This report addresses the problem statement regarding the prediction of property values for a given dataset, which includes two properties and various controllable parameters across different systems. A detailed exploratory data analysis was conducted, including data cleaning and conversion of categorical data into numerical form, pairwise correlation to identify dependencies between features, and principal component analysis to reduce dimensionality and identify influential features.

For Property 1, various regression techniques were applied:

- Linear Regression
- Multilinear Regression
- Principal Component Regression
- Support Vector Regression

For Property 2, classification models were considered:

- Logistic Regression
- Support Vector Machine

The models were compared based on accuracy, robustness, sensitivity, specificity, and generalizability when applied to an unseen test set. The report also explored the models' sensitivity to parameter variations.

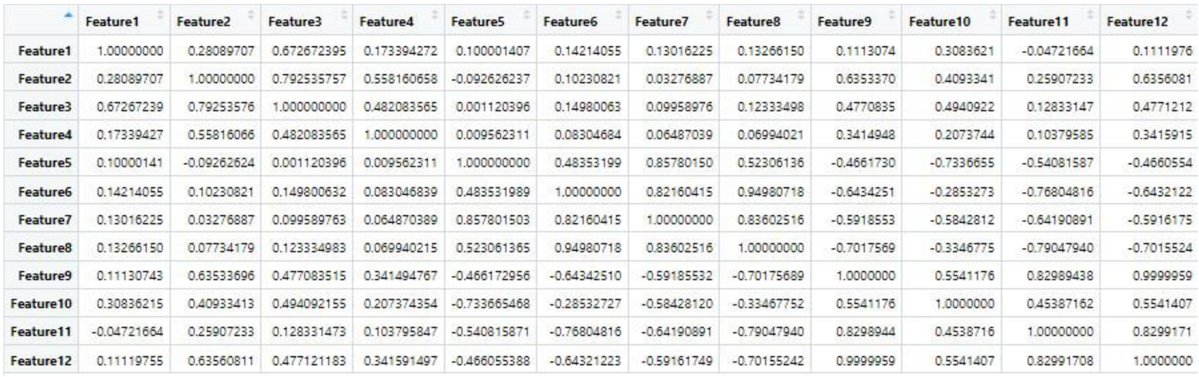
The final selection of models was justified based on their statistical performance metrics:

For Property 1, Support Vector Regression was chosen due to its high R-squared value and low RMSE, indicating a strong fit to the data.

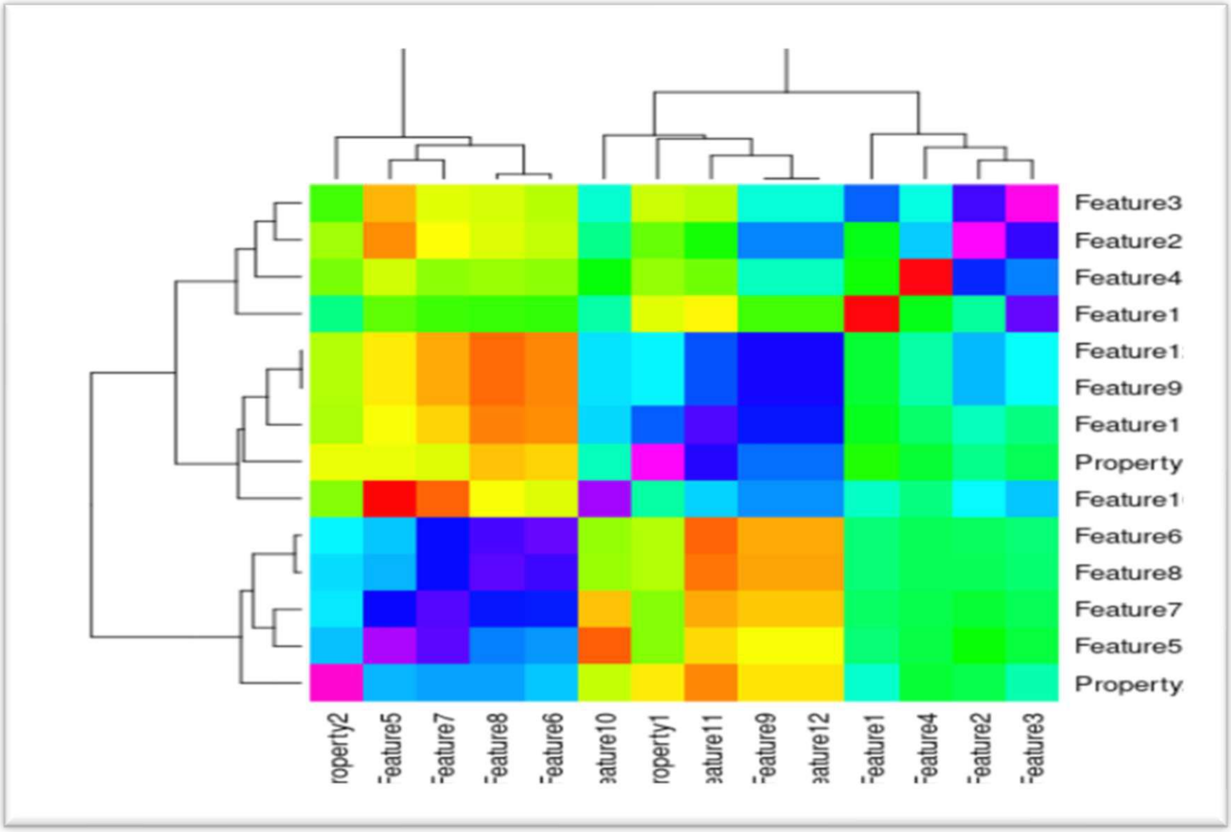
For Property 2, Logistic Regression was preferred for its balanced prediction capabilities and strong sensitivity, suggesting reliable performance and potential for generalization.

