

**Phase -2**

**Student Name: M. Kaviya**

**Register Number:620123106051**

**Institution:Avsengineeringcollege**

**Department: E C E**

**Date of Submission**:10/05/2025

**Github Repository link:**

[**https://github.com/kaviyaM24/Kaviya-M-naan-**](https://github.com/kaviyaM24/Kaviya-M-naan-Mudhlavan-project.git)

[**Mudhlavan-project.git**](https://github.com/kaviyaM24/Kaviya-M-naan-Mudhlavan-project.git)

**1.Problem statement**

●*Accurately forecasting house prices is crucial for buyers, sellers, and real estateinvestorsto make informed financial decisions. The challenge liesin capturingthecomplex,non-linearrelationshipsamongnumerousvariables like location, size, amenities, and economic conditions.*

●*TypeofProblem:Regression(predictingacontinuousvariablehouseprice).*

●*Why It Matters: Enhances decision-making in real estate markets, supports*  *financialinstitutionsinloanprocessing,andaidsurbanplanninginitiatives.*

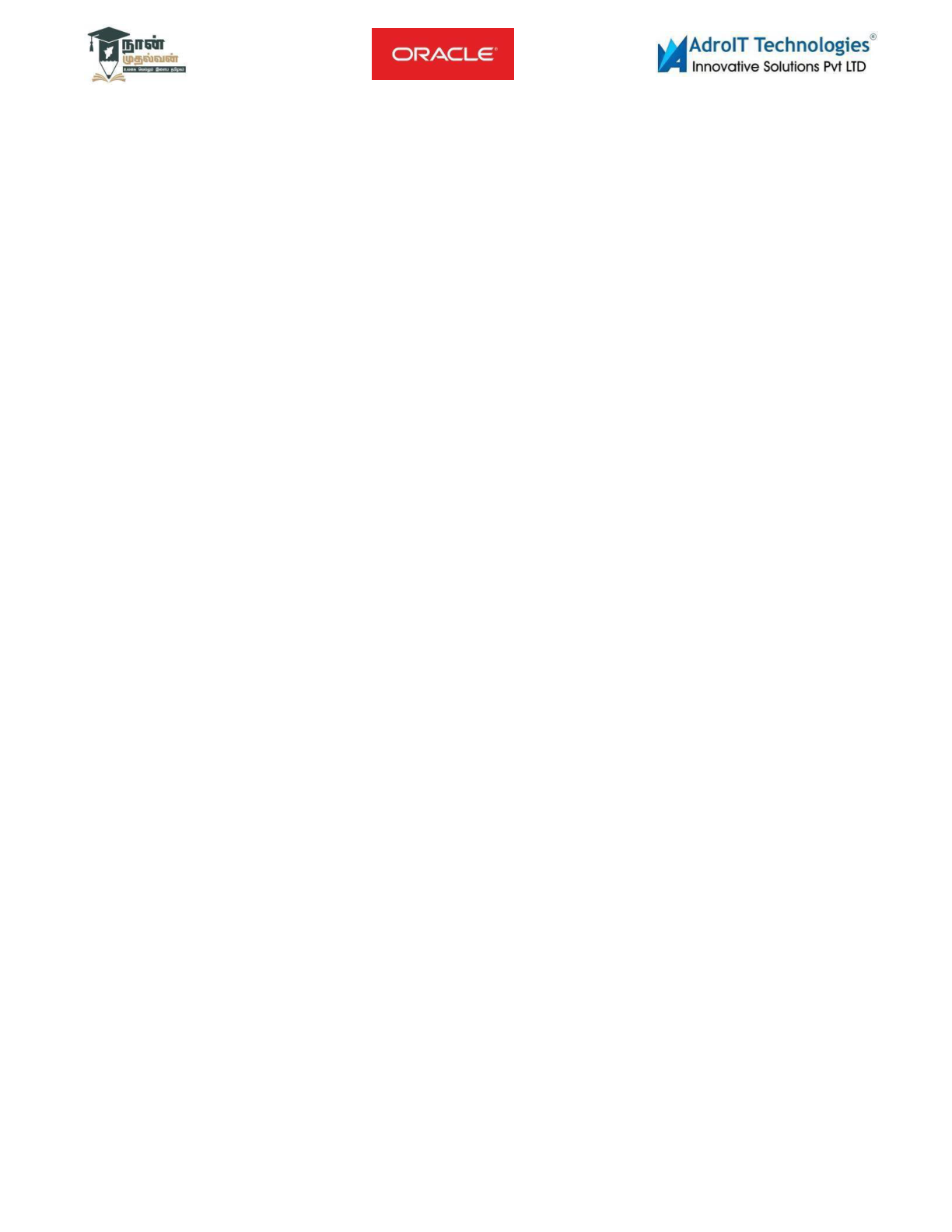
**2. ProjectObjectives**

●*PrimaryGoal:Developarobust,interpretable,andaccurateregression*  *model for house price prediction.*

●*TechnicalObjectives:*

●*Analyzethedatasettoidentifysignificantpredictors.*

●*Comparemultipleregressiontechniques.*



●*Optimizeperformanceusingfeatureengineeringandhyperparametertuning.*

●*Updated Goal: After initial EDA, emphasis shifted to improving model*  *interpretabilitywhileretainingaccuracyduetomulticollinearityinfeatures.*

●*Assess model fairness and bias, ensuring that the model does not*  *systematicallyunderpredictoroverpredictbasedonlocationorhousetype.*

●*UpdatedFocus:AfterinitialEDA,emphasisshiftedtowardimprovingmodel interpretability while maintaining accuracy, due to multicollinearity observed among features.*

**3. FlowchartoftheProjectWorkflow**

1. DataCollection   
 ↓  
2. DataPreprocessing   
 ↓  
3. ExploratoryDataAnalysis(EDA)   
 ↓  
4. FeatureEngineering   
 ↓  
5. ModelBuilding   
 ↓  
6. Evaluation   
 ↓  
7. Visualization   
 ↓  
8. Conclusion

