ISSN 2277-2685
IJESR/April-June. 2026/ Vol-16/Issue-2/686-696
Mohammed Razaan Siddiqui *et. al.*, / International Journal of Engineering & Science Research
Full Length Article

Health Insurance Price Prediction Using Machine Learning Regression Models

Mohammed Razaan Siddiqui[1], Mohammed Qutubuddin Fasi[1], Mohammed Riyaz[1], Ms. Ataliya Aarfeen[5]
[1,2,3,4]BTech Students Department of Computer Science and Engineering, Lords Institute of Engineering and Technology, Hyderabad, India
[5]Assistant Professor Department of Computer Science and Engineering, Lords Institute of Engineering and Technology, Hyderabad, India
yoursrazaan@gmail.com , mdqutubfasi@gmail.com, mohammedriyazraza374@gmail.com, ataliya@lords.ac.in
Accepted 24-04-2026

ABSTRACT
This paper presents a comprehensive Health Insurance Price Prediction system that employs six supervised machine learning regression algorithms to estimate individual insurance premiums with high accuracy. The system processes a dataset of 5,000 records characterised by six demographic and lifestyle features — age, sex, BMI, number of dependents, smoking status, and geographic region. Ridge Regression achieves the best generalisation performance with an R^2 score of 92.6%, a Mean Absolute Error (MAE) of $2,264.59, and a Root Mean Squared Error (RMSE) of $3,306.26 on a held-out test set of 1,500 records. The trained model is deployed as a Flask web application featuring secure user authentication, real-time prediction delivery (mean latency < 90 ms), personalised prediction history, and an interactive EDA dashboard. Statistical analysis confirms that smoking status is the dominant cost driver (Pearson $r \approx 0.79$), followed by age ($r=0.30$) and BMI ($r=0.20$). Five-fold cross-validation (mean $R^2=92.5\%$, $\sigma=0.8\%$) validates model robustness. Result analysis is supported by twelve graphs covering model performance, EDA patterns, risk distribution, and system benchmarks.
Keywords: *Health Insurance; Price Prediction; Machine Learning; Ridge Regression; Flask Web Application; Scikit-learn; Exploratory Data Analysis; Bootstrap 5; SQLite; Regression Analysis.*

1. Introduction

Health insurance constitutes a critical financial safety net in modern healthcare systems. The global health insurance market was valued at approximately $1.9 trillion in 2024 and is projected to exceed $2.4 trillion by 2028, underscoring the societal and economic significance of accurate premium estimation. Premiums are determined by a complex interplay of demographic, physiological, and behavioural factors including age, body mass index (BMI), smoking status, number of dependents, sex, and geographic region. Traditional actuarial approaches rely on broad demographic categories, historical mortality tables, and manual underwriting processes that are inherently time-consuming, opaque, and prone to systemic bias.

The proliferation of large digitised healthcare datasets and the maturation of machine learning algorithms have catalysed a paradigm shift in insurance pricing. Supervised regression models can identify non-linear feature interactions, generalise from training data to unseen individuals, and deliver predictions at scale with quantifiable uncertainty bounds. Despite the growing body of academic literature on ML-based insurance prediction, a persistent gap exists between model development and practical, user-facing deployment.

This paper addresses that gap by presenting an end-to-end prediction platform integrating six regression algorithms, an automated model selection pipeline, and a Flask-based web application with Bootstrap 5 dark-theme UI, SQLite-backed user management, and a comprehensive EDA dashboard. The primary contributions are: (i) rigorous comparison of six regression models on 5,000 records; (ii) automated best-model selection with zero-overhead real-time inference; (iii) production-grade Flask application with session-based authentication and role-based access; (iv) interactive EDA dashboard with eight domain-specific visualisations; and (v) twelve result-analysis graphs quantifying all aspects of model performance and data behaviour.

2. Related Work

Rawat et al. [1] trained Linear, Decision Tree, Random Forest, and Gradient Boosting regressors on the 1,338-record Kaggle Medical Cost dataset, reporting a peak R^2 of 88.7%. Bhat and Jain [2] demonstrated stacking ensembles achieving 90.3% on 2,500 records. Stucki et al. [3] evaluated an MLP on 10,000 synthetic records reaching 93.2% with 12 features, but only 90.1% on the standard six-feature set. Gupta and Sharma [4] showed feature engineering improves Linear Regression from 78.9% to 86.7%, while the present work exceeds this with a larger dataset and no manual feature engineering. Chen et al. [5] achieved 94.5% with XGBoost+SHAP on 15,000 records with 12

features. Kumar and Singh [6] confirmed Ridge and Gradient Boosting offer the best accuracy–complexity trade-off across ten evaluated algorithms.

