May 2025

PREDICTION OF GLOBAL SUICIDE RATES

University of Yalova

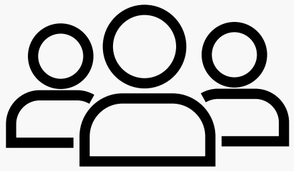
Faculty of Engineering

Computer Engineering Department

Maya Yagan

More than **720 000** people die due to suicide every year.

Over **one in every 100** deaths in 2019 were the result of suicide



The global suicide rate is over **twice as high** among men than women

INTRODUCTION

Suicide Statistics

Age Groups

- Suicide is the third leading cause of death among **15–29-year**-olds.
- Over half (58%) of all deaths by suicide occur **before the age of 50 years old.**

Regions

- Suicide occurs across all regions in the world, however, **over three quarters** (77%) of global suicides in 2019 occurred in **low and middle income** countries.

Mental Health

- An individual with depression is **twenty times** more likely to die by suicide than someone without the disorder.

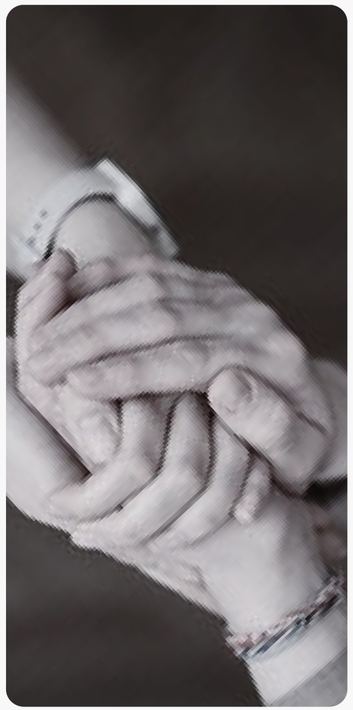
Identifying key risk factors and predicting suicide rates can support early intervention and informed policymaking.

- Traditional statistical methods often fall short with complex, non-linear relationships between socioeconomic and demographic factors.

- Machine Learning (ML) offers a powerful approach to model these relationships and improve prediction accuracy.

Objective
of this study

To apply and compare various ML models, enhanced by feature selection techniques (GA and PCA), to **predict suicide rates across countries** using socioeconomic and demographic data.



DATASET 1

Suicide Rates Overview from Kaggle

Features:

1. Country
2. Age group
3. Gender
4. Year
5. GDP per year
6. GDP per capita
7. Generation

Target: **Suicide rate per 100 000 population**

Years Covered: **1985 - 2021**

Rows Count: **31 Thousand**

This dataset lacked key mental health and socioeconomic indicators such as depression rates, alcohol consumption, or unemployment. Furthermore, inconsistencies and missing values limited its predictive strength, prompting the need for a more comprehensive dataset.

DATASET 2

Custom made dataset

Features:

1. Country
2. Age group
3. Gender
4. Year
5. GDP per capita
6. HDI
7. Depression rate
8. Alcohol Consumption
9. Unemployment rate
10. Life expectancy

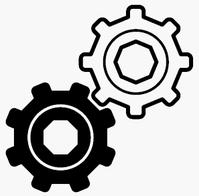
Target: **Suicide rate per 100 000 population**

Years Covered: **2000 - 2019**

Rows Count: **84 Thousand**

While combining data from different sources led to a shorter time frame due to gaps in availability, the dataset provided a more holistic view of potential predictors of suicide rates.

Preprocessing



Standardization

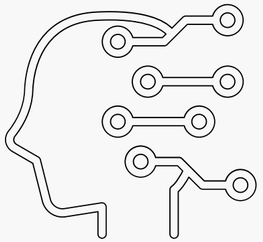
applied to numeric features for models sensitive to scale (KNN, MLP, SVM)

- Used **StandardScaler (Z-score normalization)**
  - Ensures features have zero mean and unit variance

Categorical Feature Encoding:

- Gender → **Label Encoding** (binary)
- Age Group → **Ordinal Encoding** (reflects natural order)
- Generation → **One-Hot Encoding** (6 binary columns)
- Country → **Target Encoding** using nested CV
  - Avoids data leakage
  - Captures country-level average suicide rates from training folds

Used Machine Learning Algorithms



1. K-Nearest neighbors (KNN)
2. Random forest
3. Decision tree
4. Multilayer perceptron (MLP)
5. Linear regression
6. Ridge regression
7. Support vector regression (Linear kernel)
8. Support vector regression (Polynomial kernel)
9. Support vector regression (RBF kernel)

Performance metrics

Used Evaluation Metrics

- **MAE (Mean Absolute Error):**

Average of the absolute differences between predictions and actual values. Measures overall accuracy.

- MSE (Mean Squared Error):

Average of squared prediction errors. Penalizes larger errors more than MAE.

