

Optimization of Machine Learning Datasets through Evolutionary Computing and Symbolic Regression

Alejandro Iabin Arteaga Hernandez

Universidad Carlos III de Madrid

Avda. de la Universidad, 30. 28911 Leganes (Madrid). Spain

1. Executive Summary

Supervised regression often faces a trade-off between interpretability and predictive performance. Linear models are transparent but limited by strong structural assumptions, while ensemble methods such as Random Forests improve accuracy at the cost of opacity. This project studies a hybrid framework that combines Random Forest-based feature selection with Genetic Programming (GP) Symbolic Regression to reduce this gap.

The configuration is applied to two benchmark problems: the Diabetes dataset, representing a noisy quasi-linear biomedical system, and the California Housing dataset, representing a non-linear spatial economic system. The pipeline operates in two stages: (1) a feature selection step using Gini importance from a Random Forest Regressor to prune low-importance features and shrink the search space; and (2) a GP-based Symbolic Regression stage (via gplearn) with an extended function set (including sin and cos) and relatively high mutation pressure to explore non-linear expressions. For the GP stage, both inputs and targets are standardized and predictions are mapped back to the original target scale before evaluation.

Across both domains, the symbolic models improve over standard Linear Regression baselines:

- **Diabetes:** Test MAE of 40.21 vs. 42.79 for Linear Regression (6.03 % improvement).
- **California Housing:** Test MAE of 0.4751 vs. 0.5332 for Linear Regression (10.89 % improvement).

This suggests that combining model-based feature selection with GP can yield compact, interpretable formulas with better accuracy than a purely linear hypothesis class.

2. Introduction and Theoretical Overview

In supervised regression, the learner must approximate an unknown mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ from data. The choice of hypothesis space \mathcal{H} constrains what the model can express; when $f \notin \mathcal{H}$, the model incurs irreducible bias. Ordinary Least Squares Linear Regression assumes that the target is a

linear combination of features plus noise,

$$y = \beta_0 + \sum_{j=1}^n \beta_j x_j + \epsilon,$$

implicitly enforcing linearity, additivity, and monotonic effects. In domains such as physiology or spatial housing prices, these assumptions are often violated, leading to underfitting.

Symbolic Regression (SR) relaxes these constraints by searching over the space of mathematical expressions instead of fixing an equation form in advance. In this project, SR is implemented via Genetic Programming: candidate models are represented as expression trees with terminals (input features and constants) and functions (e.g. add, mul, sin, log). Using gplearn, the population is initialized with random trees and evolved for a fixed number of generations (80) under a cycle of fitness evaluation (MAE on the training set), tournament selection (size = 20), subtree crossover and mutation. This yields compact, non-linear formulas tuned directly to the data.

However, GP is sensitive to the curse of dimensionality: many weak or irrelevant features inflate the search space, promote tree bloat, and encourage overfitting. To mitigate this, the framework applies a model-based filter using Random Forest feature importance. A Random Forest Regressor provides a Mean Decrease in Impurity (MDI) score for each feature by aggregating the variance reduction at split nodes using that feature, averaged across trees. Unlike simple linear correlation, this measure captures non-linear dependencies and interactions. Retaining only the highest-importance features produces a reduced, high-signal input set on which GP can search more efficiently, enabling the evolved symbolic models to reach better accuracy with lower complexity.

3. Methodology: System Architecture

The proposed system follows a two-stage pipeline: (1) model-based feature selection using Random Forests, and (2) Symbolic Regression via Genetic Programming (GP) with a trigonometric function set.

3.1. Data Preprocessing and Baseline Models

Raw data are loaded from CSV files and numeric columns are coerced to `float64` to avoid type inconsistencies. The dataset is split into training and test partitions using an 80/20 split with a fixed random seed (`seed = 42`) to ensure reproducibility.

Two baseline models are trained using `scikit-learn` pipelines:

- **Linear Regression:** median imputation (`SimpleImputer`), feature standardization (`StandardScaler`), and `LinearRegression`.
- **Random Forest Regressor:** median imputation, followed by a `RandomForestRegressor` with `n_estimators = 100` and maximum depth 10.

These baselines provide reference MAE values for both training and test sets.

For the GP stage, only the features selected in Phase 1 are used. These inputs are standardized to zero mean and unit variance using `StandardScaler`. The target variable is also standardized to zero mean and unit variance, and the GP model is trained in this normalized space. At prediction time, GP outputs are inverse-transformed back to the original target units before computing MAE. All MAE values reported in the case studies are expressed in the original target scale.

