC_SUBJECT_Commerce	HSC_SUBJECT_Science	UNDERGRAD_DEGREE_Comm&Mgmt	UNI
0.0	1.0	0.0	1.0	0.0	...	1.0	0.0	0.0	
0.0	1.0	1.0	0.0	0.0	...	0.0	1.0	0.0	
0.0	1.0	1.0	0.0	1.0	...	0.0	0.0	1.0	
0.0	1.0	1.0	0.0	1.0	...	0.0	1.0	0.0	
0.0	1.0	1.0	0.0	1.0	...	1.0	0.0	1.0	
...	...	...	...	...	...	...	...	...	
0.0	1.0	0.0	1.0	0.0	...	1.0	0.0	1.0	
0.0	1.0	0.0	1.0	0.0	...	0.0	1.0	0.0	
0.0	1.0	0.0	1.0	0.0	...	1.0	0.0	1.0	
1.0	0.0	0.0	1.0	0.0	...	1.0	0.0	1.0	
0.0	1.0	1.0	0.0	0.0	...	0.0	1.0	1.0	

EDA (Numeric): Relationship between numeric variables

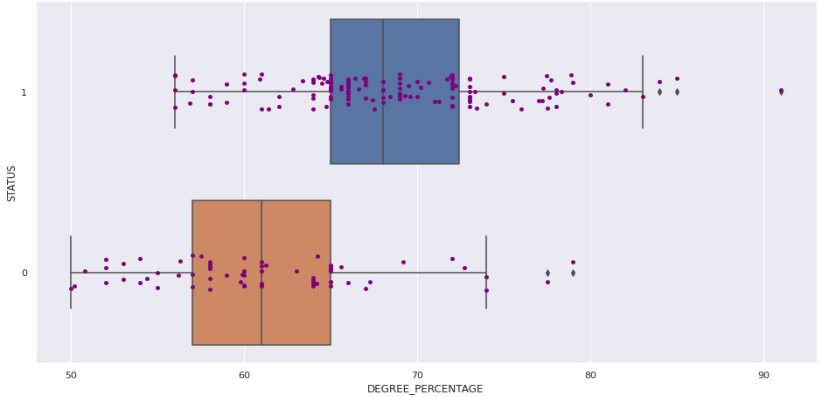


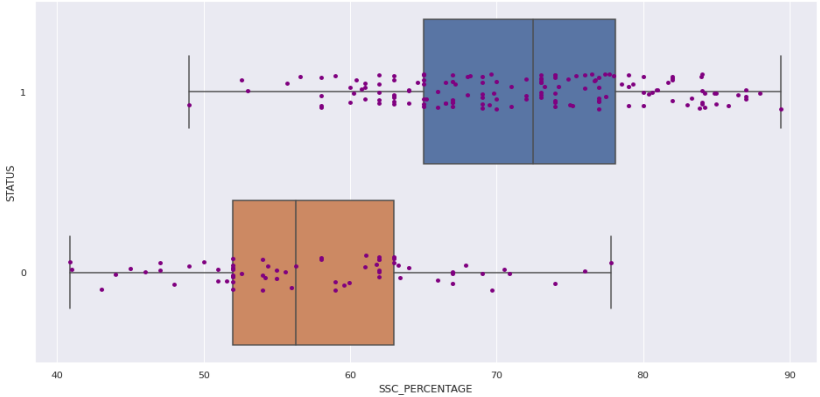
High correlations (in descending order):

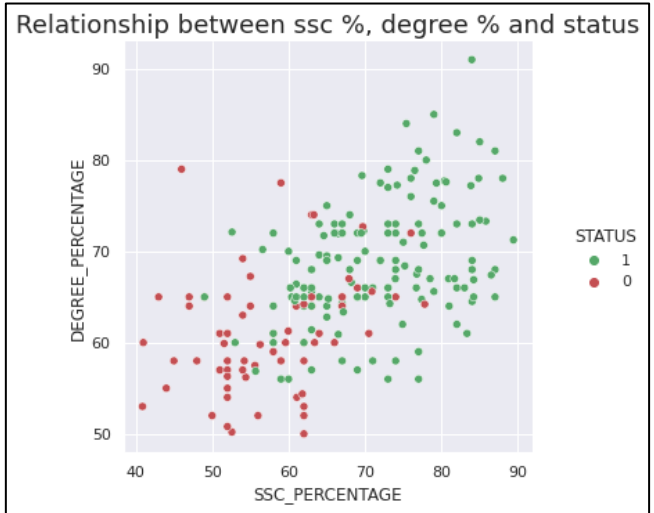
1. Degree % and Senior Secondary % : **0.54**
2. Higher Secondary % and Senior Secondary % : **0.51**
3. Degree % and Higher Secondary %: **0.43**

Degree %, Higher Secondary %, Senior Secondary % are more strongly correlated with each other.