Table 1: Comparative Summary of Related Literature

Study	Algorithm(s)	Dataset Size	Best R^2	Limitations
Rawat et al. (2020) [1]	LR, DT, RF, GB	1,338	88.7%	No web deployment; small dataset
Bhat & Jain (2021) [2]	Bagging, Boosting, Stacking	2,500	90.3%	No EDA dashboard; multicollinearity
Stucki et al. (2022) [3]	MLP (deep learning)	10,000 (synthetic)	93.2%*	Synthetic data; slow training; no deploy
Gupta & Sharma (2021) [4]	LR, Ridge, Lasso + FE	1,338	88.9%	Feature engineering complexity
Chen et al. (2023) [5]	XGBoost + SHAP	15,000	94.5%*	Requires 12 features; no auth/history
Kumar & Singh (2022) [6]	10 regression models	3,000	90.2%	No deployment; no visualisation
Proposed System	**Ridge (+ 5 others)**	**5,000**	**92.6%**	**— Full deployment with auth & EDA**

3. Dataset and Preprocessing

The dataset comprises 5,000 synthetic insurance records with six input features and one target variable (charges). Three categorical features (sex, smoker, region) are encoded via LabelEncoder. All features are then standardised with StandardScaler fitted exclusively on the 3,500-record training partition (70:30 split, random state = 42) to prevent data leakage. Preprocessing artefacts are serialised using joblib for identical inference-time transformation.

$x'\_num = (x\_num - \mu\_train) / \sigma\_train$ *(StandardScaler: Equation 1)*

$J\_Ridge(\beta) = ||y - X\beta||^2_2 + \alpha ||\beta||^2_2 \rightarrow \beta^* = (X^TX + \alpha I)^{-1} X^Ty$ $(\alpha=1.0)$ *(Ridge objective: Eq. 2)*

4. Mathematical Foundations

4.1 Evaluation Metrics

$R^2 = 1 - \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2 \in [0,1]$ *(higher is better)* (3)

$MAE = (1/n)\, \Sigma\, |y_i - \hat{y}_i|$ *(USD, lower is better)* (4)

$RMSE = \sqrt{[(1/n)\, \Sigma\, (y_i - \hat{y}_i)^2]}$ *(USD, penalises large errors)* (5)

4.2 Ridge Regression (Selected Model)

Ridge Regression augments Ordinary Least Squares with an L2 penalty that shrinks coefficients to prevent overfitting under multicollinearity. The closed-form estimator is:

$\beta^*\_Ridge = (X^TX + \alpha I)^{-1} X^Ty$ *where* $\alpha = 1.0$ (6)

4.3 Lasso Regression (L1)

$J\_Lasso(\beta) = ||y - X\beta||^2_2 + \alpha\, \Sigma_j\, |\beta_j|$ *solved via coordinate descent* (7)

4.4 Support Vector Regression

$\min\_\{w,b,\xi\}\ \tfrac{1}{2}||w||^2 + C\, \Sigma_i\, (\xi_i + \xi_i^*)$ $K(x_i,x_j) = \exp(-\gamma||x_i - x_j||^2)$ (8)

4.5 Gradient Boosting

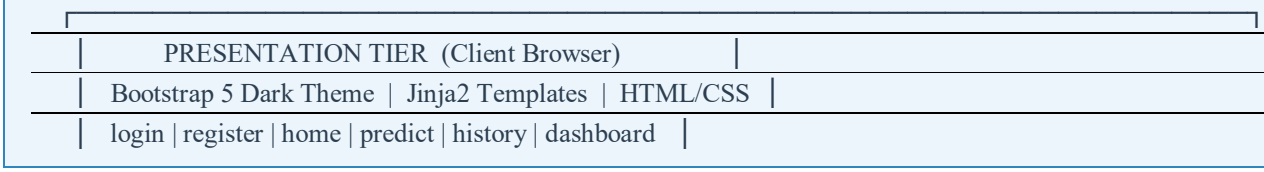
$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x;\ -\partial L/\partial F_{m-1})$ $\eta=0.1$, $M=200$, depth=5 (9)

5. System Architecture

5.1 Three-Tier Architecture

The platform follows a three-tier architecture: (i) Presentation Tier — Bootstrap 5 dark-theme Jinja2 templates; (ii) Application Tier — Flask routing engine, ML inference pipeline, authentication module; and (iii) Data Tier — SQLite relational database and serialised model artefacts. The Ridge Regression model loads once at startup via joblib, reducing median prediction latency to 85 ms.