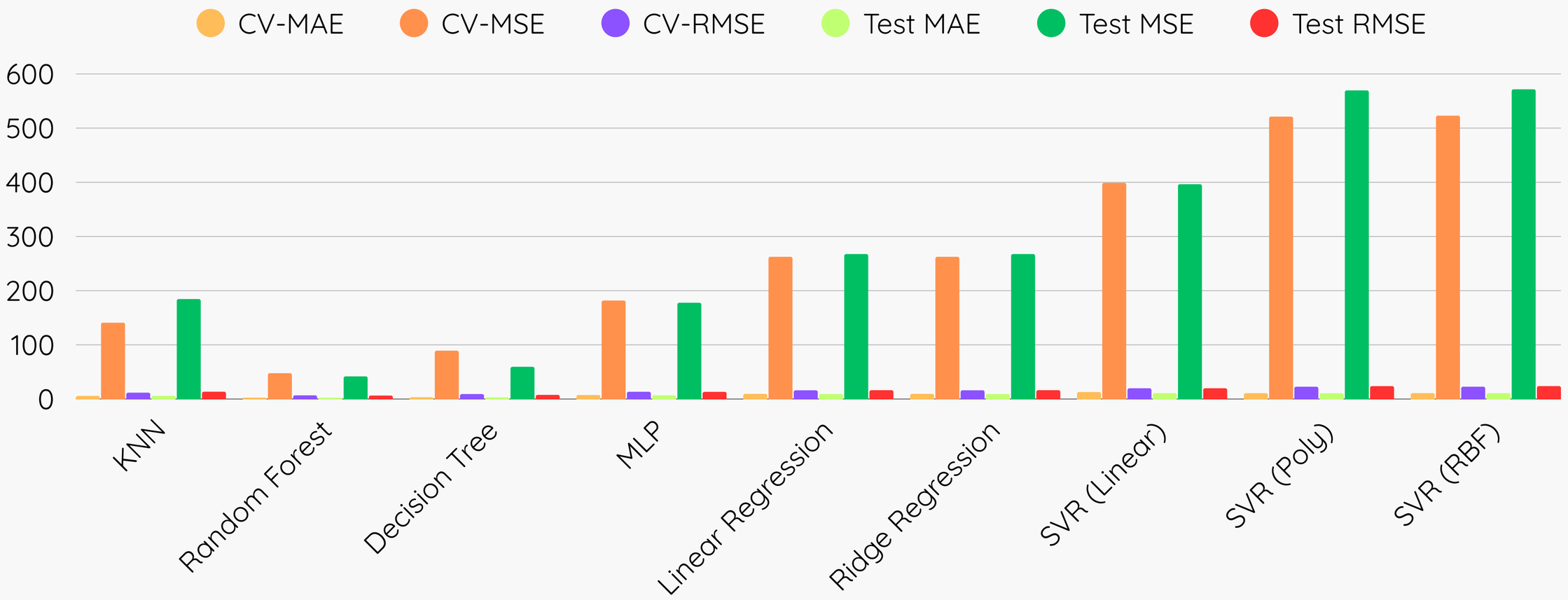
- **RMSE (Root Mean Squared Error):**

Square root of MSE. Interpretable in the same units as the target variable.

DATASET 1
Baseline Results

Model	CV-MAE	CV-MSE	CV-RMSE	Test MAE	Test MSE	Test RMSE
KNN	5.76	140.87	11.79	5.91	184.54	13.58
Random Forest	2.65	47.74	6.81	2.47	41.83	6.46
Decision Tree	3.32	89.3	9.27	3.01	59.53	7.71
MLP	7.41	181.83	13.46	6.88	177.72	13.33
Linear Regression	9.54	262.5	16.17	9.33	267.59	16.35
Ridge Regression	9.54	262.5	16.17	9.33	267.59	16.35
SVR (Linear)	12.95	399.03	19.87	10.89	396.45	19.91
SVR (Poly)	10.87	521.19	22.78	10.8	569.56	23.86
SVR (RBF)	10.96	522.94	22.82	10.89	571.62	23.9