EDA (Numeric): Relationship between numeric variables and 'STATUS'



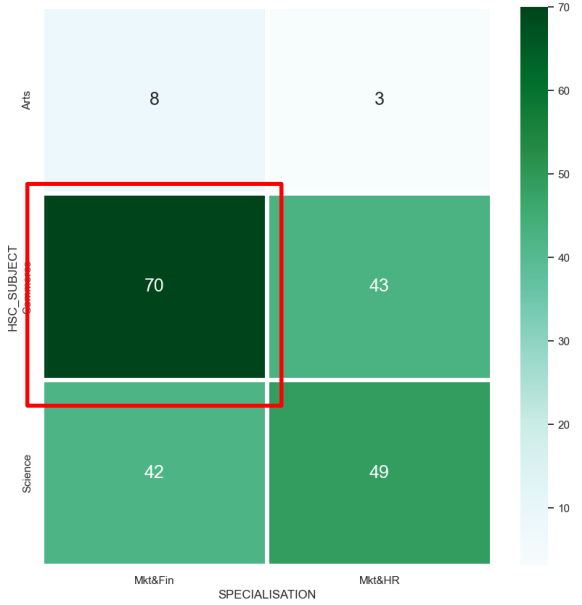




1. Boxplots of numeric variables against status
   - boxplots of SSC % and Degree % are more distinctly different (lesser overlap)
2. Relationship plot verifies importance of these variables
   - Larger distribution of points labelled positive placement status for higher SSC % and Degree %

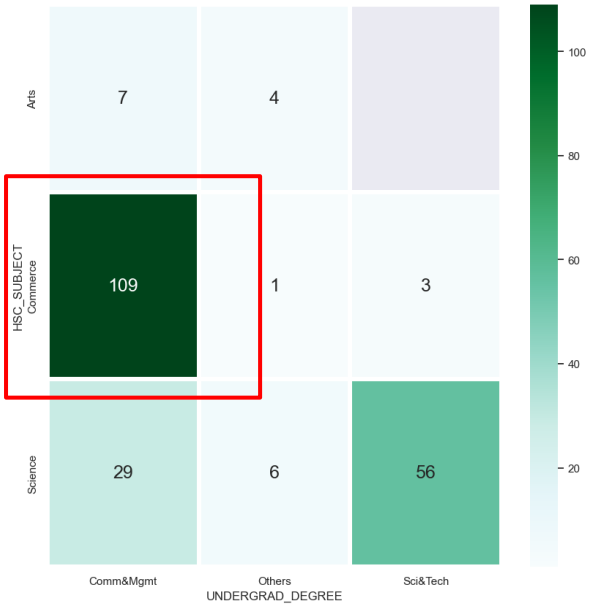
SSC % and Degree % have stronger relationship with status

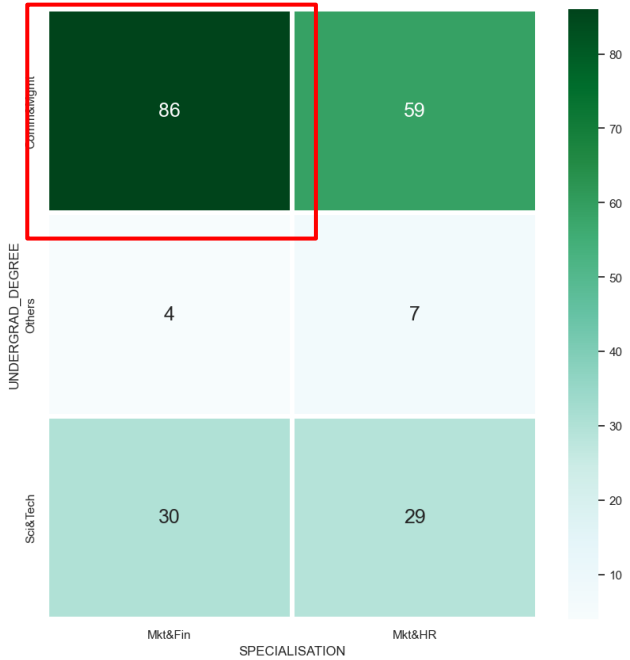
EDA (Categoric): Relationship between categorical variables