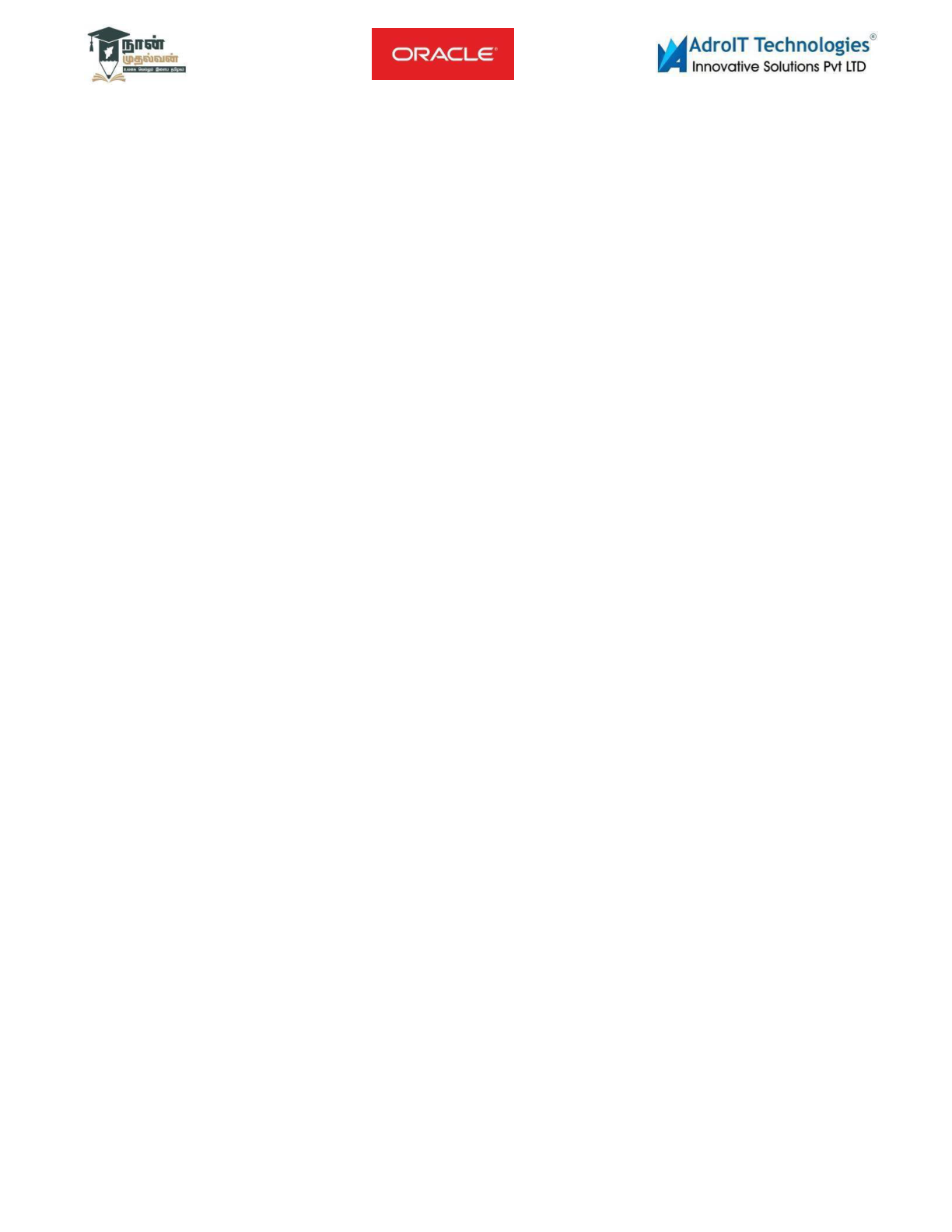
**4. DataDescription**

●*Source:Kaggle–HousePrices:AdvancedRegressionTechniques*

●*Type:Structureddata(tabular)*

●*Records&Features:~1460rows,80+features*

●*DatasetNature:Static*



●*TargetVariable: SalePrice*

**5. DataPreprocessing**

●*Handledmissingvaluesusingmean/medianordomain-specificlogic.*

●*Removedduplicaterecordsandverifieduniqueidentifiers.*

●*DetectedoutliersusingIQRandvisualmethods(boxplots).*

●*Convertedcategoricalcolumnstonumericalusingone-hotencoding.*

●*StandardizednumericfeaturesusingStandardScaler.*

●*Ensureddatatypeswereconsistentacrosscolumns.*

**6. ExploratoryDataAnalysis(EDA)**

●*UnivariateAnalysis:*