DATASET 1
Baseline Results

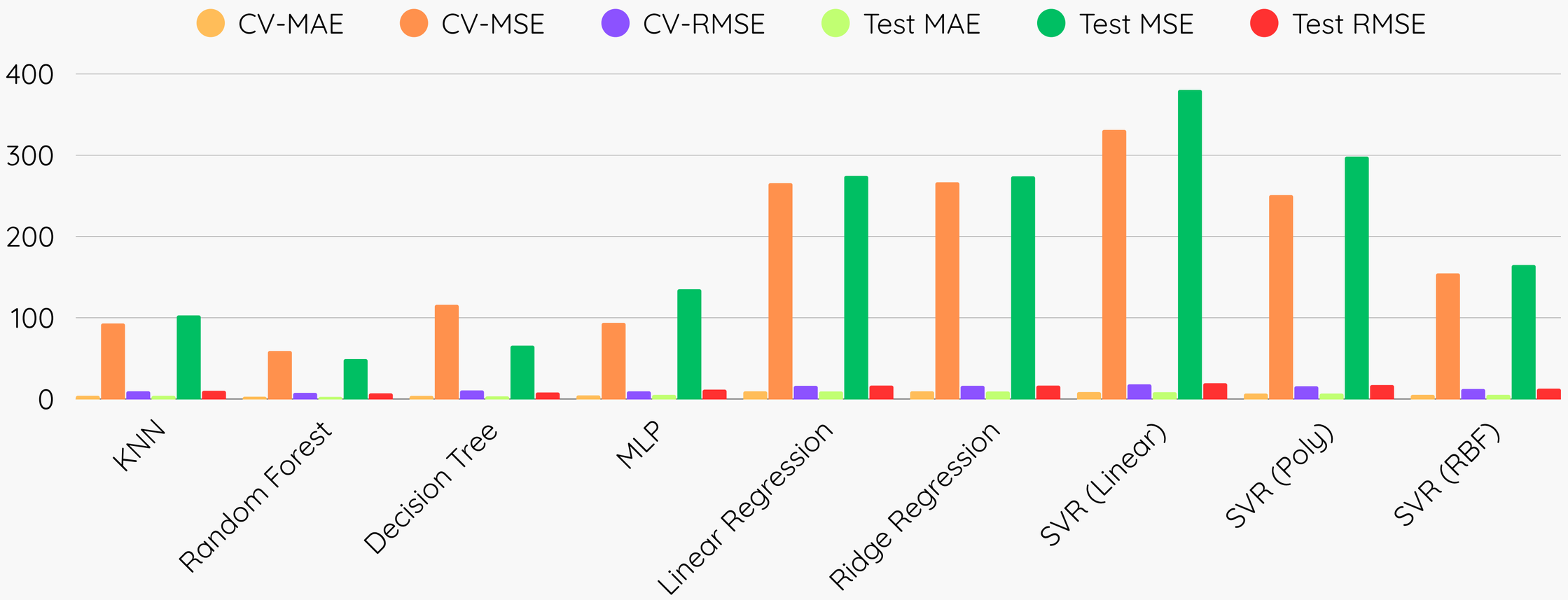


DATASET 1
GA Feature Selection Results

Model	CV-MAE	CV-MSE	CV-RMSE	Test MAE	Test MSE	Test RMSE
KNN	4.02	92.98	9.59	3.99	102.92	10.14
Random Forest	2.92	59.12	7.61	2.7	49.18	7.01
Decision Tree	3.89	115.97	10.64	3.34	65.72	8.1
MLP	4.48	93.73	9.54	5.17	135.14	11.62
Linear Regression	9.55	265.69	16.27	9.35	274.62	16.57
Ridge Regression	9.57	266.7	16.3	9.34	274.06	16.55
SVR (Linear)	8.64	331.12	18.13	8.49	380.29	19.5
SVR (Poly)	6.77	250.95	15.74	6.77	298.29	17.27
SVR (RBF)	5.22	154.61	12.41	5.21	164.95	12.84