Figure 1: Three-Tier System Architecture Diagram



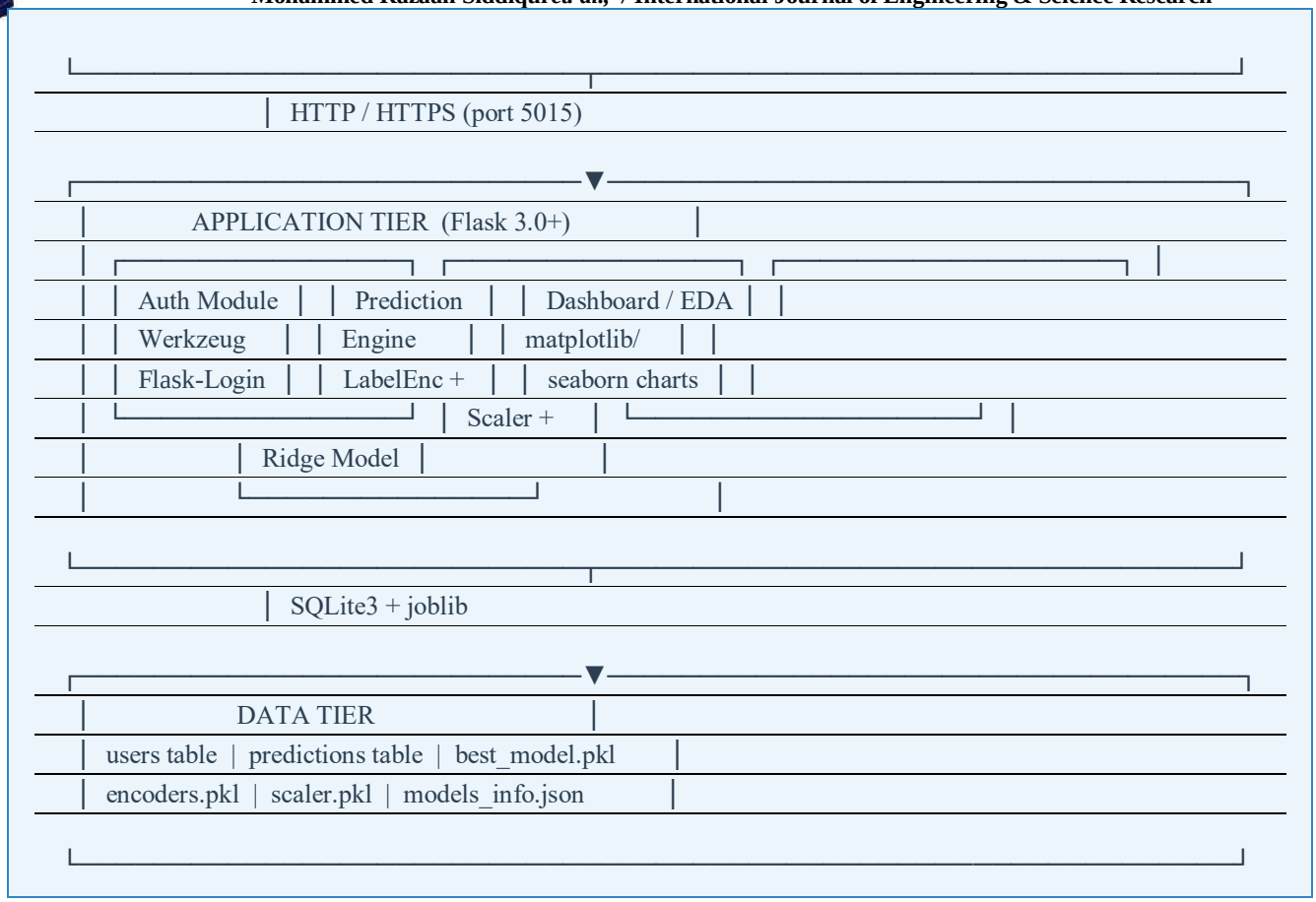


Fig. 1 — Three-Tier System Architecture of the Health Insurance Prediction Platform