Distribution of hsc subjects and mba specialisation

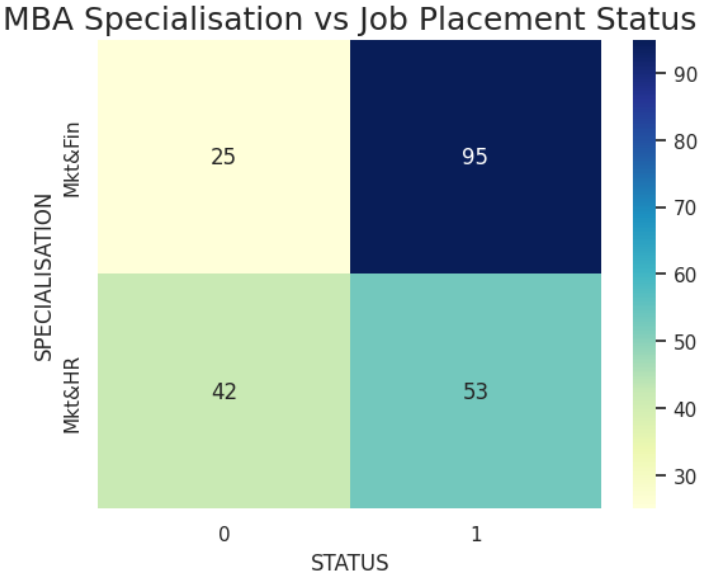
Distribution of hsc subjects and undergrad degrees





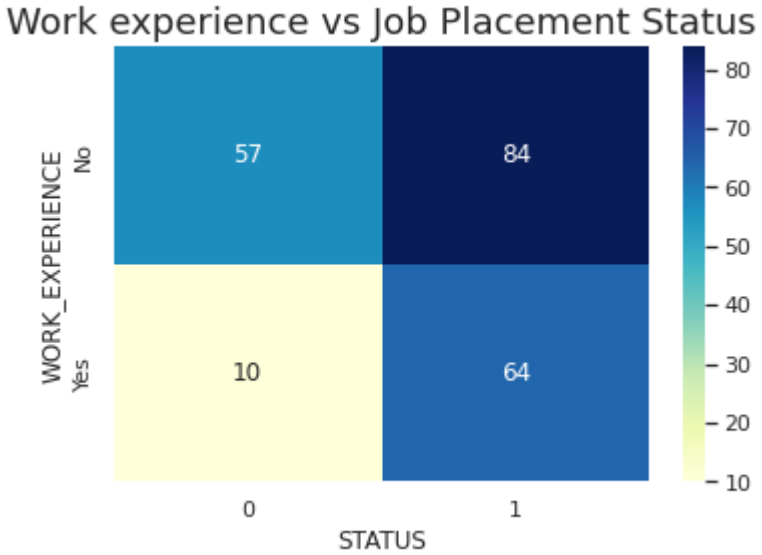
Distribution of undergrad degree and specialisation

EDA (Categoric): Relationship between categorical variables and 'STATUS'



Chi-square = 12.440> critical value 3.841

Chi-square = 15.154> critical value 3.841

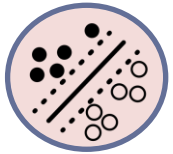


Machine Learning Models

Binary Classification problem

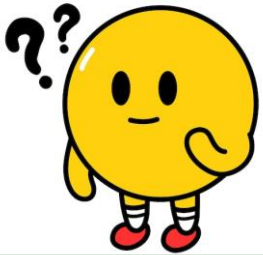
Decision Tree

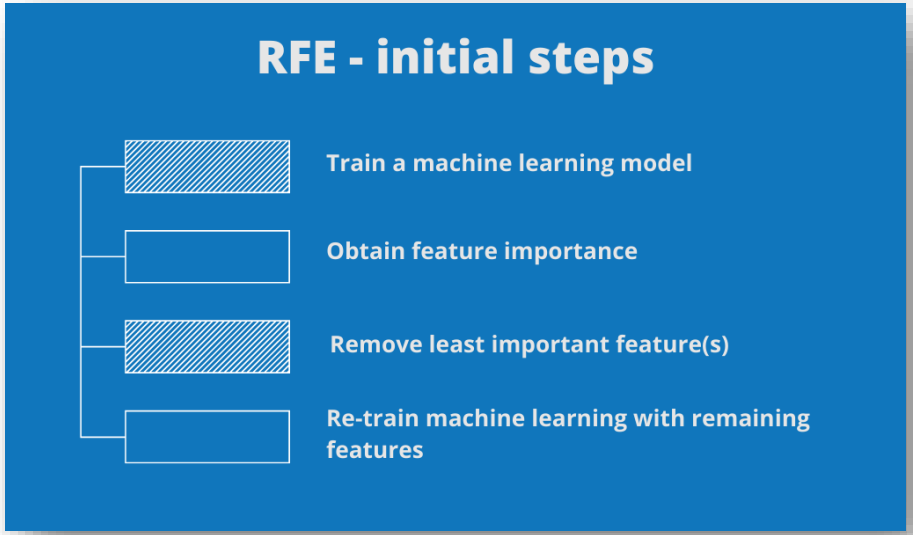
Logistic Regression



Support Vector Machines (SVM)

