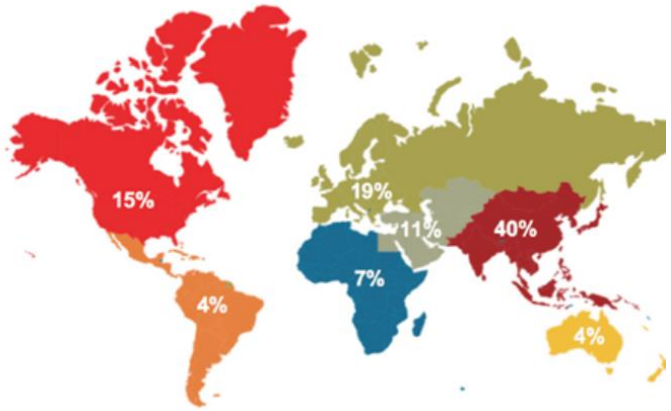
Click to add text

FRAUD FIRM REGRESSION AND CLASSIFICATION

- AISHWARYA BHAVSAR (029371509)
- MADGULA SARMA (029339438)
- JUSTIN KIEU (014151382)
- JANKI PATEL (029356143)
- MUDRA CHAUDHARY (029475821)



MOTIVATION

Increasing Fraud around the globe

How to identify Frauds?

Types of Fraud Detection Techniques

DATA SET

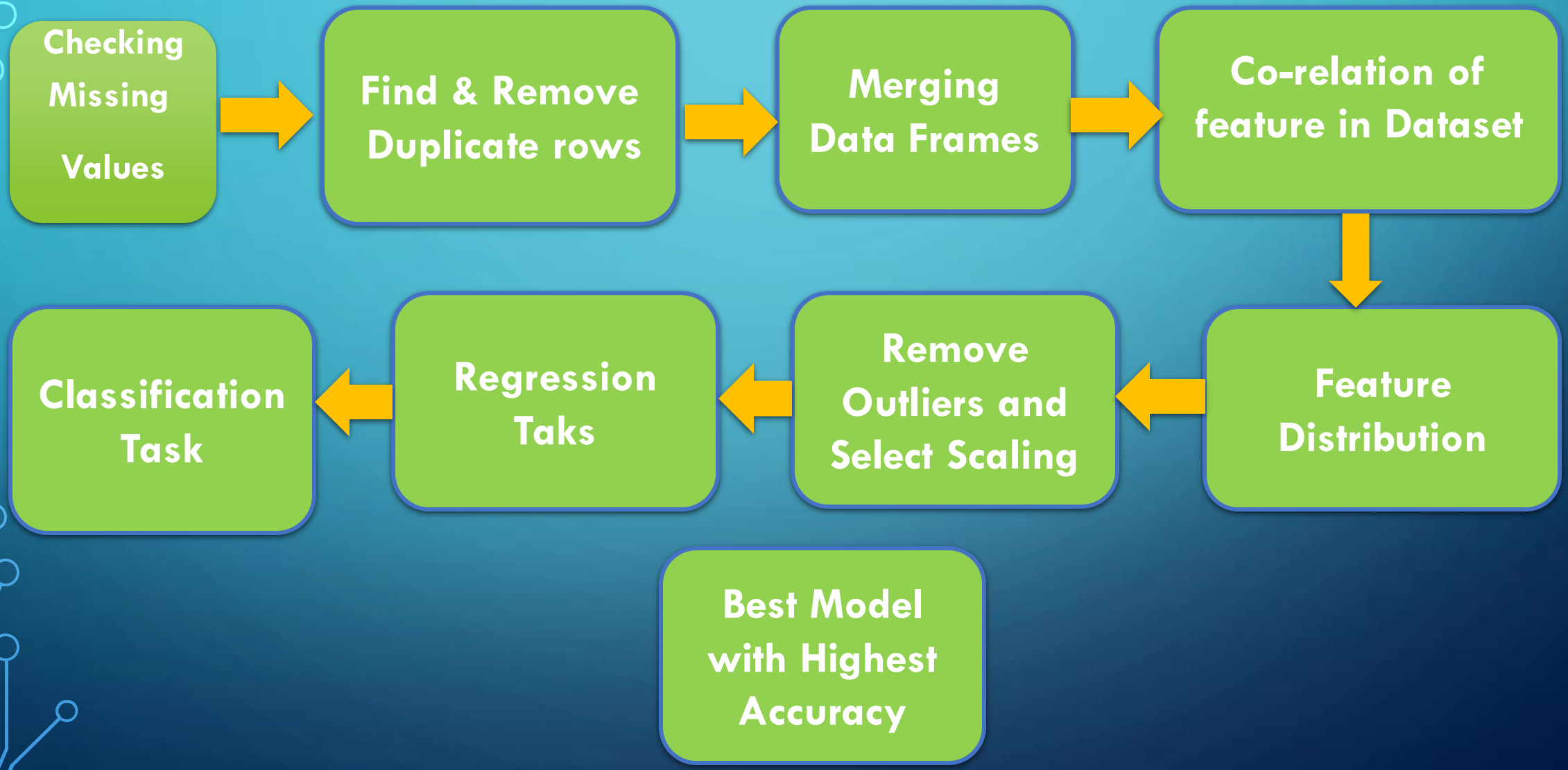
Information on firms :

- Public Health
- Buildings and roads
- Corporate

Contains values such as:

- Historical risk value of target
- Unique ID of a city
- Score and risk values of every report

PROJECT ROAD MAP



REGRESSION TASK

For estimating the relationships between a dependent variable and one or more independent variables.