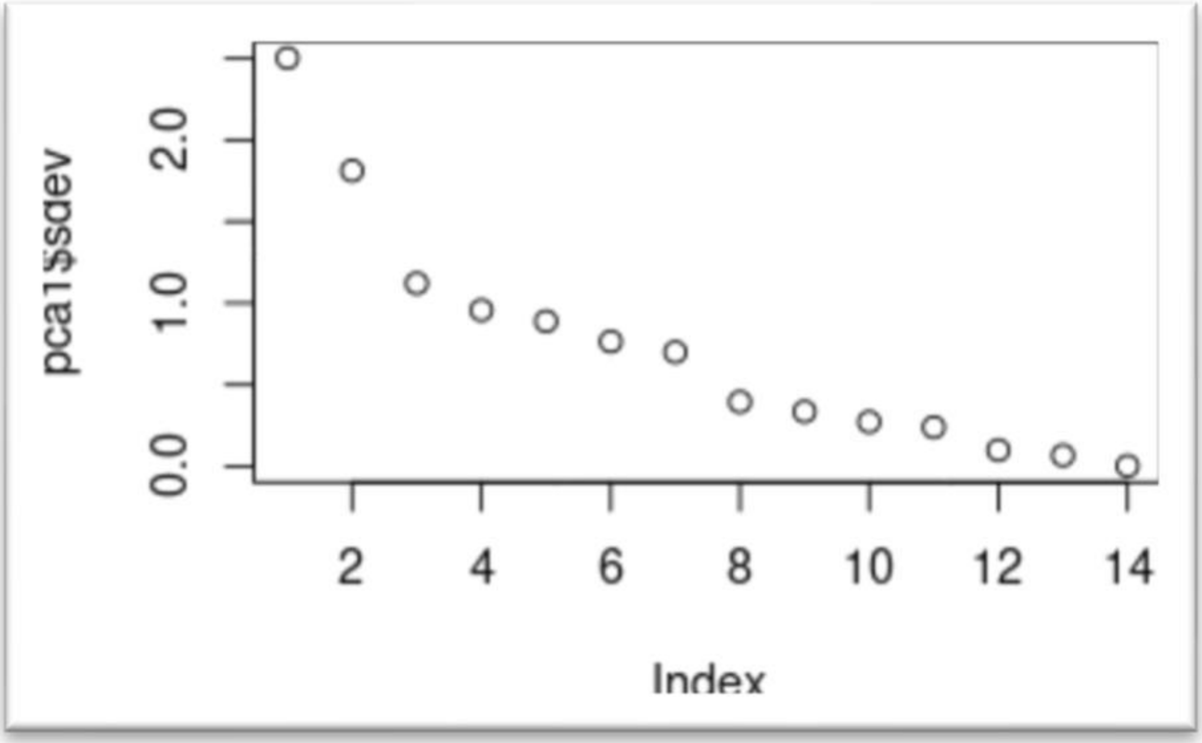
The report concludes with confidence in the chosen models for deployment and provides predicted values for the first ten rows of data where property measurements were previously unknown. The chosen methods outperformed others due to their mathematical robustness and the ability to handle the dataset's characteristics, leading to more reliable and generalizable models.