Feature Selection Technique:
Recursive Feature Elimination (RFE)



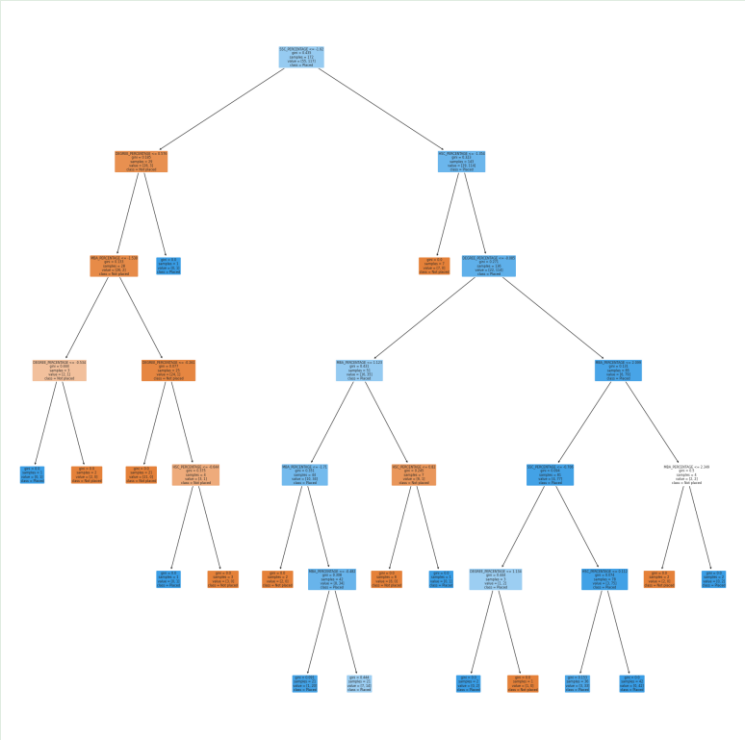


Why choose RFE over Select K-Best?

- RFE considers whether features are related to each other, but SelectKBest does not.
- RFE is more accurate in measuring feature importance since it uses the model's performance, while SelectKBest only looks at how each feature is related to the target variable.
- RFE works better with Support Vector Machines, a non-linear model that we will be using later, but Select K-best may not work as well with SVMs that do not support univariate statistical tests.

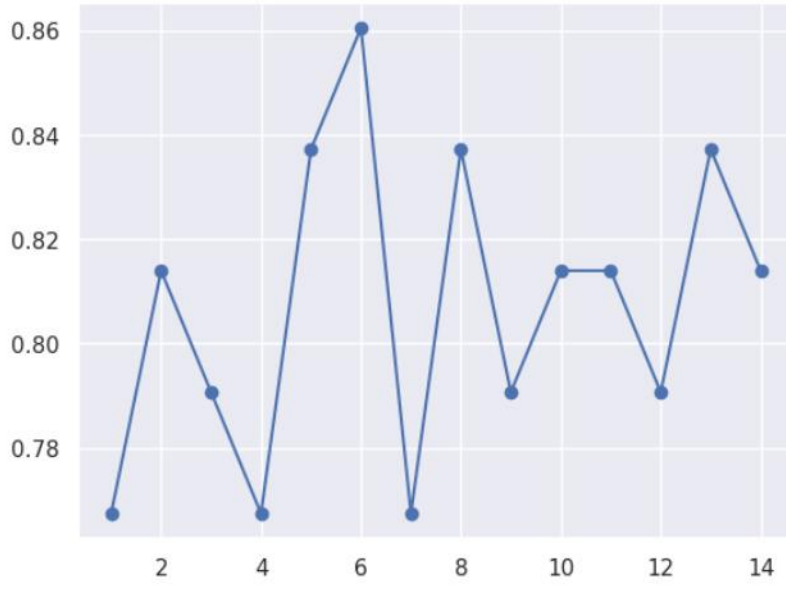
Decision Tree

used to predict the class of the target variable by learning simple decision rules inferred from previous data



Best depth

```
# finding best tree depth to do the model

list = []

models = []

for i in range(1, 15):
    dectree = DecisionTreeClassifier(max_depth = i)  # create the decision tree object
    dectree.fit(X_train, y_train)
    y_train_pred = dectree.predict(X_train)
    y_test_pred = dectree.predict(X_test)

    # Check the Goodness of Fit (on Test Data)
    list.append(dectree.score(X_test, y_test))
    models.append(dectree)
```




```
[ ]  # Best depth
     print(f"Best accuracy: {max(list)}\nDepth: {list.index(max(list))+1}")
```

```
Best accuracy: 0.8604651162790697
Depth: 6
```