1) KNN Regressor

2) Linear Regression

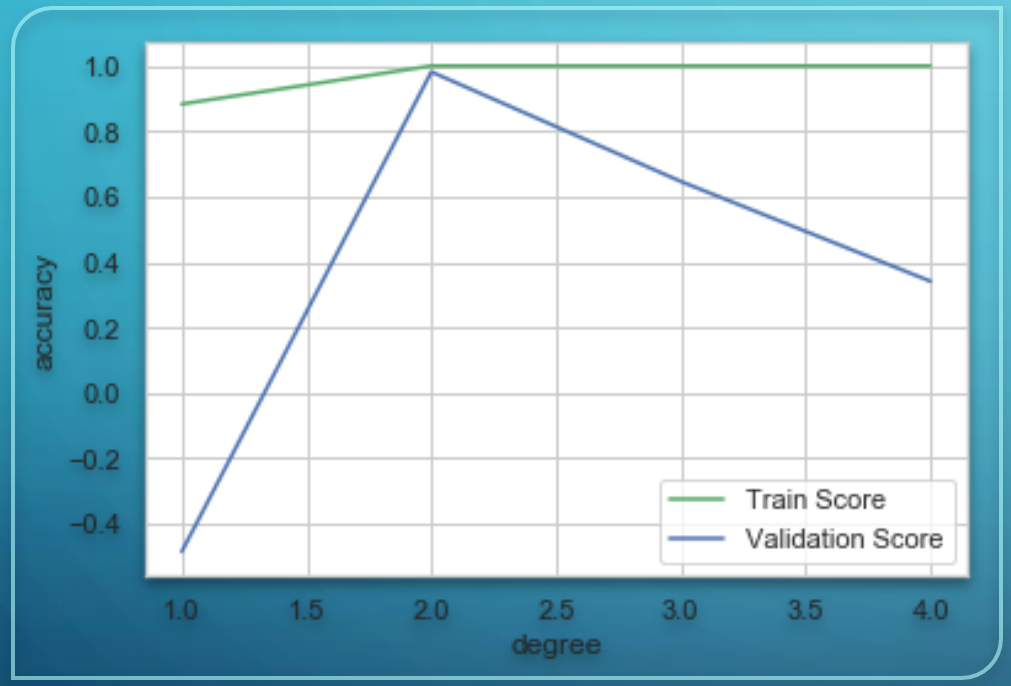
3) Polynomial Regression

4) Ridge Regression

5) Lasso Regression

6) Support Vector Machine with Kernel trick – Linear, Rbf, Poly

POLYNOMIAL REGRESSION

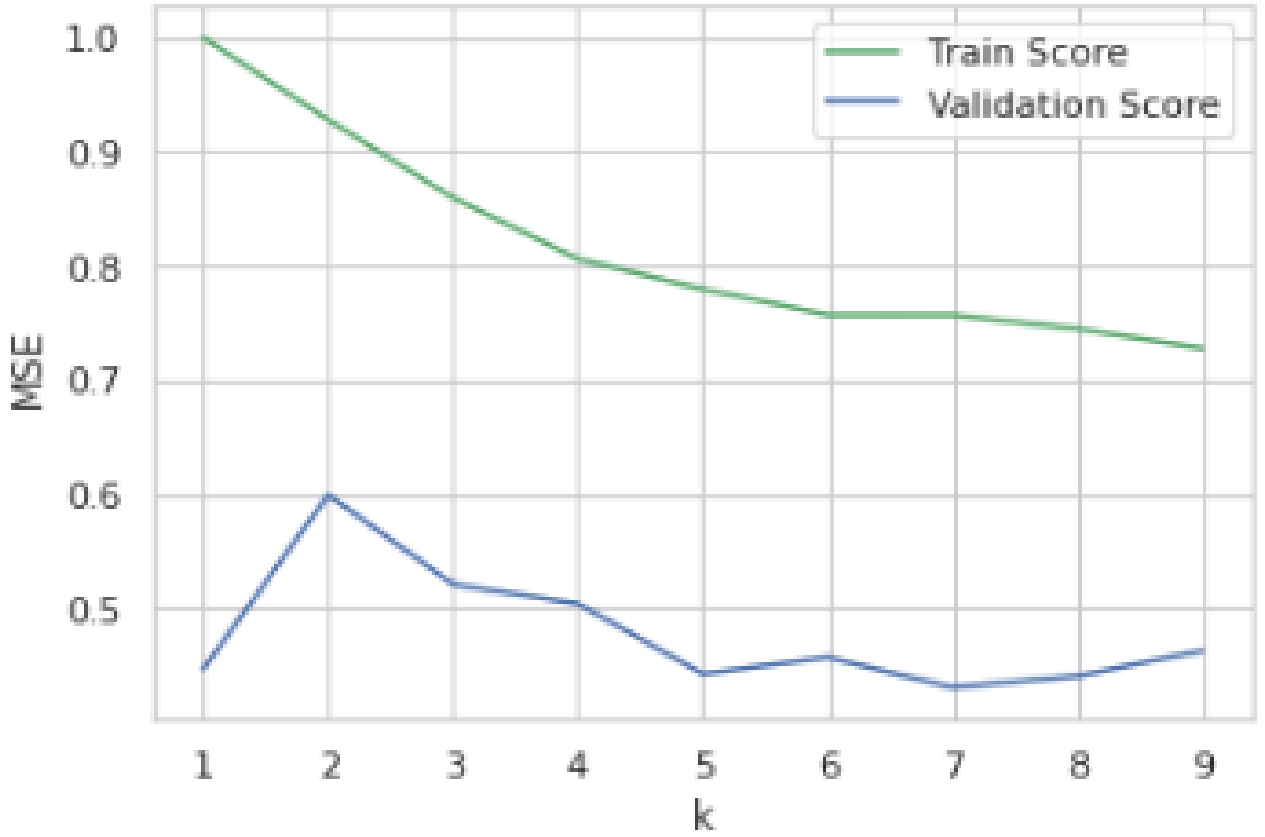


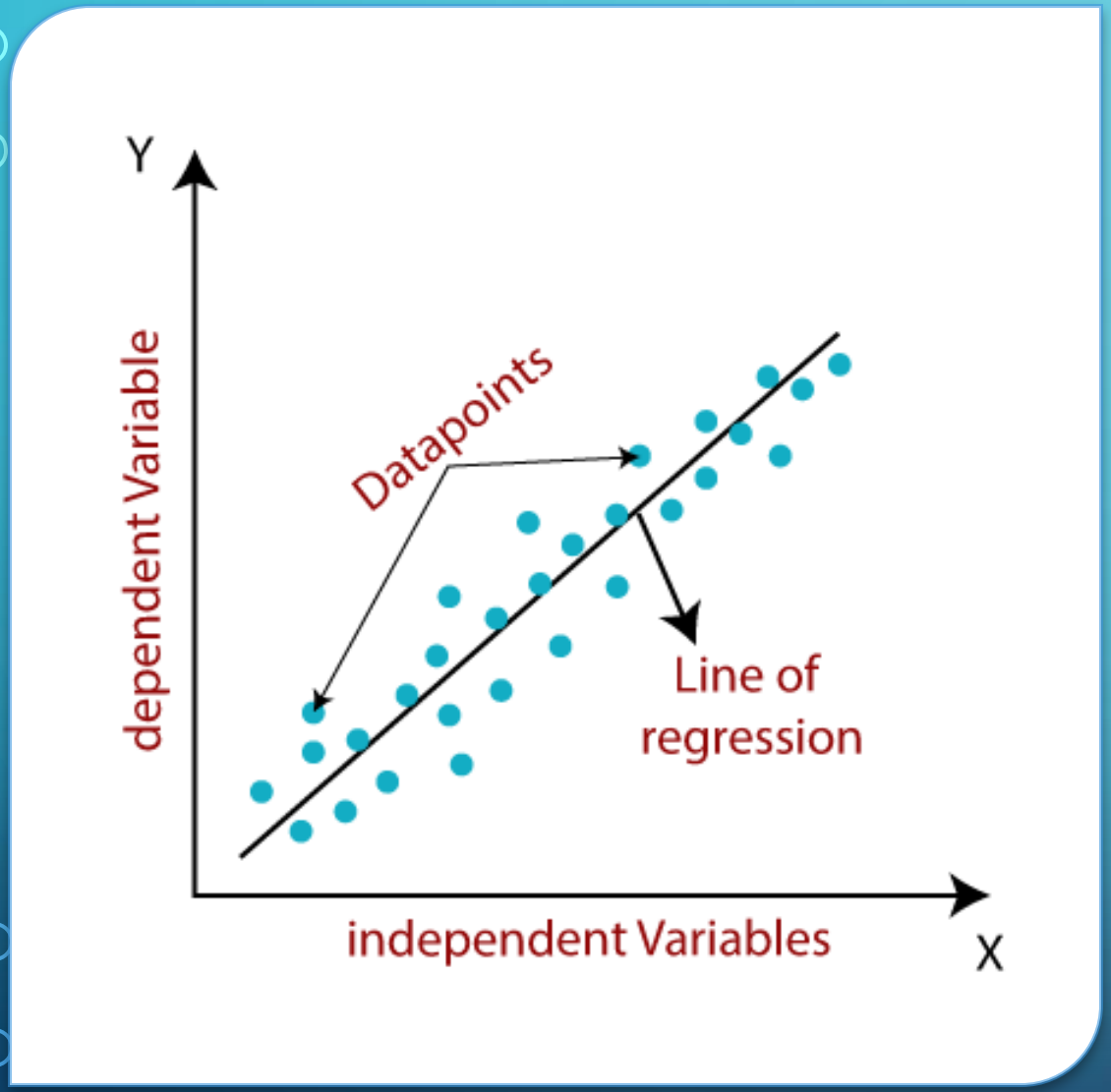
- Form of regression that models the relationship as the *nth* degree between the independent variable x and dependent variable y
- Best parameter value is 2nd degree with score of 0.98
- Grid Search CV returns 3rd degree is better, but causes overfitting

KNN REGRESSION

- To predict a new record, this method relies on finding “similar” records in the training data.
- The KNN algorithm uses ‘**feature similarity**’ to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set.

```
Best parameters: {'K': 2}
Best score: 0.60
```





LINEAR REGRESSION

- In statistics, **linear regression** is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent & independent variables).