DATASET 1

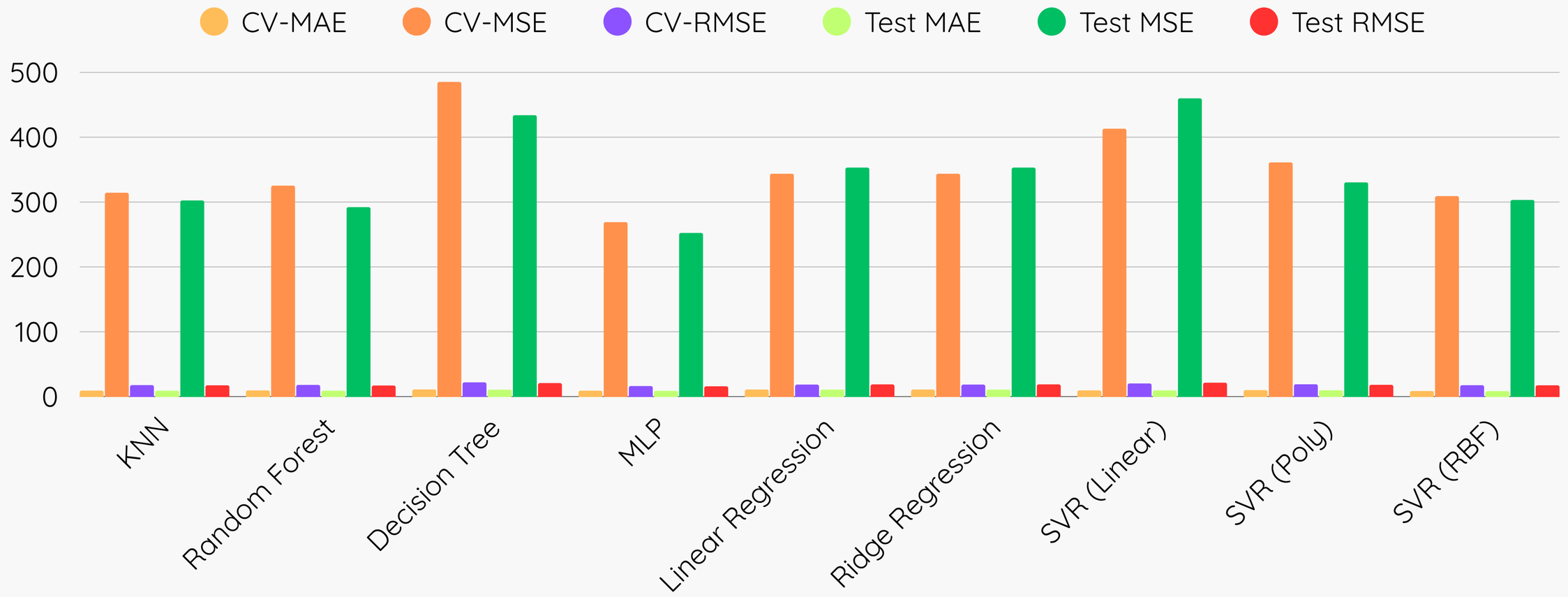
GA Feature Selection Results



DATASET 1
PCA Feature Reduction Results

Model	CV-MAE	CV-MSE	CV-RMSE	Test MAE	Test MSE	Test RMSE
KNN	9.28	314.33	17.71	9.11	302.44	17.39
Random Forest	9.53	325.28	18.01	9.13	292.11	17.09
Decision Tree	11.03	485.18	21.95	10.66	433.86	20.82
MLP	9.14	268.96	16.28	8.86	252.34	15.88
Linear Regression	10.9	343.59	18.51	10.78	353.1	18.79
Ridge Regression	10.9	343.59	18.51	10.78	353.1	18.79
SVR (Linear)	9.6	413.04	20.27	9.56	459.86	21.42
SVR (Poly)	9.98	361.09	18.99	9.75	330.3	18.17
SVR (RBF)	8.57	309.17	17.57	8.44	303.21	17.41