According to RFE, our selected features are SSC %, HSC % , Degree %, MBA % , Science HSC Subject

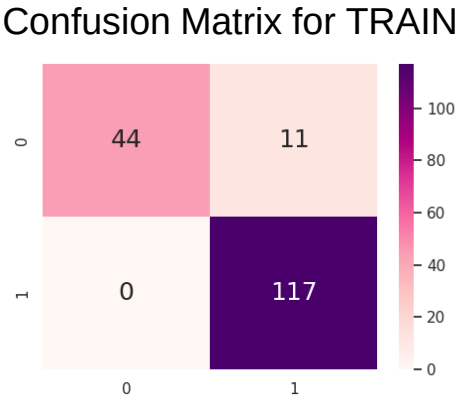
```
# Print 5 features using RFE for Decision Tree model
dectree = DecisionTreeClassifier(max_depth=6)
print('Decision Tree Features\n')
rfe_selection(dectree, X_train, y_train)
```

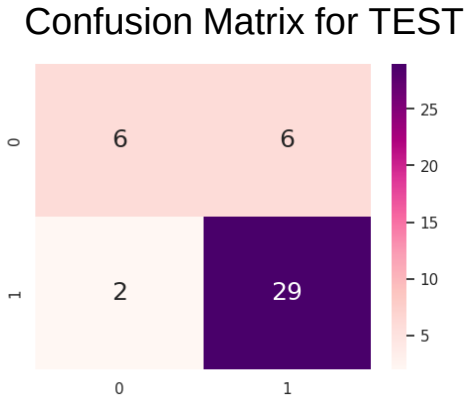
```
Decision Tree Features

[ True  True  True False  True False False False False False False
 False  True False False False False False False False]
[ 1  1  1 11  1 12 10  8 14 16 15 13  9  1  7  6  5  4  3  2 17]

 Selected features are:  Index(['SSC_PERCENTAGE', 'HSC_PERCENTAGE', 'DEGREE_PERCENTAGE',
       'MBA_PERCENTAGE', 'HSC_SUBJECT_Science'],
      dtype='object')
```

Decision Tree

1. 10-fold Cross Validation average accuracy score: 76.76 %

2.





Goodness of fit: Train Data

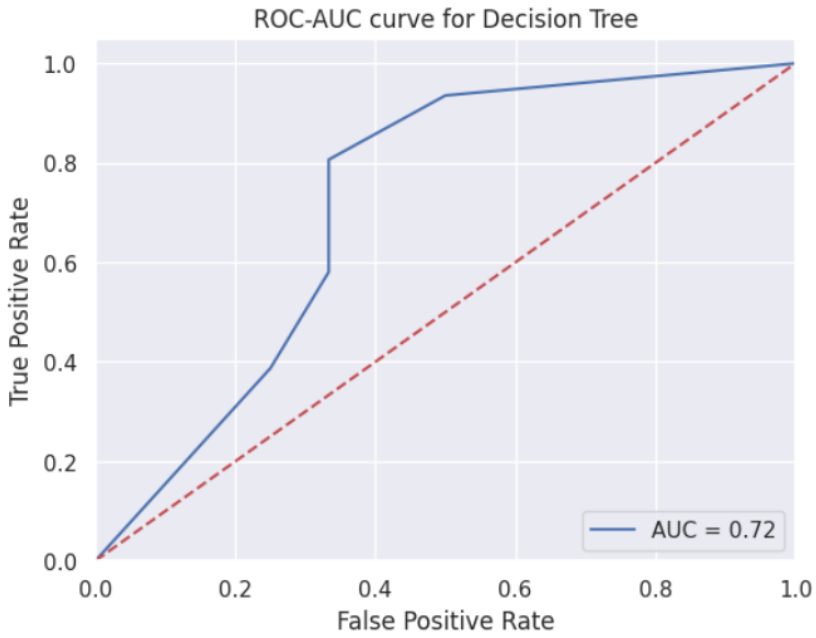
- Classification accuracy = 94%
- True Positive rate/ Recall = 0.91
- False Positive rate = 0.0
- True Negative rate = 1.0
- False Negative rate = 0.085

Goodness of fit: Test Data

- Classification accuracy = 81%
- True Positive rate/ Recall = 0.83
- False Positive rate = 0.25
- True Negative rate 0.75
- False Negative rate = 0.17