- Average cross-validation score is 0.71

- Turned out to be better than KNN Regression

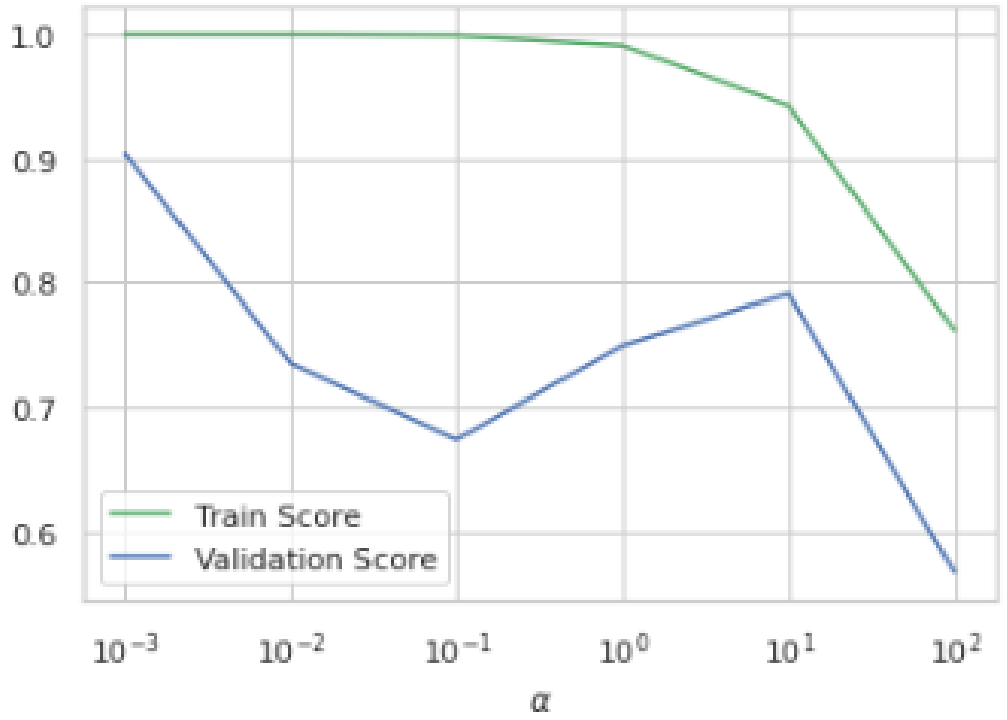
RIDGE REGRESSION

IT USES L2 REGULARIZATION TECHNIQUE.

ALPHA - PENALTY TERM.

HIGH ALPHA – HIGH PENALTY.

```
Best parameters: {'alpha': 0.001}
Best score: 0.90
```



