**Capstone Project**

For the capstone project I have chosen the Chennai housing sale dataset.

The sales price needs to be predicted from this dataset, SALES\_PRICE is the dependent variable. This is a regression problem and it is supervised learning

**Proprocessing**

There are 3 fields [N\_BEDROOM ,N\_BATHROOM,QS\_OVERALL]which are having null values and are handled using interpolate method

For other 6 [AREA,SALE\_COND,PARK\_FACIL,BUILDTYPE,UTILITY\_AVAIL,STREET] columns the field values were not proper, so the column values are replaced to properly

**Univariate and Bivariate**

Univariate and Bivariate analysis are done for this dataset

**Feature Selection**

The feature is selection is done using Select K and RFE algorithms

**Select K**

A screenshot of a computer

Description automatically generated

The 5 best features using Select K are ['INT\_SQFT', 'DIST\_MAINROAD', 'REG\_FEE', 'COMMIS', 'AREA\_T Nagar']

**RFE**

**A screenshot of a computer

Description automatically generated**

**The random forest algorithm gives the best score**

**Model Creation**

Linear Regression

A screenshot of a computer program

Description automatically generated

The R score using linear regression is **0.963**

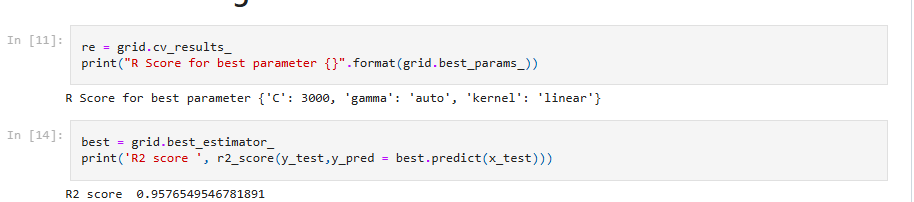
**Decision Tree**

A screenshot of a computer code

Description automatically generated

The R score using Decision tree is **0.959** and the best parameter **is {'criterion': 'friedman\_mse', 'max\_features': None, 'splitter': 'random'}**

**SVM**



The R score using SVM is **0.957** and the best parameter **is *{'C': 3000, 'gamma': 'auto', 'kernel': 'linear'}***

**Random forest**

A screenshot of a computer code

Description automatically generated

The R score using Random forest algorithm is **0.979** and the best parameter is **{'criterion': 'squared\_error', 'max\_features': 'sqrt', 'n\_estimators': 100}**

**Final Model**

The best model is **Random forest** for the Chennai housing sale dataset as the R score is higher for this model