3.2. Phase 1: Intelligent Dimensionality Reduction

The first phase transforms the raw input matrix X_t into a reduced matrix X'_t by pruning low-importance features using a Random Forest Regressor:

1. **Pre-training:** A `RandomForestRegressor` (`n_estimators = 100`) is trained on the full feature set.
2. **Importance extraction:** The `feature_importances_` vector is obtained.
3. **Thresholding:** Features with importance below 0,04 are removed:

If $\text{Imp}(X_j) < 0,04$, then X_j is discarded.
4. **Safety fallback:** If fewer than three features satisfy the threshold, the top three features by importance are retained.

This procedure supplies GP with a compact, high-signal subset of variables, mitigating the curse of dimensionality and reducing search-space bloat.

3.3. Phase 2: Evolutionary Synthesis via GP

The reduced dataset X'_t is then passed to `gplearn.SymbolicRegressor`. The main hyperparameters are:

Parameter	Value
Population size	5000
Generations	80
Tournament size	20
Parsimony coefficient	0.001
Metric	MAE

The fitness function used is

$$\text{Fitness} = \text{MAE} + 0,001 \times \text{Length},$$

which penalizes excessively long expressions while allowing sufficient complexity.

Function Set Candidate expressions are represented as trees with terminals (input features and constants) and a function set

$$\mathcal{F} = \{+, -, \times, \div, \sqrt{}, \log, |\cdot|, \sin, \cos\}.$$

Trigonometric functions are included to model:

- non-linear spatial patterns (e.g. periodic behaviour in latitude/longitude),
- soft saturation effects in physiological variables.

The tangent function is excluded to avoid numerical instabilities due to its vertical asymptotes.

Genetic Operators and Mutation Strategy Expression trees are evolved under a standard GP cycle: initialization, fitness evaluation on training data (MAE), tournament selection, crossover, mutation, and termination after 80 generations.

Compared to typical GP defaults, the configuration increases the point-mutation probability to encourage finer adjustment of numeric constants. Concretely, crossover is set to probability 0,6 and point mutation to 0,25. This maintains structural exploration while continuously perturbing constants, improving the numerical fit of the resulting symbolic models.

4. Case Study 1: Diabetes (Quasi-linearity and Noise)

The Diabetes dataset from `sklearn.datasets` contains 442 samples with 10 standardized clinical features (age, sex, BMI, blood pressure, and 6 serum measures) and a continuous target measuring disease progression after one year.

Baseline performance on the held-out test set was:

- Linear Regression MAE: 42.79
- Random Forest MAE: 44.38

The fact that Linear Regression slightly outperforms Random Forest suggests a predominantly linear, high-noise regime where complex trees tend to overfit.

Random Forest feature importance showed a concentrated signal, with `bmi` and `s5` (serum triglycerides) dominating. Applying the 4 % importance threshold reduced the feature set from 10 to 8, discarding very low-importance variables such as `sex` and `s4`. This removes features that mainly contribute noise to the GP search.

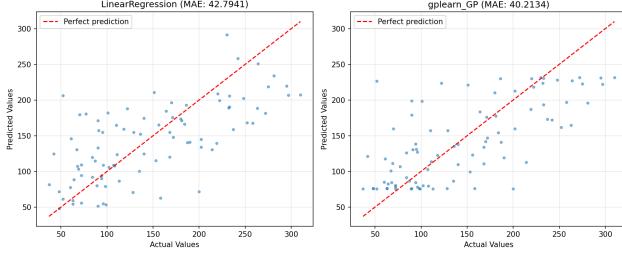


Figura 1: Actual vs. predicted values (Linear Regression vs. GP model) on 20 % held-out Diabetes data.

4.1. Evolved Symbolic Model

Let z_{bmi} and z_{s5} denote the standardized versions of BMI and s5, i.e.,

$$z_{\text{bmi}} = \frac{\text{bmi} - \mu_{\text{bmi}}}{\sigma_{\text{bmi}}}, \quad z_{\text{s5}} = \frac{\text{s5} - \mu_{\text{s5}}}{\sigma_{\text{s5}}},$$

and let z_y be the standardized target,

$$z_y = \frac{y - \mu_y}{\sigma_y}.$$

In these standardized coordinates, the best GP program learned for Diabetes is

$$\hat{z}_y = \sin(\sin(\sin(\sin(\sin(z_{\text{bmi}})))) + \sin(z_{\text{s5}})).$$

Predictions in the original target scale are then recovered by the inverse standardization

$$\hat{y} = \mu_y + \sigma_y \hat{z}_y.$$

Thus, the final symbolic model for disease progression depends only on BMI and triglycerides through a nested sine composition, acting as a smooth saturation of these two standardized variables.