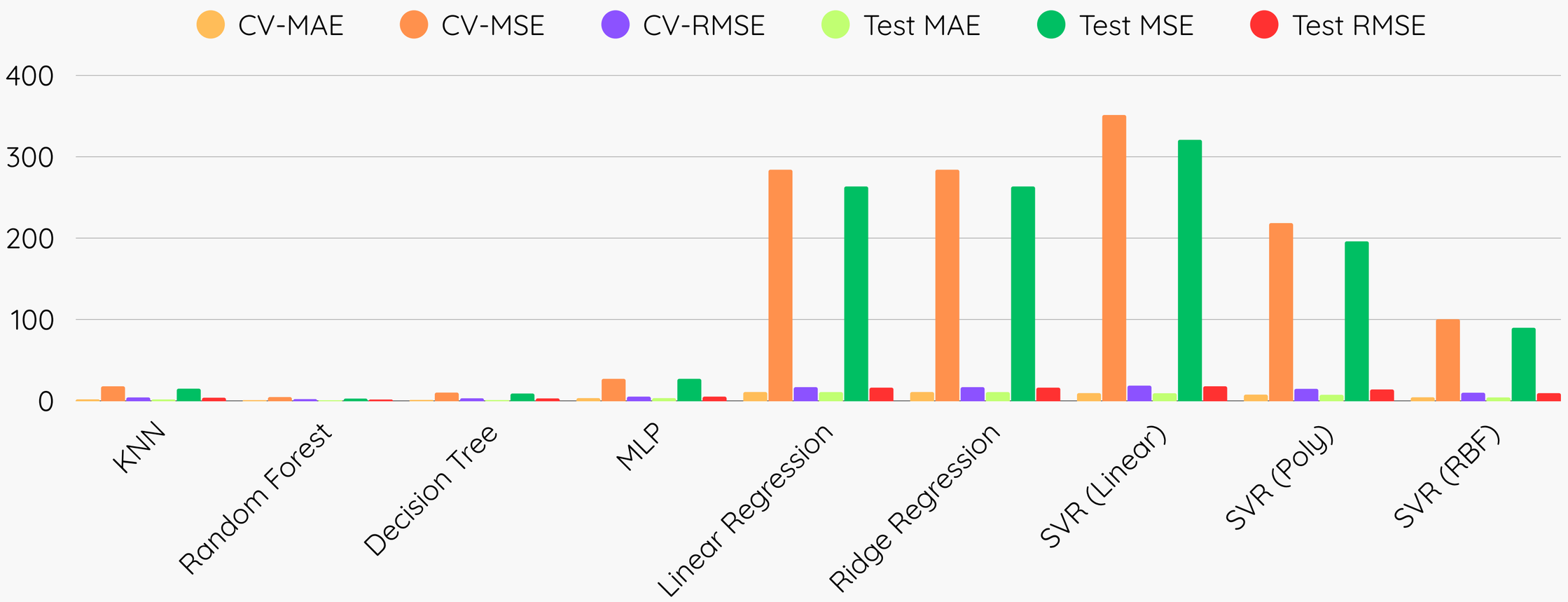
DATASET 1
PCA Feature Reduction Results



DATASET 2
Baseline Results

Model	CV-MAE	CV-MSE	CV-RMSE	Test MAE	Test MSE	Test RMSE
KNN	1.85	17.89	4.22	1.73	14.97	3.86
Random Forest	0.91	4.65	2.14	0.75	2.77	1.66
Decision Tree	1.16	10.18	3.15	0.98	9.02	3
MLP	3.34	27.07	5.2	3.34	27.08	5.2
Linear Regression	10.87	284.07	16.85	10.79	263.48	16.23
Ridge Regression	10.87	284.07	16.85	10.79	263.48	16.23
SVR (Linear)	9.53	351.31	18.73	9.36	320.82	17.91
SVR (Poly)	7.68	218.46	14.77	7.45	196.02	14
SVR (RBF)	4.38	100.35	10.01	4.22	89.83	9.47

DATASET 2
Baseline Results

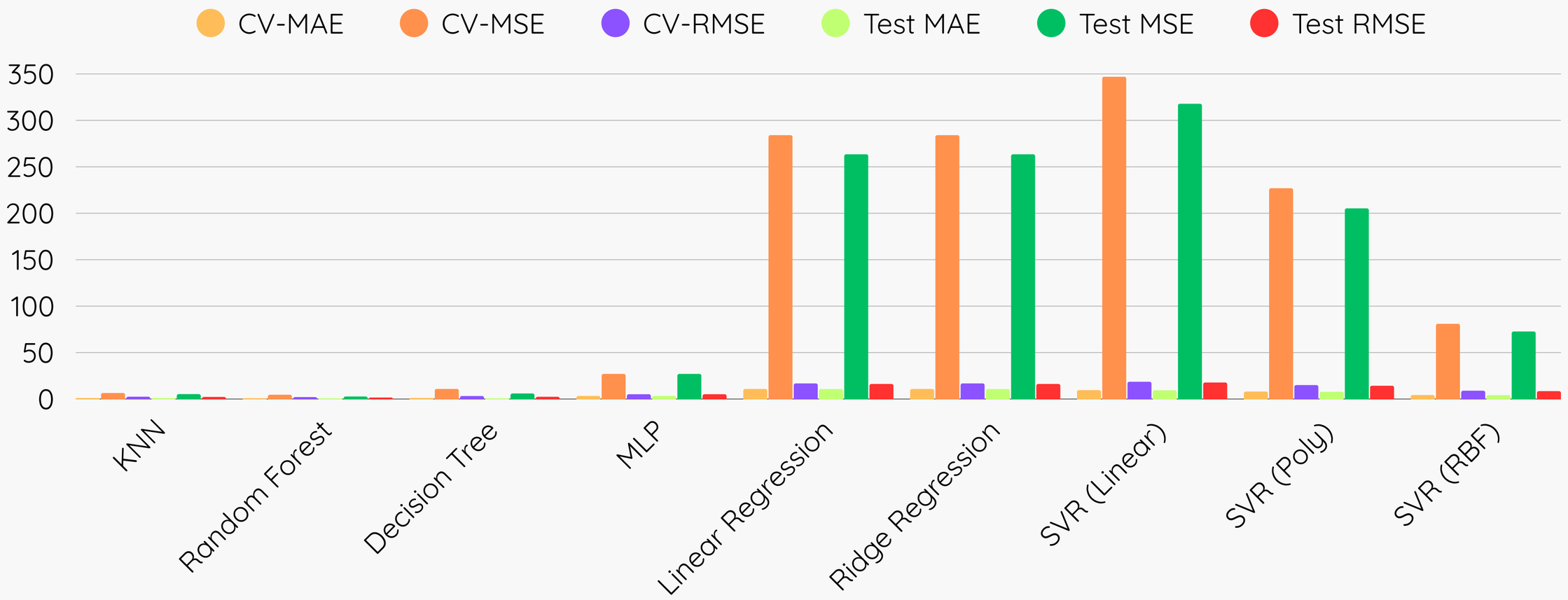


DATASET 2
GA Feature Selection Results

Model	CV-MAE	CV-MSE	CV-RMSE	Test MAE	Test MSE	Test RMSE
KNN	1.09	6.52	2.54	0.98	5.34	2.31
Random Forest	0.91	4.65	2.14	0.75	2.77	1.66
Decision Tree	1.17	10.89	3.27	0.75	6.01	2.45
MLP	3.34	27.07	5.2	3.34	27.08	5.2
Linear Regression	10.87	284.07	16.84	10.79	263.47	16.23
Ridge Regression	10.87	284.07	16.84	10.79	263.47	16.23
SVR (Linear)	9.6	346.91	18.62	9.43	317.86	17.82
SVR (Poly)	8.08	226.95	15.06	7.87	205.26	14.32
SVR (RBF)	4.32	80.96	8.99	4.22	72.73	8.52