Decision Tree

3

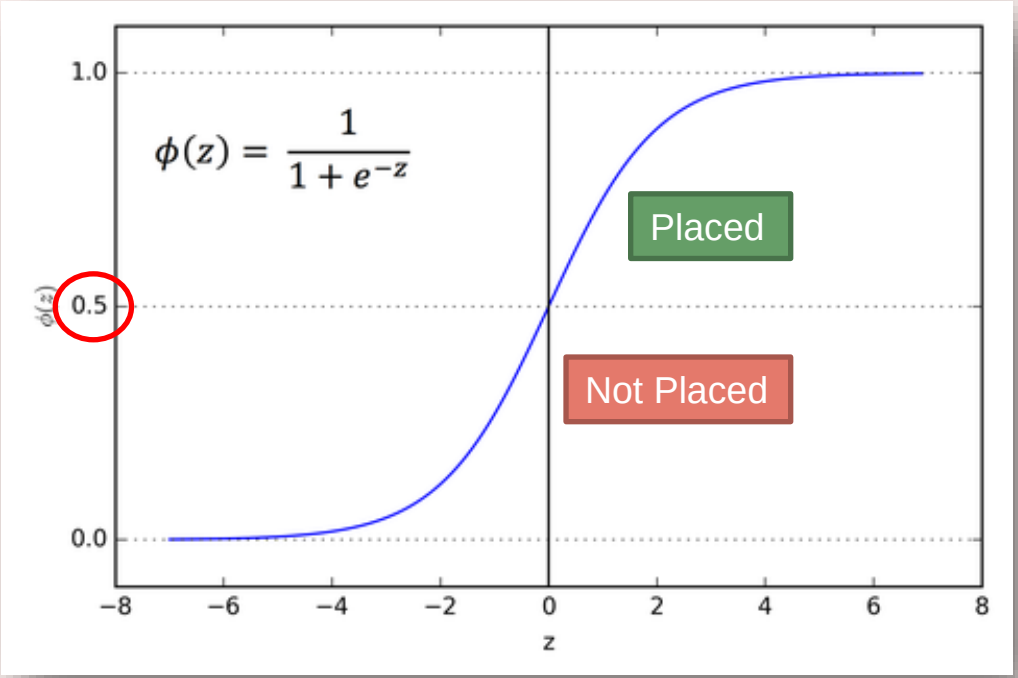


Deduction:

ROC-AUC score of 0.72 indicates that decision tree model is moderately good in distinguishing between 'placed' and 'not placed' classes.

Logistic Regression

predicts the probability of a binary outcome (1 or 0) based on input features applied into the sigmoid function



According to RFE, our selected features are SSC %, HSC % , Degree %, MBA %, No Work Experience

```
# Print 5 features using RFE for LogReg Model
print('Logistic Regression Features\n')
rfe_selection(logreg, X_train, y_train)
```

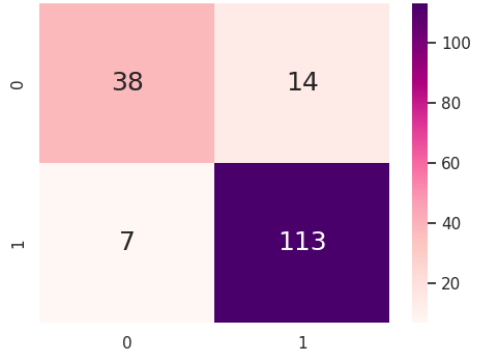
```
Logistic Regression Features

[ True  True  True False  True False False False False False False False
 False False False False False False False  True False]
[ 1  1  1 10  1  9  5 11 14  8  2  6  7 12  3 13 15 16 17  1  4]

 Selected features are:  Index(['SSC_PERCENTAGE', 'HSC_PERCENTAGE', 'DEGREE_PERCENTAGE',
       'MBA_PERCENTAGE', 'WORK_EXPERIENCE_No'],
      dtype='object')
```

Logistic Regression

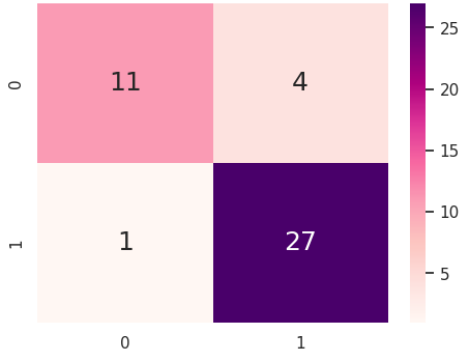
1. 10-fold Cross Validation average accuracy score: 85.42 %

2.