○*Usedhistogramsandboxplotsfornumericfeatures.*

●*Bivariate/MultivariateAnalysis:*

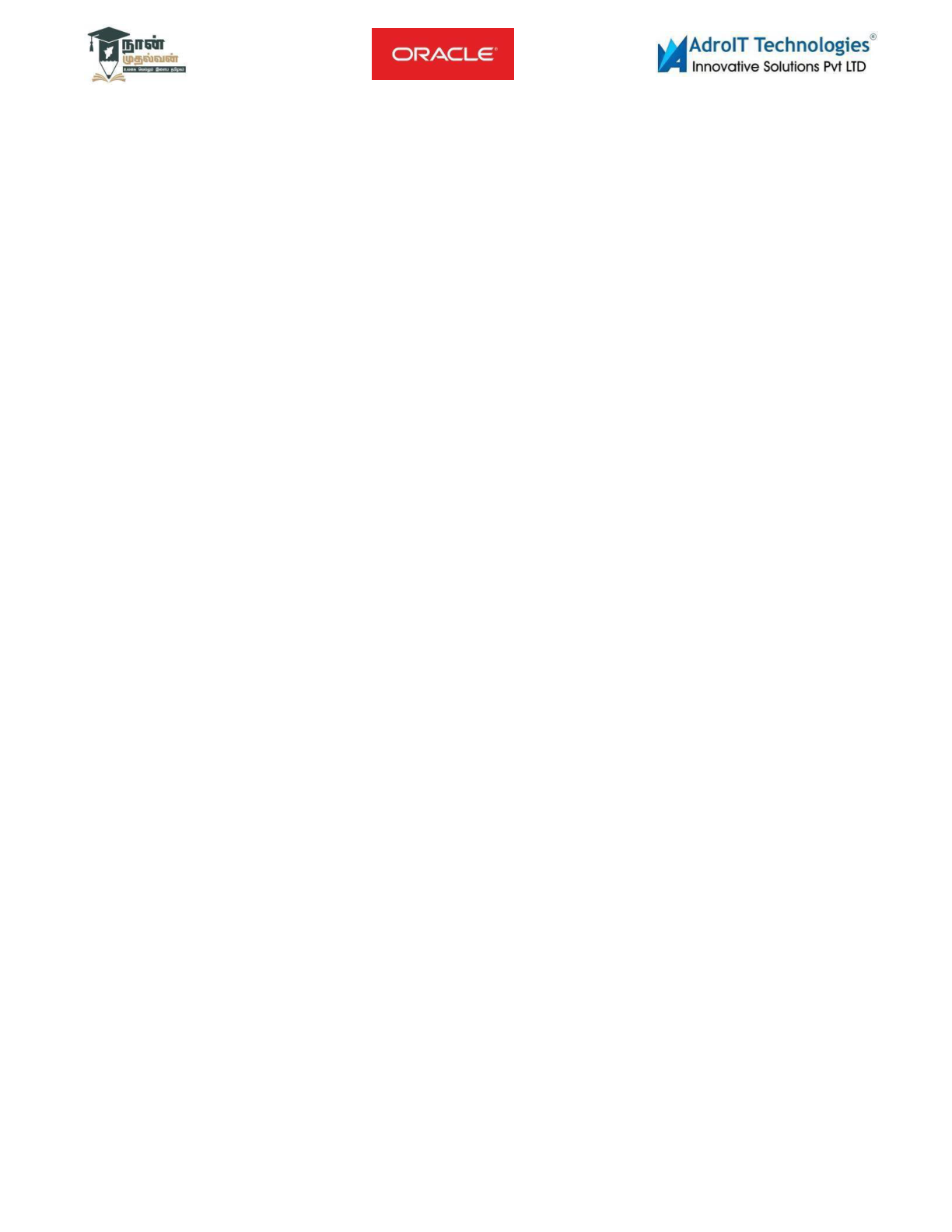
○*Correlationmatrixandpairplotsforkeyvariablesvs.SalePrice.*

●*InsightsSummary:*

○*StrongpositivecorrelationwithfeatureslikeOverallQual,*

*GrLivArea.*

○*Location(Neighborhood)playsamajorrole.*



○*Somefeaturesarehighlyskewedandneedtransformation.*

**7. FeatureEngineering**

●*Creatednewfeaturessuchas“TotalBathrooms”,“HouseAge”,and“Is*  *Remodeled”.*

●*Appliedlogtransformationonskewedfeatures.*

●*Binnedcontinuousvariables(e.g.,YearBuiltintodecades).*

●*Removedfeatureswithhighcollinearorlowvariance.*

●*Createdinteractionterms(e.g.,OverallQual\*GrLivArea)tocapture*  *combined effects.*