DATASET 2

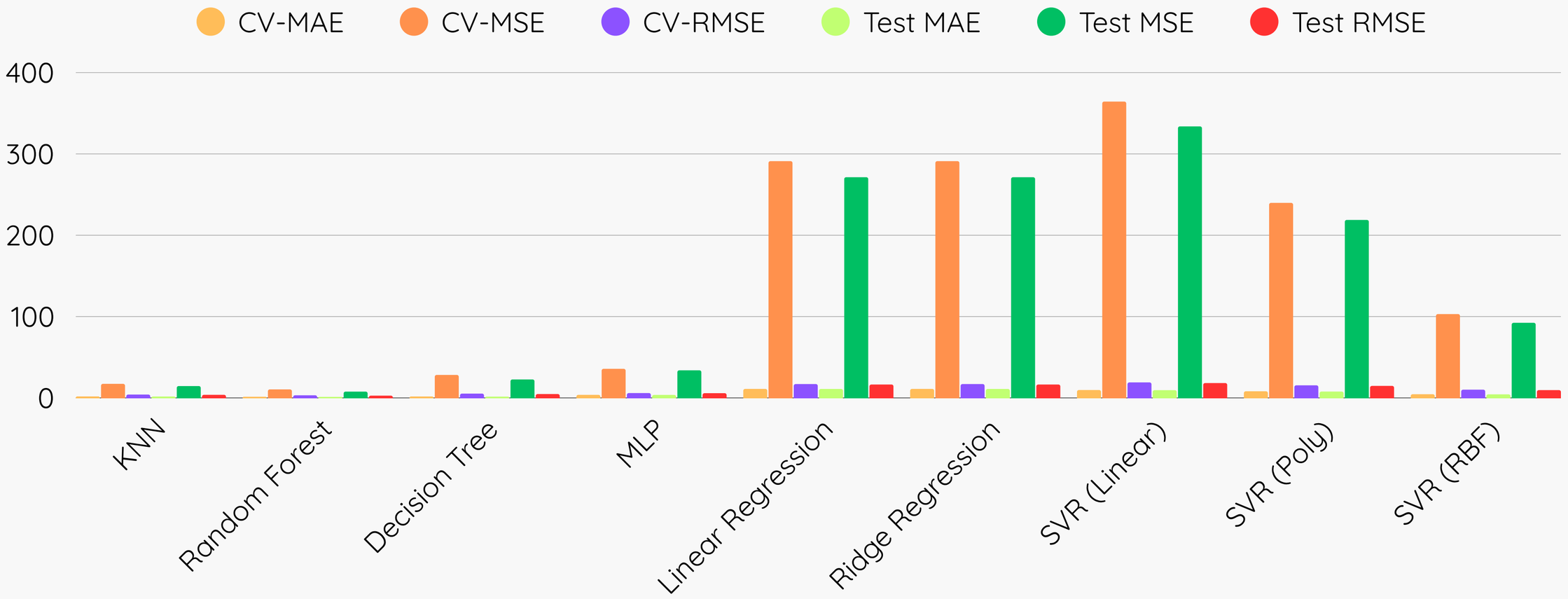
GA Feature Selection Results



DATASET 2
PCA Feature Reduction Results

Model	CV-MAE	CV-MSE	CV-RMSE	Test MAE	Test MSE	Test RMSE
KNN	1.83	17.3	4.15	1.72	14.57	3.81
Random Forest	1.44	10.46	3.23	1.28	7.75	2.78
Decision Tree	1.77	28.26	5.3	1.58	22.7	4.76
MLP	3.78	35.77	5.97	3.7	33.92	5.82
Linear Regression	11.09	291.09	17.05	11.02	271.33	16.47
Ridge Regression	11.09	291.09	17.05	11.02	271.33	16.47
SVR (Linear)	9.71	364.31	19.08	9.52	333.8	18.27
SVR (Poly)	8.15	239.78	15.48	7.82	218.79	14.79
SVR (RBF)	4.55	103.07	10.15	4.4	92.34	9.6