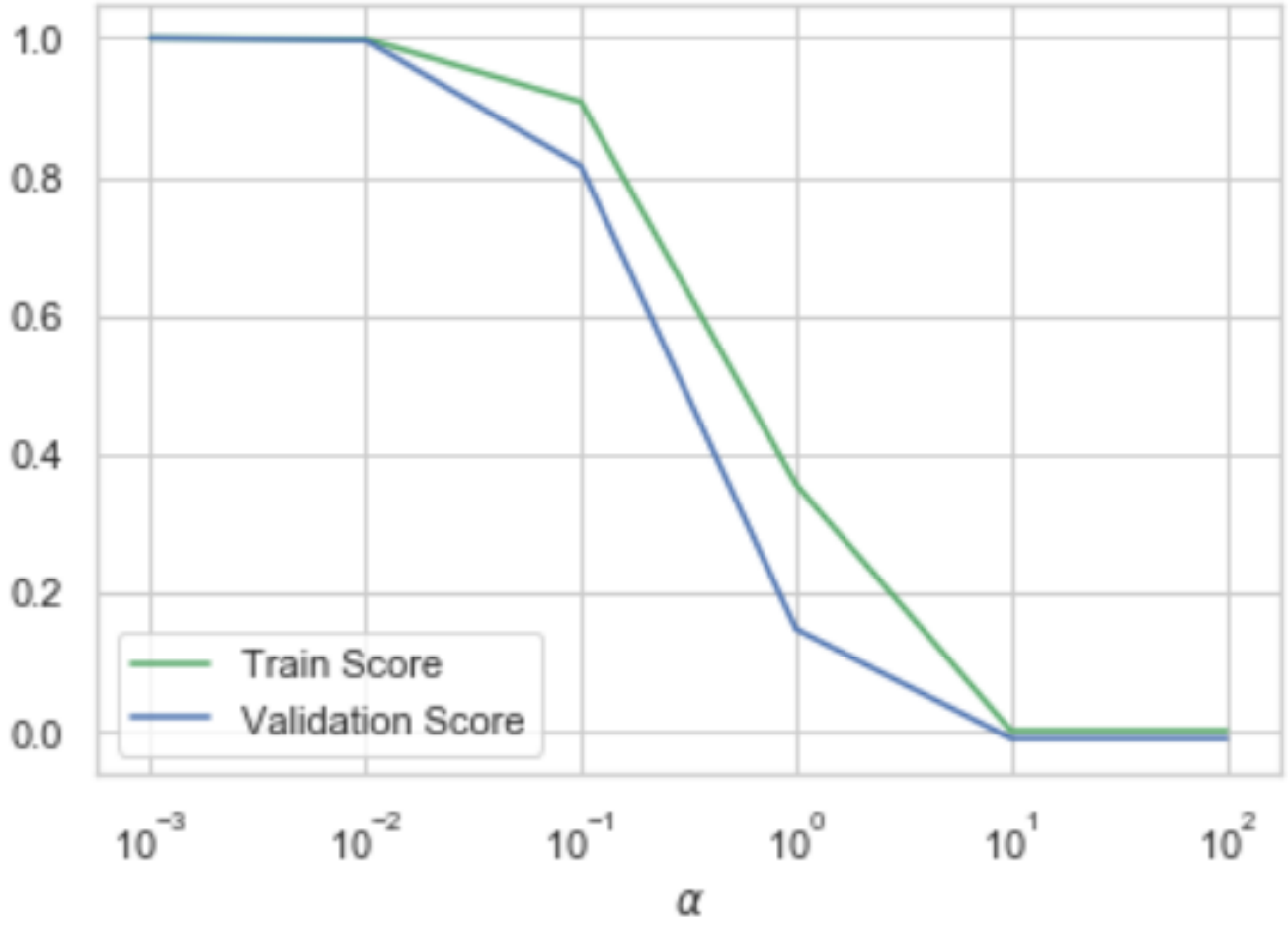
SVM WITH KERNEL TRICK

- Features having a linear or non-linear decision boundary.
- **SVM – rbf (Radial Basis Function) Best Score : 0.65, Gamma = 0.1 & C=100**
- SVM – Poly Best Score : 0.66, Degree = 2 & C=100

LASSO REGRESSION

- Shrinkage and variable selection method for linear regression models
- The best parameter value of alpha for this model is 0.001 giving a perfect score of 1.00 on the validation dataset.
- GridSearchCV gives the same alpha value as the best parameter with a score of 0.99.

```
Best score: 1.00
Best parameters: {'alpha': 0.001}
```

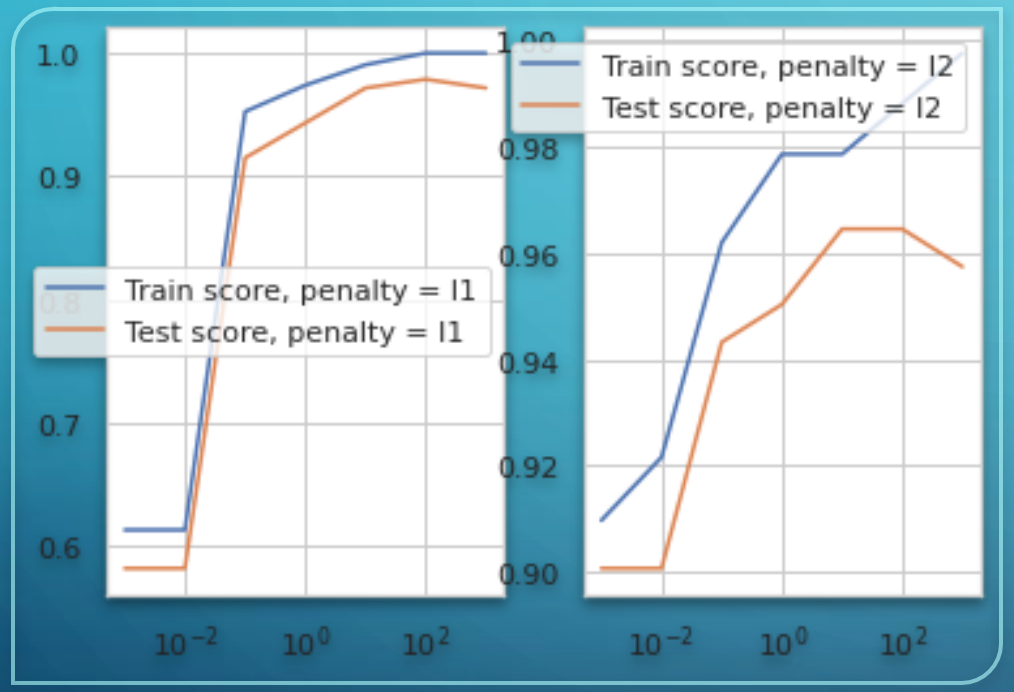


CLASSIFICATION TASK

The classification is carried out with the help of a model obtained using a learning procedure. According to the type of learning used, there are two categories of classification, one using **supervised learning** and the other using **unsupervised learning. Some of the models :**