EDA

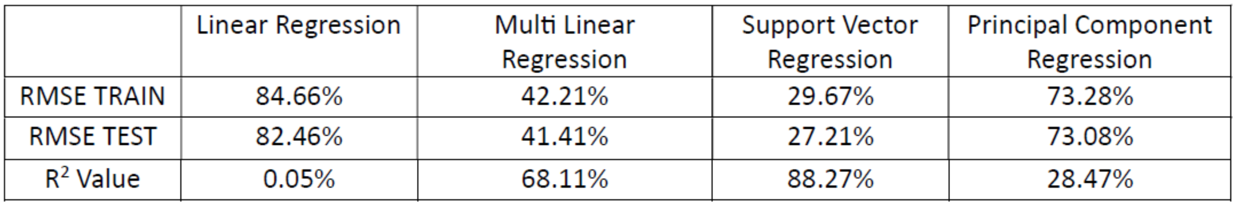


	Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7	Feature8	Feature9	Feature10	Feature11	Feature12
Feature1	1.00000000	0.28089707	0.672672395	0.173394272	0.100001407	0.14214055	0.13016225	0.13266150	0.1113074	0.3083621	-0.04721664	0.1111976
Feature2	0.28089707	1.00000000	0.792535757	0.558160658	-0.092626237	0.10230821	0.03276887	0.07734179	0.6353370	0.4093341	0.25907233	0.6356081
Feature3	0.67267239	0.79253576	1.000000000	0.482083565	0.001120396	0.14980063	0.09958976	0.12333498	0.4770835	0.4940922	0.12833147	0.4771212
Feature4	0.17339427	0.55816066	0.482083565	1.000000000	0.009562311	0.08304684	0.06487039	0.06994021	0.3414948	0.2073744	0.10379585	0.3415915
Feature5	0.10000141	-0.09262624	0.001120396	0.009562311	1.000000000	0.48353199	0.85780150	0.52306136	-0.4661730	-0.7336655	-0.54081587	-0.4660554
Feature6	0.14214055	0.10230821	0.149800632	0.083046839	0.483531989	1.00000000	0.82160415	0.94980718	-0.6434251	-0.2853273	-0.76804816	-0.6432122
Feature7	0.13016225	0.03276887	0.099589763	0.064870389	0.857801503	0.82160415	1.00000000	0.83602516	-0.5918553	-0.5842812	-0.64190891	-0.5916175
Feature8	0.13266150	0.07734179	0.123334983	0.069940215	0.523061365	0.94980718	0.83602516	1.00000000	-0.7017569	-0.3346775	-0.79047940	-0.7015524
Feature9	0.11130743	0.63533696	0.477083515	0.341494767	-0.466172956	-0.64342510	-0.59185532	-0.70175689	1.0000000	0.5541176	0.82989438	0.9999959
Feature10	0.30836215	0.40933413	0.494092155	0.207374354	-0.733665468	-0.28532727	-0.58428120	-0.33467752	0.5541176	1.0000000	0.45387162	0.5541407
Feature11	-0.04721664	0.25907233	0.128331473	0.103795847	-0.540815871	-0.76804816	-0.64190891	-0.79047940	0.8298944	0.4538716	1.00000000	0.8299171
Feature12	0.11119755	0.63560811	0.477121183	0.341591497	-0.466055388	-0.64321223	-0.59161749	-0.70155242	0.9999959	0.5541407	0.82991708	1.0000000


Feature3
Feature2
Feature4
Feature1
Feature1
Feature9
Feature1
Property
Feature1
Feature6
Feature8
Feature7
Feature5
Property
roperty2
Feature5
Feature7
Feature8
Feature6
eature10
roperty1
eature11
Feature9
eature12
Feature1
Feature4
Feature2
Feature3


pca1$sdev
0.0
1.0
2.0
2
4
6
8
10
12
14
Index

The following table represents the comparison of each model in terms of accuracy with trained and unseen test data set.



	Linear Regression	Multi Linear Regression	Support Vector Regression	Principal Component Regression
RMSE TRAIN	84.66%	42.21%	29.67%	73.28%
RMSE TEST	82.46%	41.41%	27.21%	73.08%
R^2 Value	0.05%	68.11%	88.27%	28.47%

<u>Support Vector Machine:</u>

Predicted as class 0 (Negative class):

TN (True Negatives): 39 instances were correctly predicted as class zero.

FN (False Negatives): 3 instances were incorrectly predicted as class zero when they belonged to class 1.

Predicted as class 1 (Positive class):

FP (False Positives): 56 instances were incorrectly predicted as class 1 when they belonged to class 0.

TP (True Positives): 166 instances were correctly predicted as class 1.

For test data set:

Predicted as class 0 (Negative class):

TN (True Negatives): 7 instances were correctly predicted as class zero.

FN (False Negatives): no instances were incorrectly predicted as class zero when they belonged to class 1.

Predicted as class 1 (Positive class):

FP (False Positives): 13 instances were incorrectly predicted as class 1

Logistic Regression:

For training data set:

Predicted as class 0 (Negative class):
TN (True Negatives): 59 instances were correctly predicted as class zero.
FN (False Negatives): 5 instances were incorrectly predicted as class zero when they belonged to class 1.

Predicted as class 1 (Positive class):
FP (False Positives): 26 instances were incorrectly predicted as class 1 when they belonged to class 0.
TP (True Positives): 174 instances were correctly predicted as class 1.

For test data set:

Predicted as class 0 (Negative class):
TN (True Negatives): 22 instances were correctly predicted as class zero.
FN (False Negatives): 1 instance was incorrectly predicted as class zero when they belonged to class 1.

Predicted as class 1 (Positive class):
FP (False Positives): 8 instances were incorrectly predicted as class 1