DATASET 2
PCA Feature Reduction Results



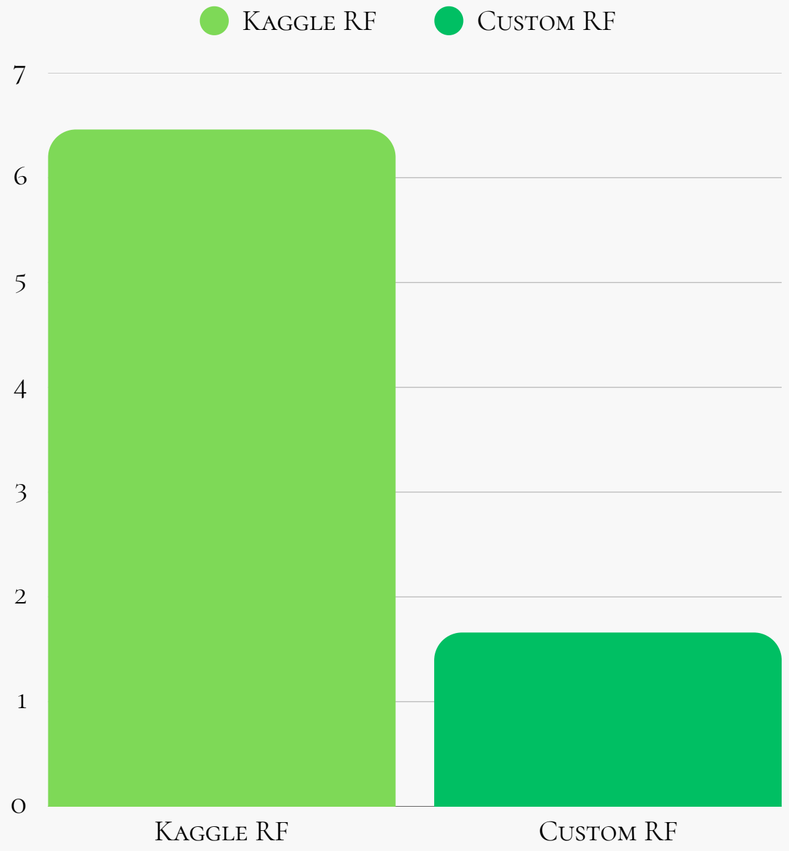
Dataset Comparison

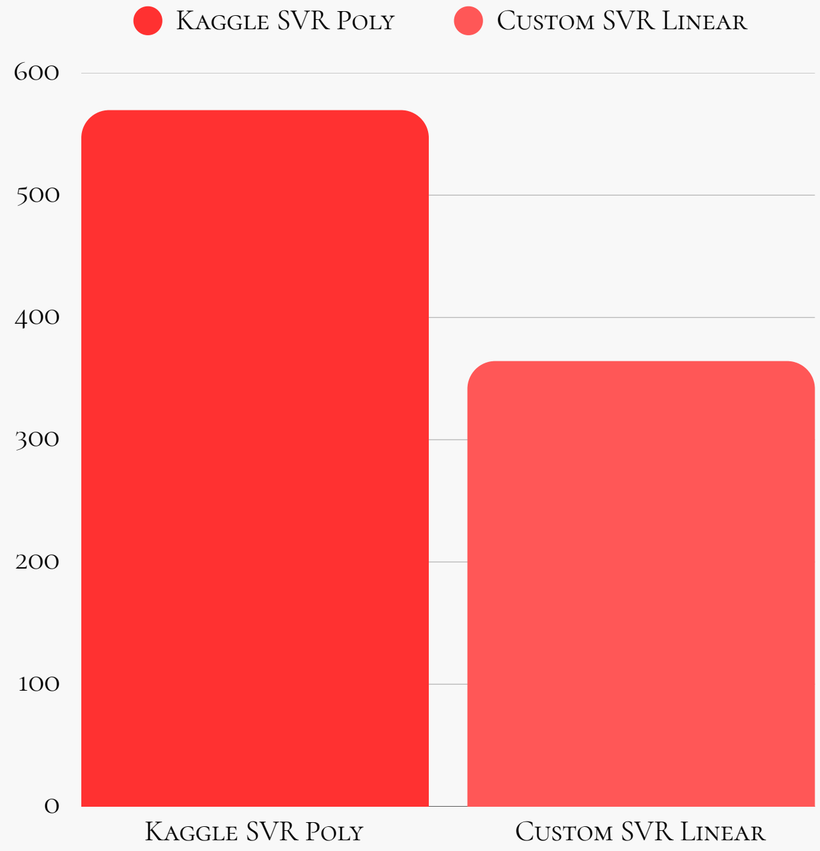
Custom Dataset vs Kaggle Dataset

Accuracy Gains

- **Best model (Random Forest):**
  - Kaggle → RMSE: 6.46
  - Custom → RMSE: 1.66

- **Worst Model (SVR):**
  - Kaggle → RMSE: 569.56
  - Custom → RMSE: 364.31





External Comparison

How Our Results Compare to Previous Studies

Peer-Reviewed Study

"Suicide Prediction Using Machine Learning Techniques"

Used Kaggle dataset with:

- **Random Forest → CV RMSE: 4.00**
- Decision Tree → CV RMSE: 6.34
- SVR → CV RMSE: 21.99

Our Kaggle Dataset

- **Random Forest → CV RMSE: 6.46**
- Decision Tree → CV RMSE: 7.71
- SVR → CV RMSE: 154.61