The GP run converged to a compact model with low and stable tree length, reflecting the parsimony penalty. The final performance was:

- GP Test MAE: 40.21
- Linear Regression Test MAE: 42.79
- Relative improvement: 6.03 %

For standardized inputs near zero, $\sin(x) \approx x$, but for larger magnitudes the growth slows, producing a soft clipping effect: progression increases with BMI and triglycerides, but not indefinitely, which Linear Regression tends to overestimate for outliers.

5. Case Study 2: California Housing (Spatial Non-linearity)

The California Housing dataset contains 20,640 samples and 8 features (Median Income, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude), with the target being median house value in units of \$100,000.

Baseline performance:

■ Linear Regression MAE: 0.5332

■ Random Forest MAE: 0.3663

Here Random Forest clearly dominates, indicating strong non-linear and interaction effects, especially of location.

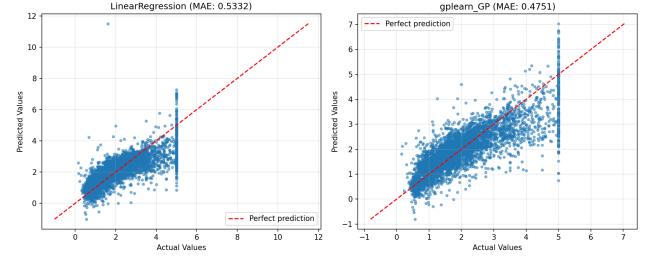


Figura 2: Actual vs. predicted values (Linear Regression vs. GP model) on 20 % held-out California Housing data.

Random Forest importance revealed a highly skewed structure: MedInc is the main driver, followed by AveOccup and the spatial coordinates (Latitude, Longitude). Applying the 4 % threshold removed AveRooms, AveBedrms, and Population. This reduction shrank the search space and removed variables that are largely redundant with income, forcing GP to exploit orthogonal information from location.

5.1. Evolved Symbolic Model

Let z_{MedInc} , z_{AveOcc} , z_{Lat} , and z_{Long} be the standardized versions of Median Income, Average Occupancy, Latitude, and Longitude, respectively, and let z_y denote the standardized target:

$$z_{\text{MedInc}} = \frac{\text{MedInc} - \mu_{\text{MedInc}}}{\sigma_{\text{MedInc}}}, \quad z_{\text{AveOcc}} = \frac{\text{AveOccup} - \mu_{\text{AveOcc}}}{\sigma_{\text{AveOcc}}},$$

$$z_{\text{Lat}} = \frac{\text{Latitude} - \mu_{\text{Lat}}}{\sigma_{\text{Lat}}}, \quad z_{\text{Long}} = \frac{\text{Longitude} - \mu_{\text{Long}}}{\sigma_{\text{Long}}}, \quad z_y = \frac{y - \mu_y}{\sigma_y}.$$

In these coordinates, the best GP program found for California can be written (in the `gplearn` functional notation) as:

$$\hat{z}_y = \text{add}\left(z_{\text{MedInc}}, \text{div}\left(T(z_{\text{MedInc}}, z_{\text{AveOcc}}, z_{\text{Lat}}, z_{\text{Long}}), -5,350\right)\right),$$

where

$$T = \text{add}\left(\text{add}\left(\text{add}\left(z_{\text{MedInc}}, \sin(\sqrt{z_{\text{Lat}}} + 6,203 z_{\text{Long}})\right), z_{\text{Lat}} + z_{\text{Long}}\right), \text{add}\left(\sin(\log z_{\text{AveOcc}}), S(z_{\text{MedInc}}, z_{\text{AveOcc}}, z_{\text{Lat}}, z_{\text{Long}})\right)\right),$$

and the inner correction term S is

$$S = \text{add}\left(\text{add}\left(\text{add}\left(\text{add}\left(\text{add}\left(z_{\text{MedInc}}, z_{\text{Long}} + z_{\text{Lat}}\right), 4,778 \sin(z_{\text{Lat}})\right), z_{\text{MedInc}} + z_{\text{Long}}\right), z_{\text{MedInc}} + z_{\text{Lat}}\right), z_{\text{MedInc}} + z_{\text{Long}}\right), z_{\text{MedInc}} + z_{\text{Lat}}\right).$$

As in the Diabetes case, predictions in the original scale are obtained via

$$\hat{y} = \mu_y + \sigma_y \hat{z}_y.$$

Although algebraically redundant terms appear in S (multiple repetitions of $z_{\text{Lat}} + z_{\text{Long}}$ and $\sin(z_{\text{AveOcc}})$ with the same coefficient), the structure is clear: a baseline linear term

in income, plus a trigonometric correction that combines latitude, longitude, and occupancy. The component

$$\sin(\sqrt{z_{\text{Lat}}} + 6,203 z_{\text{Long}})$$

acts as a spatial warp aligned with the coastline, while repeated additions of $(z_{\text{Lat}} + z_{\text{Long}})$ and $\sin(z_{\text{AveOcc}})$ control coarse regional and density effects.