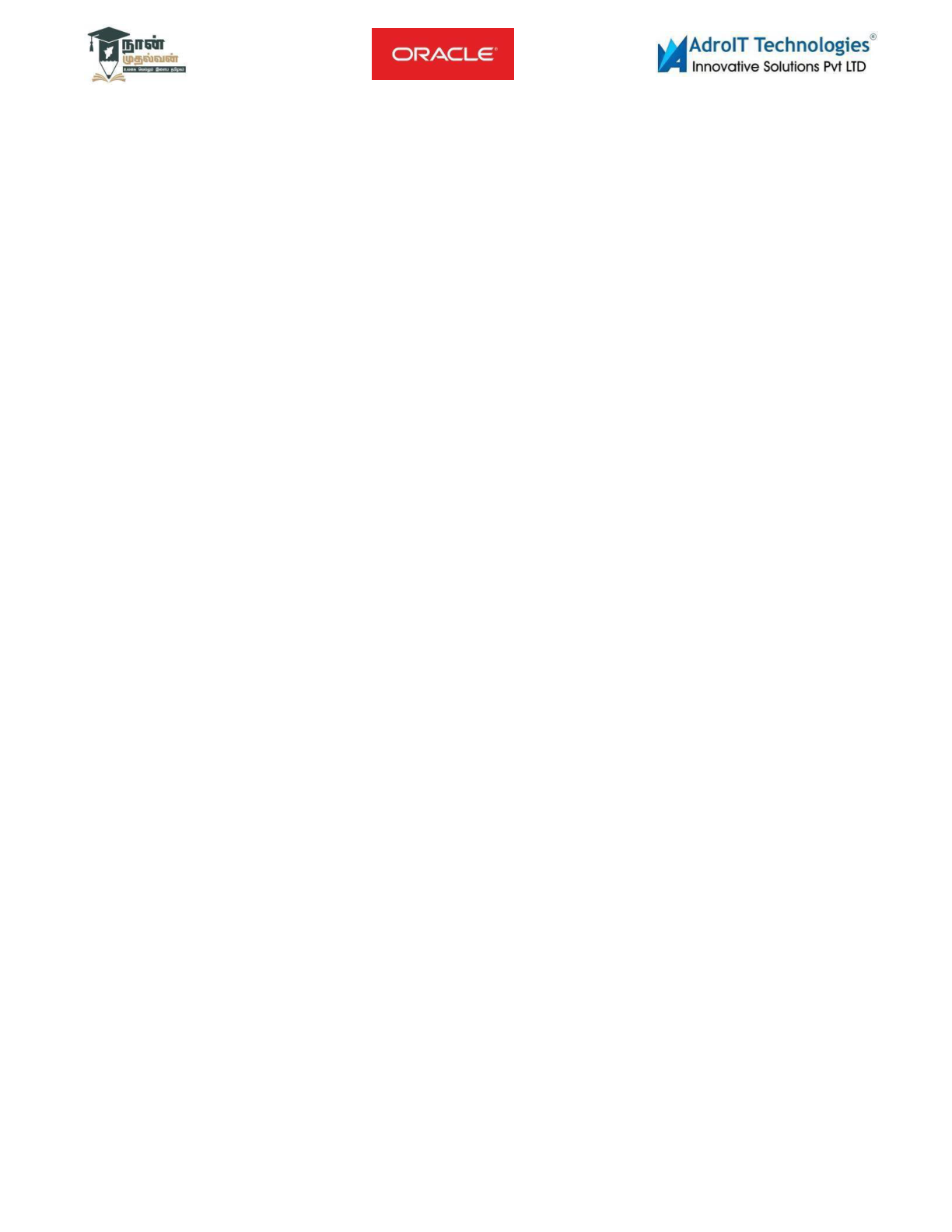
●*IntroducedpolynomialfeaturesforimportantvariableslikeGrLivAreato*  *model non-linear patterns.*

**8. ModelBuilding**

●*ChoiceofModels:SelectedLinearRegressionforbaselineinterpretability and Random Forest Regressor for handling non-linear relationships and feature interactions.*

●*Data Split: Divided the dataset into 80% training and 20% testing sets to evaluatemodelgeneralizationperformance.Stratificationwasnotrequired for continuous target variables.*

●*EvaluationMetrics:UsedMAE,RMSE,andR²Scoretoobjectivelycompare*  *models' accuracy and reliability for regression tasks.*