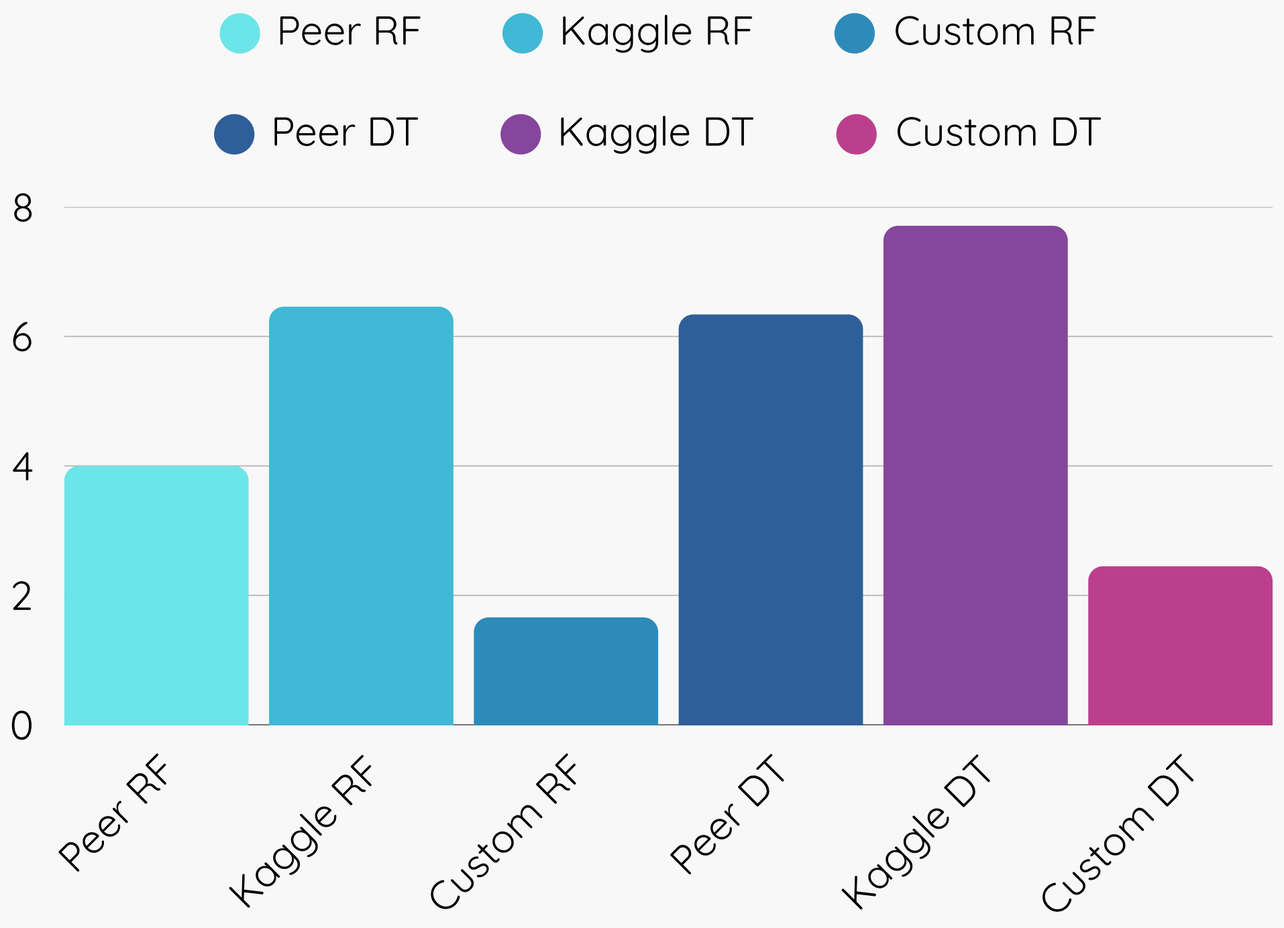
Our Custom Dataset

- **Random Forest → CV RMSE: 1.66**
- **Decision Tree → CV RMSE: 2.45**
- SVR → CV RMSE: 72.73

External Comparison

How Our Results Compare to Previous Studies

Our hand-curated dataset and pipeline **surpassed prior benchmarks**, confirming the strength of our data integration and modeling strategy.



Summary

What we did

- Trained 9 regression models on 2 datasets
- Applied 3 preprocessing strategies:

➤ Baseline

➤ Genetic Algorithm feature selection

➤ PCA feature reduction

Key Findings

- Custom Dataset (2000–2019) outperformed Kaggle data → Best RMSE: 1.66 (Random Forest)
- Random Forest was consistently the top performer
- PCA often reduced accuracy

Takeaway

- High-quality, multi-source features lead to better predictions
- Our pipeline surpassed prior benchmarks using the Kaggle dataset

Behind every data point is a life.

This study was conducted in the hope the insights turn into action to prevent the next tragedy.

Thank You

for your attention

issue | social issue | social issue | social issue | social issue | social issue | social issue |