In this setting, the resulting performance was:

- GP Test MAE: 0.4751
- Linear Regression Test MAE: 0.5332
- Relative improvement: 10.89 %

Trigonometric functions here act as a crude Fourier basis, allowing the model to capture the “waves” of high and low value across the California landscape (e.g. coastal vs. inland, metropolitan vs. rural zones) that a single linear plane cannot represent.

6. Comparative Analysis: Why the Hybrid RF+GP Approach Works

Across both datasets, standard Linear Regression is limited by a fixed, globally linear hypothesis, implicitly assuming constant derivatives. In Diabetes, this fails to reflect saturation of risk at extreme BMI or triglyceride levels; in California Housing, it cannot model the alternating peaks and valleys of price along spatial axes.

The hybrid approach leverages two complementary components:

- **Random Forest feature selection** uses Mean Decrease in Impurity to identify a small set of informative features, including non-linear and interaction effects, and removes low-utility or redundant variables. This sharply reduces the effective dimensionality of the GP search.
- **Genetic Programming Symbolic Regression** then operates on this reduced space, using a rich function set (including \sin , \cos , $\sqrt{\cdot}$, and \log) to construct explicit formulas that capture both mild non-linear corrections (Diabetes) and complex spatial manifolds (California Housing).

A relatively high point-mutation rate encourages fine adjustment of numeric constants inside the expressions, playing a role analogous to gradient-based tuning in parametric models. The combination of targeted feature selection, expressive function sets, and constant refinement explains why the RF+GP pipeline can surpass a purely linear baseline while preserving interpretability in the form of explicit analytical expressions.

7. Generative AI Annex

In line with the project requirements, this section briefly documents the use of Generative AI tools.

Generative AI was used in a narrow, supporting role. It helped (i) draft an initial Python skeleton for the Random Forest + GP pipeline, (ii) suggest interpretations for nested trigonometric terms on standardized inputs, and (iii) propose a reasonable range for the parsimony coefficient to control tree bloat.

Representative prompts were short, task-focused queries such as: “Design a Python class that combines `RandomForestRegressor` feature selection with `gplearn.SymbolicRegressor` using a 4 % importance threshold”, “Explain why a GP model might evolve nested $\sin(\sin(x))$ terms on standardized data”, and “Recommend a parsimony coefficient for a population of 5000 individuals with trigonometric functions.”

The responses accelerated implementation and helped frame hypotheses, but all code, hyperparameters and function-set decisions were validated experimentally. For example, although the AI suggested including \tan , manual testing revealed numerical instabilities, so \tan was excluded from the final function set.

8. Limitations and Future Work

This work is limited to two benchmark regression datasets and a single Genetic Programming configuration with a fixed random seed. As a result, the stability of the evolved expressions with respect to initialization and hyperparameters has not been systematically evaluated. In addition, while the trigonometric terms admit a functional interpretation as smooth nonlinear corrections, they are not claimed to have direct physical meaning. Future work could extend the analysis to a broader set of datasets, study the variability of the symbolic models across runs, and explore alternative feature selection criteria (e.g. permutation importance) or additional regularisation terms to further constrain model complexity.

9. Conclusion

This study indicates that the limitations of standard regression models—specifically the high bias of Linear Regression and the opacity of Random Forests—can be mitigated by a hybrid system combining feature selection and symbolic regression.

The proposed Random Forest + GP configuration, with feature selection followed by trigonometric Symbolic Regression, performed robustly on two different data regimes:

- In the Diabetes domain, it modeled diminishing returns of physiological risk factors, improving Test MAE by 6.03 % over Linear Regression.
- In the California Housing domain, it modeled a non-linear spatial correction on top of income, improving Test MAE by 10.89 %.

The project supports the view that dataset optimization via feature deletion is not only a method for speeding up computation, but also a useful step for Symbolic Regression. By removing features with low Random Forest importance, the Genetic Programming component can focus its search on estimating the main non-linear relationships present in the data.

Although the Random Forest baseline attains lower MAE than the GP model on California Housing, it does so at the cost of opacity. The hybrid Random Forest + GP pipeline recovers explicit analytical expressions that improve over a linear baseline while remaining substantially more interpretable than the ensemble model.