●*Performance Observation: Random Forest outperformed Linear Regression*

*byachievinglowererrorvaluesandahigherR²score,showingbetterability to*

*model complex patterns in the data.*

○ *E.g.,LogisticRegression,DecisionTree,RandomForest,KNN,etc.*

●*Appliedcross-validation(k-fold)tovalidatetherobustnessofmodel*

*performance.*

●*PerformedGridSearchforHyperparameterTuning(e.g.,tuningnumberof*

*trees, max depth in Random Forest).*

**9. VisualizationofResults&ModelInsights**

●*FeatureImportancePlot:IdentifiedtoppredictorslikeOverallQual,*

*GrLivArea, and GarageCars.*

●*ResidualPlots:RandomForesthadmoreuniformlydistributedresiduals.*

●*ModelComparison:VisualizedperformanceusingbarplotsforRMSEand R².*

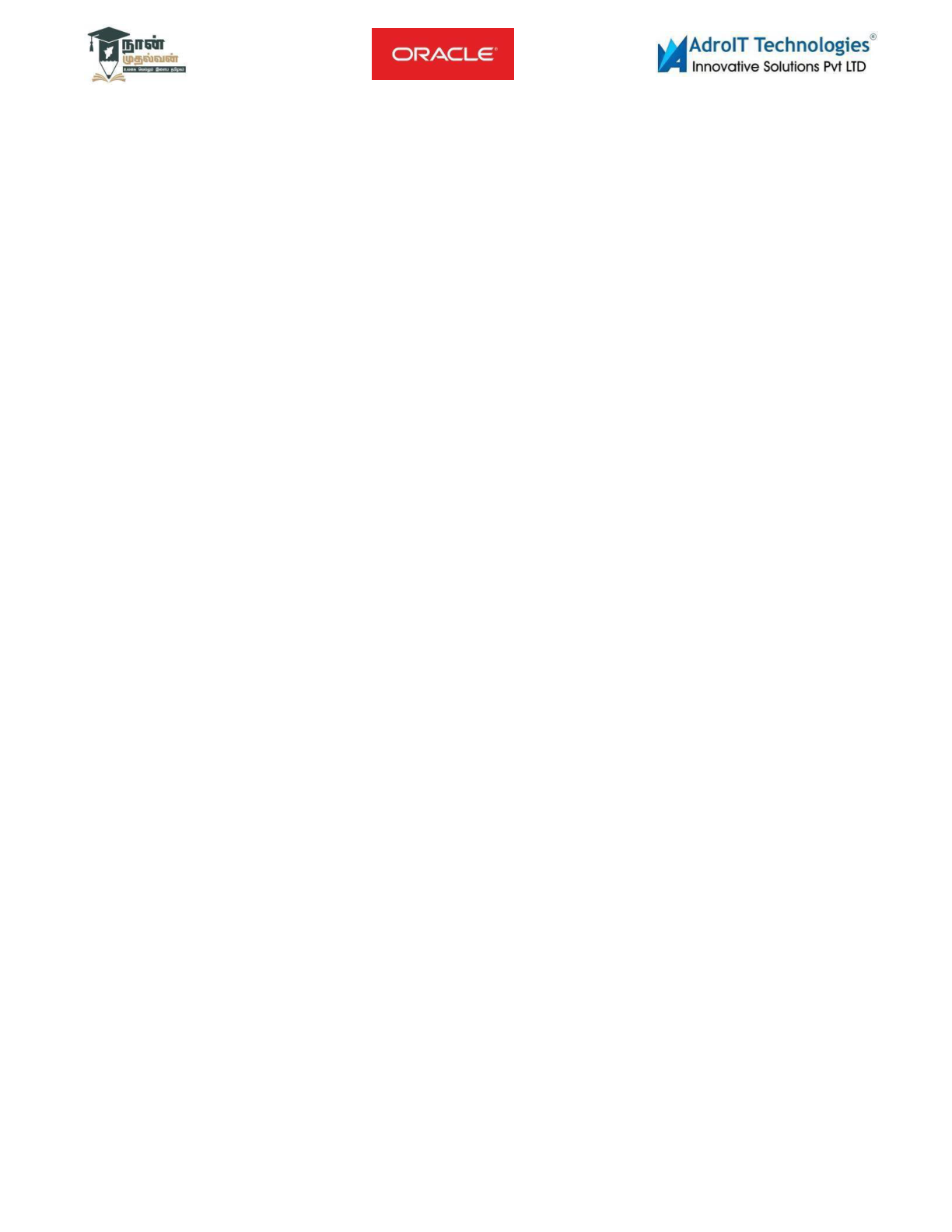
●*ActualvsPredictedPlot:*

*°ScatterplotcomparingpredictedSalePricevs.actualSalePrice, highlighting*

*model accuracy visually.*

●*DistributionofPredictionErrors:*

*°Plottedhistogramofresidualstocheckforanybiasorskewnessin predictions*



**10. ToolsandTechnologiesUsed**

●*Language: Python*

●*IDE:GoogleColab*

●*Libraries:pandas,numpy,matplotlib,seaborn,scikit-learn,XGBoost*

●*VisualizationTools:matplotlib,seaborn,Plotly*

**11. TeamMembersandContributions**

1) J.Ayishabanu:Datacleaninganddocumentation.

2) S.Anusiya:EDAandproblemobjective.

3) M.Dharani:Featureengineeringandreporting.

4) M.Kaviya: Modeldevelopmentandvisualizationof

results.