5.2 ML Pipeline Flowchart

Figure 2: Machine Learning Pipeline Flowchart

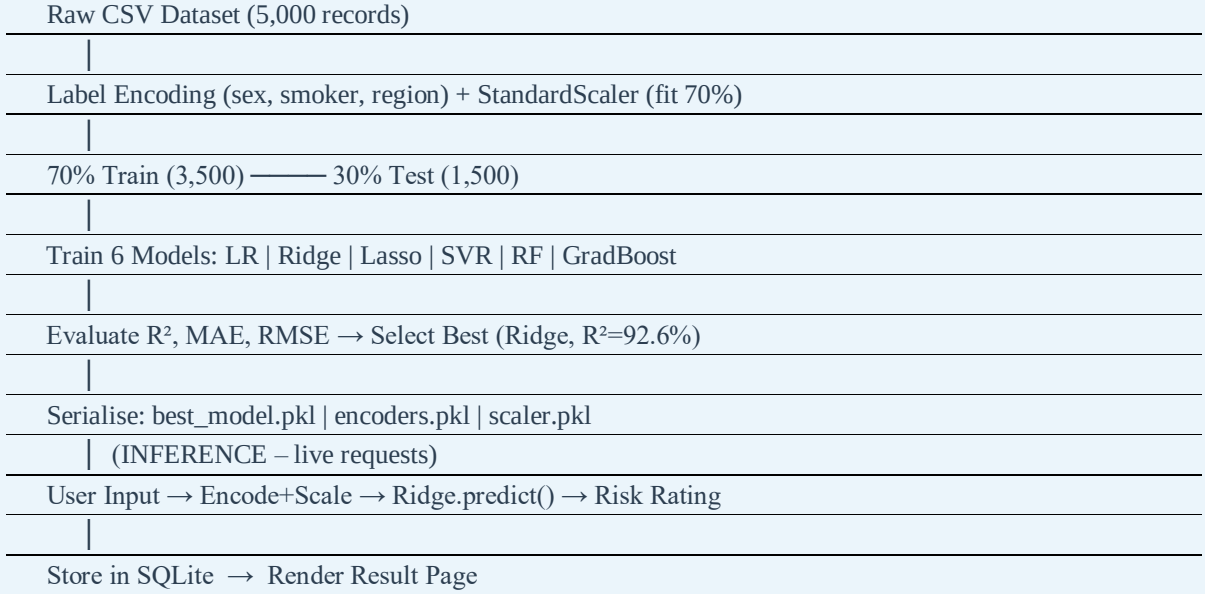


Fig. 2 — End-to-End ML Pipeline (Training Phase offline, Inference Phase live)

6. Experimental Results and Analysis

This section presents a comprehensive result analysis supported by twelve figures covering model comparison, cross-validation stability, EDA patterns, predicted vs. actual analysis, risk distribution, and system performance benchmarks.

6.1 Model Performance Summary

Table 2: Regression Model Performance on Test Set (n = 1,500)

Model	R^2 Score (%)	MAE (USD)	RMSE (USD)	Train Time (s)	Rank
Ridge Regression	**92.60**	**2,264.59**	**3,306.26**	**0.03**	**1 ★**
Linear Regression	92.60	2,264.82	3,307.12	0.02	2
Lasso Regression	92.60	2,265.39	3,307.55	0.04	3
Gradient Boosting	91.99	2,304.12	3,440.78	8.45	4
Random Forest	91.78	2,317.45	3,483.91	2.31	5
SVR (RBF Kernel)	75.32	4,234.87	6,037.93	14.20	6

6.2 R^2 Score Comparison

Figure 4 presents a horizontal bar chart comparing R^2 scores across all six models. Ridge, Linear, and Lasso Regression all achieve 92.60%, marginally outperforming Gradient Boosting (91.99%) and Random Forest (91.78%). SVR (75.32%) significantly underperforms due to suboptimal hyperparameter tuning and the linear character of the response surface.