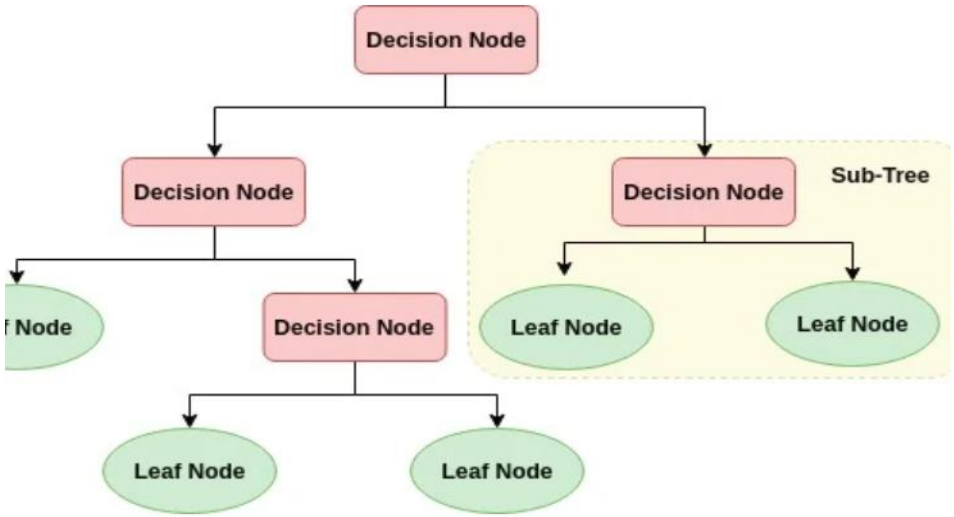
1. KNN Classification
2. Logistic Regression
3. Support Vector Machine - Linear SVC
4. Support Vector Machine with Kernel trick – Rbf, Poly
5. Decision Tree

LOGISTIC REGRESSION



- Models the possibility of an outcome based on an input
  - Binary Outcome: two values such as true/false
  - Multinomial Outcome: more than two possible outcomes
- Parameter of 100 and penalty l1 gave the best score of 0.98
- Grid Search returns the same parameter and penalty

DECISION TREE CLASSIFICATION




```
dtree_cv = DecisionTreeClassifier()
scores = cross_val_score(dtree_cv, X_trainval, y_trainval, cv = 10, scoring = 'accuracy' )
print("Cross-validation scores: {}".format(scores))

print("Average cross-validation score: {:.2f}".format(scores.mean()))
```

```
Cross-validation scores: [1.         1.         0.98214286 1.         0.98214286 1.
 1.         1.         1.         1.        ]
Average cross-validation score: 1.00
```

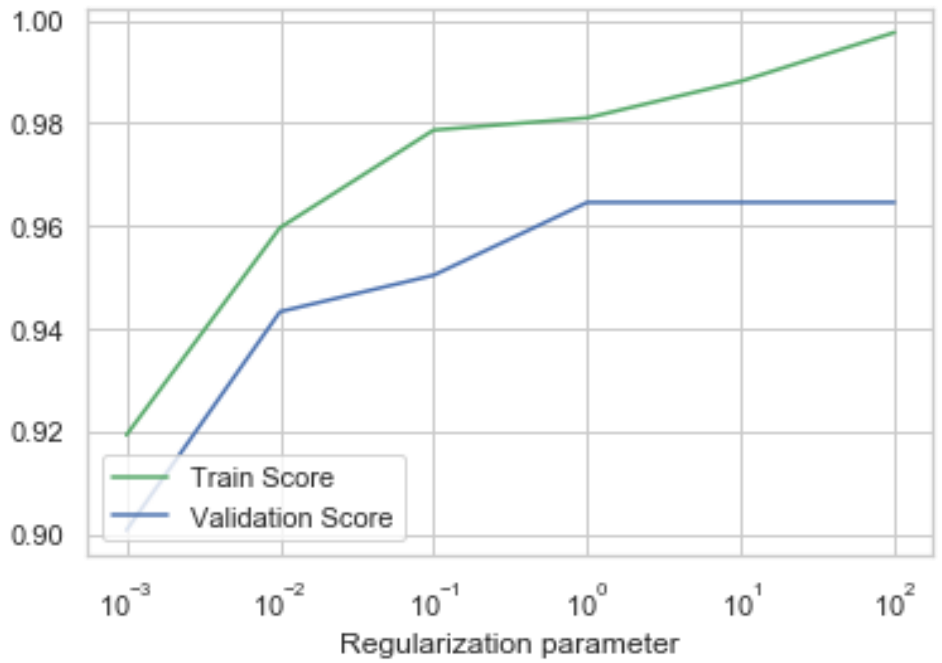
A decision tree is a flowchart-like tree structure

Accuracy on training set: 1.000
Accuracy on test set: 1.000

The average cross-validation score for this model is 1 using cross_val_score function.

Grid-Search is not required to be performed.

SUPPORT-VECTOR MACHINE



Linear SVC

- The parameter that affects is the regularization term C.

- The bigger this parameter, the less regularization is done and more features are added in the model.

Best score: 0.96

Best parameters: {'C': 1}

SVM WITH KERNEL TRICK

- way to make optimization efficient when there are features having a linear or non linear decision boundary

SVM - rbf(Radial Basis Function)

- The hyper-parameters for this model are 'gamma' and regularization term 'C'.