Confusion Matrix for TRAIN



Confusion Matrix for TEST



Goodness of fit: Train Data

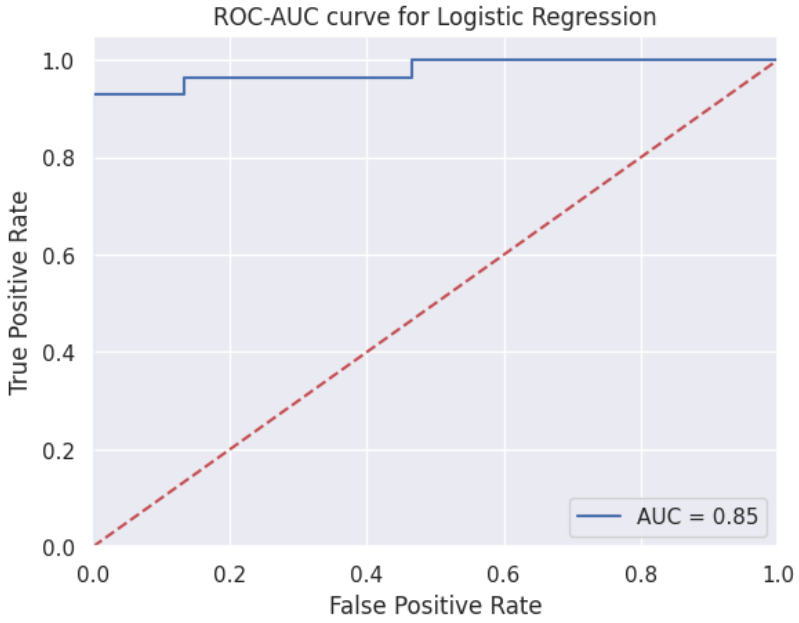
- Classification accuracy = 87.79%
- True Positive rate/ Recall = 0.94
- False Positive rate = 0.269
- True Negative rate = 0.731
- False Negative rate = 0.058

Goodness of fit: Test Data

- Classification accuracy = 88.37%
- True Positive rate/ Recall = 0.96
  False Positive rate = 0.267
- True Negative rate 0.733
- False Negative rate = 0.036

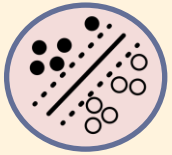
Logistic Regression

3



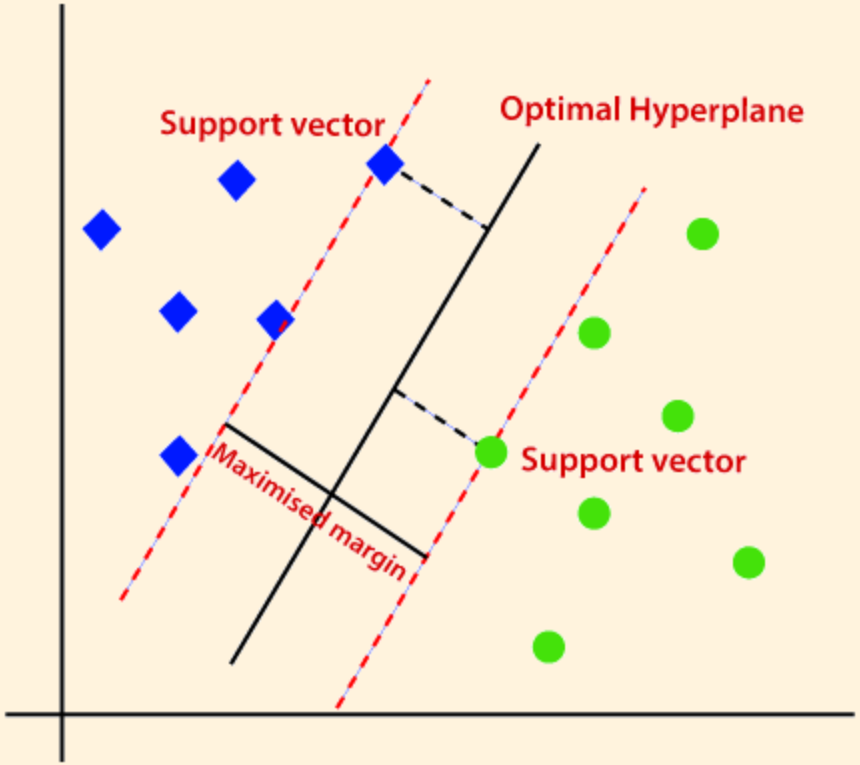
Deduction:

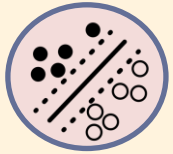
ROC-AUC score of 0.85 indicates that logistic regression model is good in distinguishing between 'placed' and 'not placed' classes.



Support Vector Machine

What is SVM?