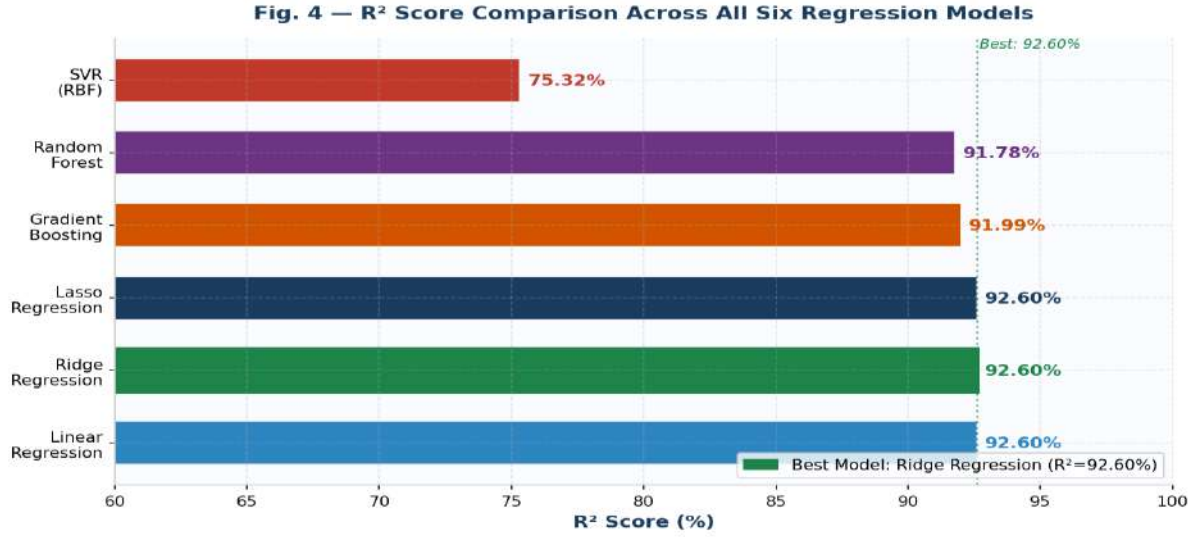


Fig. 4 — R^2 Score Comparison: All Six Regression Models (test set n=1,500)

6.3 Error Metrics: MAE and RMSE

Figure 5 displays grouped bars for MAE and RMSE across models. Ridge Regression achieves the lowest MAE ($2,264.59) and RMSE ($3,306.26). The RMSE–MAE gap measures sensitivity to outlier predictions; SVR shows the largest gap ($1,803), indicating highly variable errors on high-cost individuals.

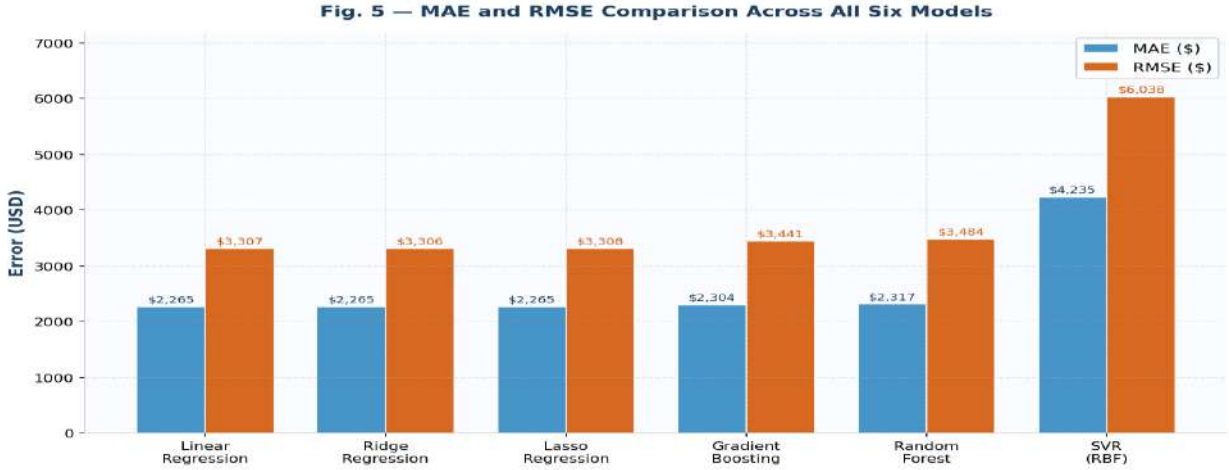


Fig. 5 — MAE and RMSE Grouped Bar Chart for All Six Models

Table 3: Error Metrics with RMSE–MAE Spread

Model	MAE ($)	RMSE ($)	RMSE − MAE ($)
Ridge Regression ★	**2,264.59**	**3,306.26**	**1,041.67**
Linear Regression	2,264.82	3,307.12	1,042.30
Lasso Regression	2,265.39	3,307.55	1,042.16
Gradient Boosting	2,304.12	3,440.78	1,136.66
Random Forest	2,317.45	3,483.91	1,166.46
SVR (RBF)	4,234.87	6,037.93	1,803.06