- Best score: 0.98

- Best parameters: {'gamma': 0.1, 'C': 100}

SVM - Poly

- This model takes into account another parameter 'degree' apart from 'gamma' and regularization term 'C'.

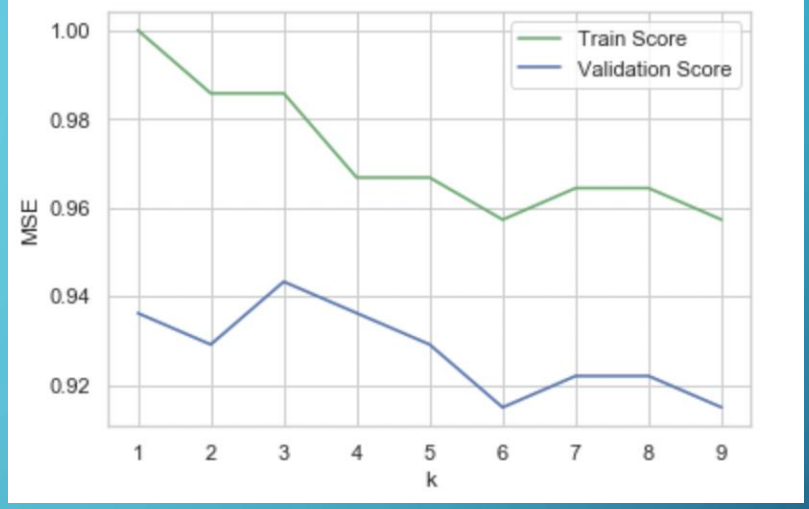
- Best score: 0.99

- Best parameters: {'degree': 1, 'C': 100, 'gamma': 100}

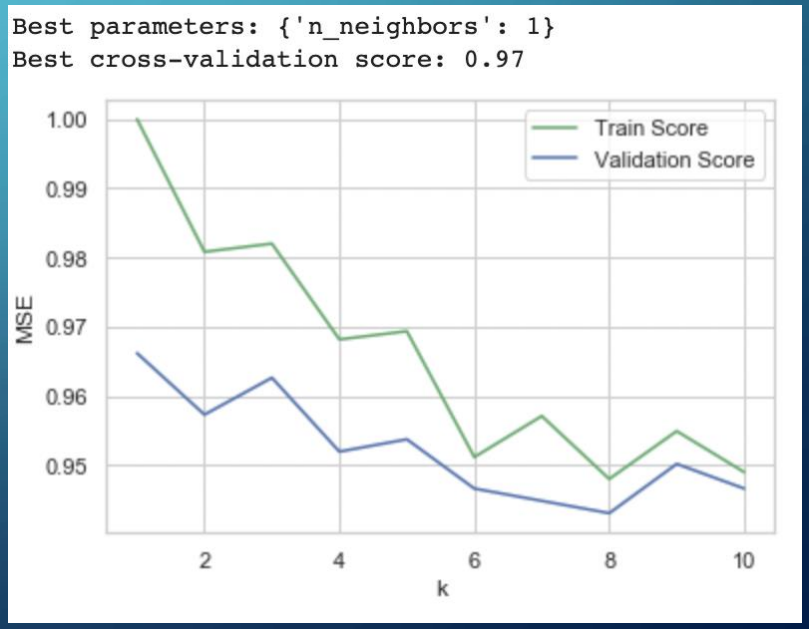
KNN CLASSIFICATION

- The idea in k-nearest-neighbors methods is to identify k records in the training dataset that are similar to a new record that we wish to classify.
- The best parameter value of K for this model is 3 which gives an accuracy of 0.94 on the validation dataset.
- The average cross-validation score for the parameter K = 3 is 0.96

```
Best score: 0.94
Best parameters: {'K': 3}
```