Support Vector Machine

According to RFE, our selected features are SSC %, Degree %, MBA %, No Work Experience and Work Experience

```
# Print 5 features using RFE for Support Vector Machine
print('Support Vector Machine Features\n')
rfe_selection(svc, X_train, y_train)
```

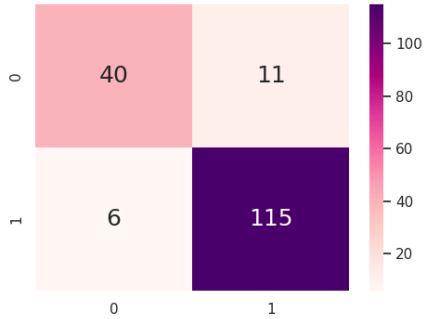
```
Support Vector Machine Features

[ True False  True False  True False False False False False False False
 False False False False False False False  True  True]
[ 1  5  1  3  1  8  2 11  7 16 13 15  6 14  4 17  9 10 12  1  1]

 Selected features are:  Index(['SSC_PERCENTAGE', 'DEGREE_PERCENTAGE', 'MBA_PERCENTAGE',
       'WORK_EXPERIENCE_No', 'WORK_EXPERIENCE_Yes'],
      dtype='object')
```

Support Vector Machine

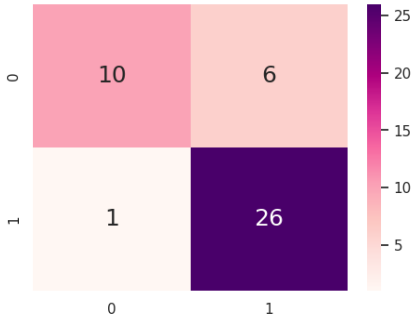
1. 10-fold Cross Validation average accuracy score: 86.01%

2.

Confusion matrix for TRAIN



Confusion matrix for TEST



Goodness of fit: Train Data

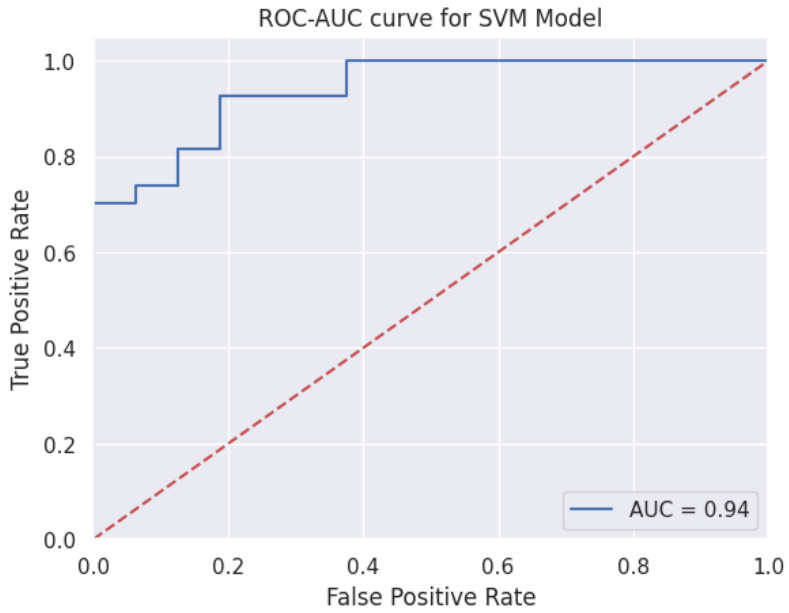
- Classification accuracy = 90.1%
- True Positive rate/ Recall = 91.2%
- False Positive rate = 13.0%
- True Negative rate = 87.0%
- False Negative rate = 8.7%

Goodness of fit: Test Data

- Classification accuracy = 83.7%
- True Positive rate/ Recall = 81.3%
- False Positive rate = 9.0%
- True Negative rate = 90.9%
- False Negative rate = 18.8%

Support Vector Machine

3



Deduction:

ROC-AUC score of 0.94 indicates that support vector machine model is very good in distinguishing between 'placed' and 'not placed' classes.

What we learnt?

New techniques used

Chi Square Test

Correlation analysis between categorical data

Recursive Feature Elimination

Feature selection technique

Logistic Regression

ML model

Support Vector Machine

ML model

Overall Model Performance

	Model	CV Accuracy	Precision	Recall	F1 Score	ROC-AUC score
0	Decision Tree	0.767647	0.828571	0.935484	0.878788	0.717742
1	Log Regression	0.854248	0.870968	0.964286	0.915254	0.848810
2	SVM (SVC)	0.860131	0.812500	0.962963	0.881356	0.939815

Log regression has the highest F1 score (good balance between precision and recall).

- model is correctly identifying the positive cases (i.e., correctly predicting who gets placed in a job)
- while minimizing false positives (i.e., wrongly predicting that someone will get placed in a job).

Data-driven insights

01 Logistic regression is the most accurate model out of the three

02 Attributes to focus on:
SSC %, HSC % , Degree %, MBA %, No work experience

03 Hiring trends:
- Increasing focus on work experience
- Less gender bias unlike we presumed

Thank you!