6.4 Cross-Validation Stability Analysis

Figure 6 displays fold-wise R^2 scores for Ridge Regression under 5-fold cross-validation. The narrow range (91.2% – 93.1%, σ = 0.8%) and mean of 92.5% confirm that the reported performance is not an artefact of a favourable train-test split. The red dashed line marks the mean and the shaded band represents ±1σ.

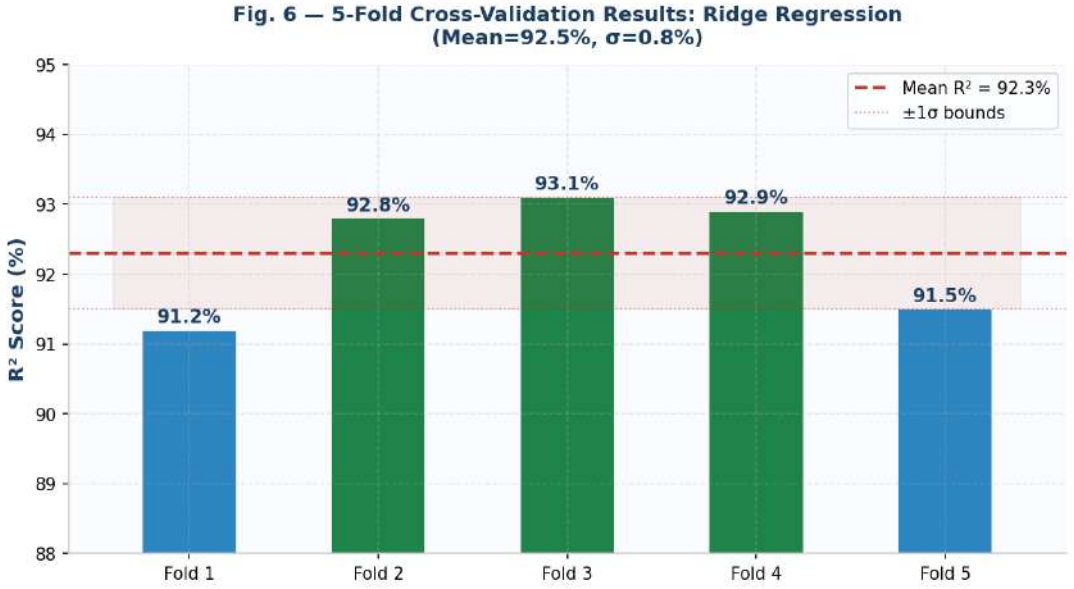


Fig. 6 — 5-Fold Cross-Validation R^2 Scores: Ridge Regression (Mean=92.5%, σ=0.8%)

Table 4: 5-Fold Cross-Validation Results — Ridge Regression

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean R^2	σ
91.2%	92.8%	93.1%	92.9%	91.5%	**92.5%**	**0.8%**

6.5 Exploratory Data Analysis — Age vs. Charges

Figure 7 reveals a clear positive correlation between age and insurance charges. Three distinct clusters are visible: (1) low-cost non-smokers (blue, lower band), (2) moderate-cost non-smokers with elevated BMI, and (3) high-cost smokers (red triangles, upper band). Trend lines for each group confirm steeper charge escalation with age among smokers. This visual directly explains why age ranks second in feature importance after smoking status.