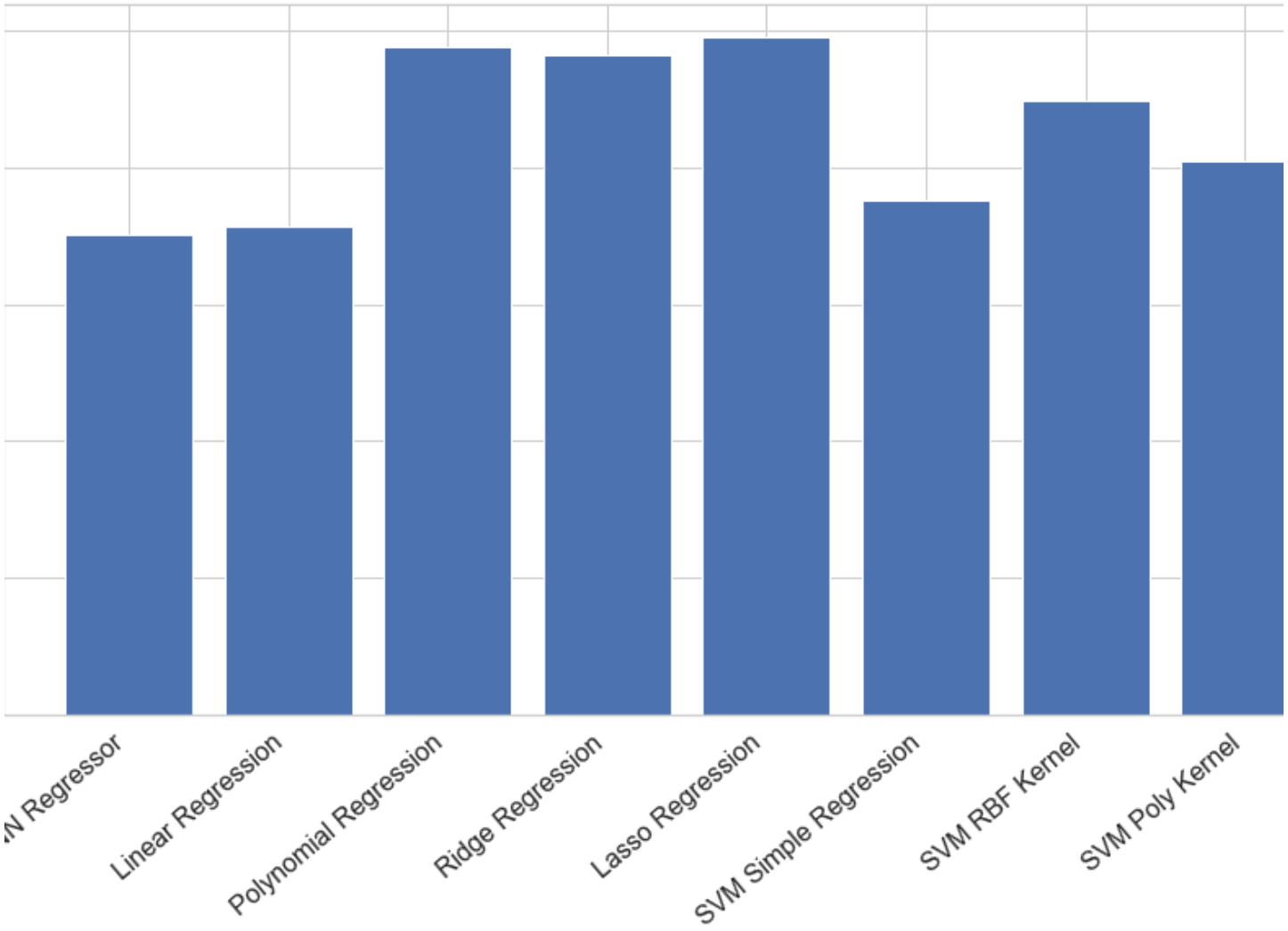

```
Best parameters: {'n_neighbors': 1}
Best cross-validation score: 0.97
```



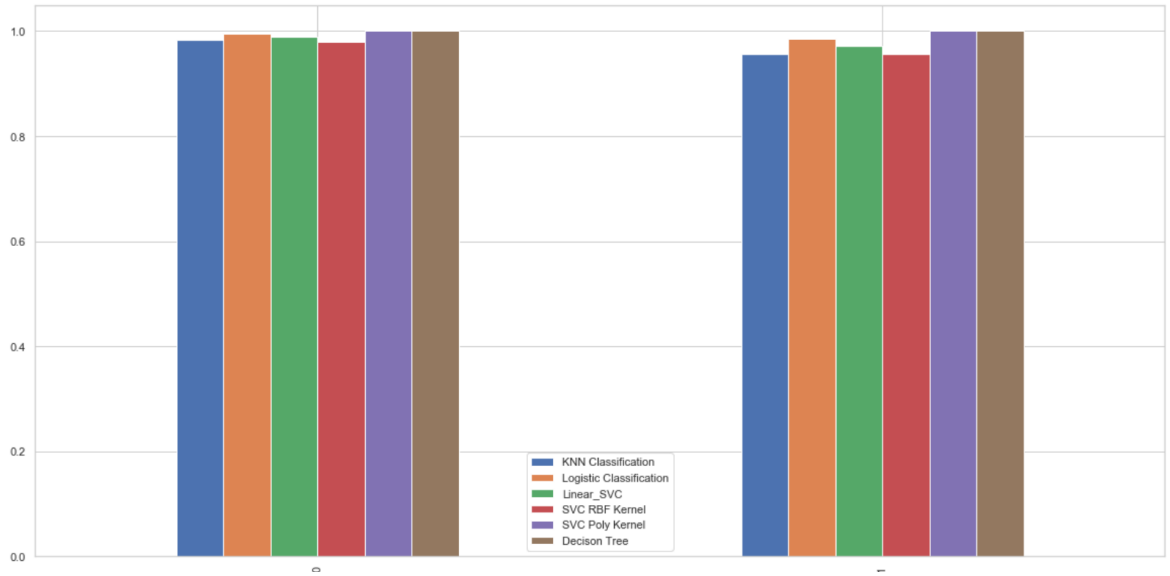
BEST REGRESSION MODEL

Lasso regression with alpha = 0.001 will be the best regressor model to predict audit_risk.

Accuracy reported is 99.86%



BEST CLASSIFIER



- Evaluation based on the having best accuracy and best recall value.
- Models that passed evaluation
  - SVM - Poly
  - Decision Tree
- What is the difference between Two algorithms?
- Which Model is more preferable?

THANK YOU
ANY QUESTIONS?