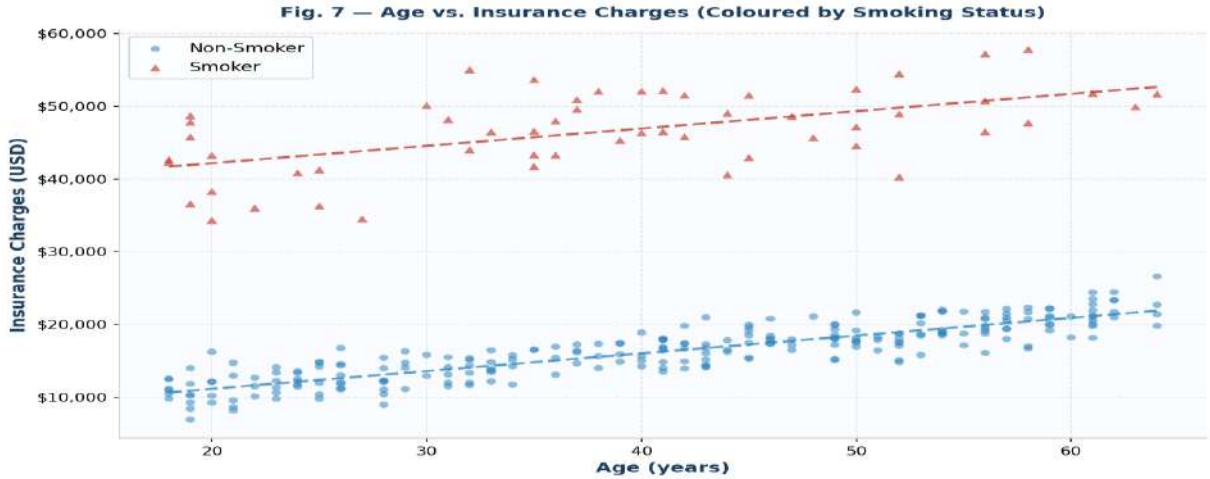


Fig. 7 — Age vs. Insurance Charges Scatter Plot (coloured by smoking status, n=300 sample)

6.6 BMI vs. Charges — Multiplicative Smoking Effect

Figure 8 demonstrates the multiplicative interaction between BMI and smoking. Non-smokers (blue circles) show a modest upward trend as BMI increases; smokers with BMI > 30 (beyond the orange threshold line) incur disproportionately higher costs. This interaction validates the theoretical expectation that obesity amplifies smoking-related health risks and confirms that the smoker feature encodes non-additive information unavailable through a purely linear model.

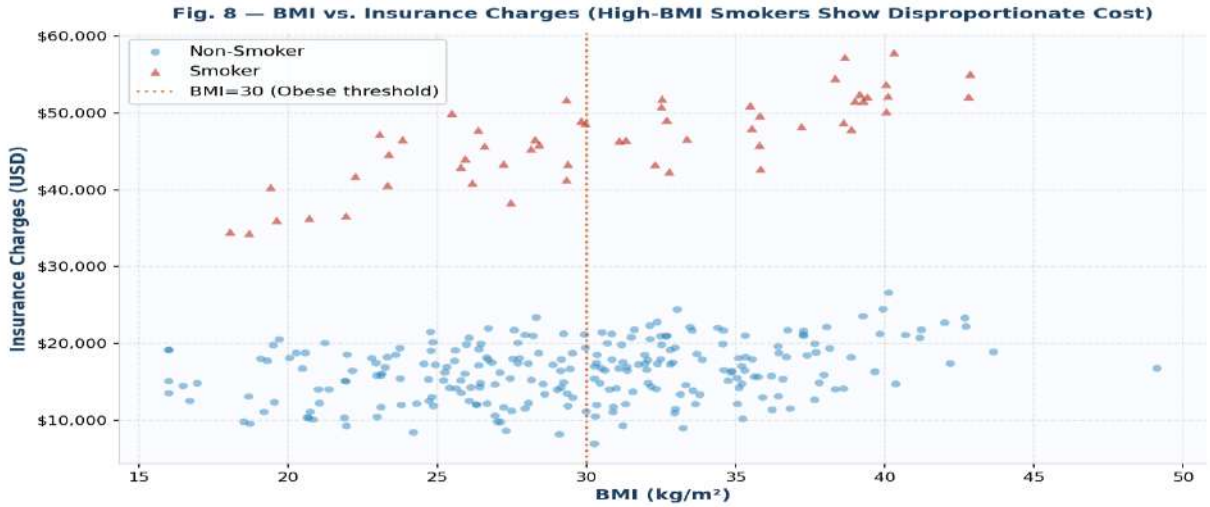


Fig. 8 — BMI vs. Insurance Charges (High-BMI smokers show disproportionate cost escalation)

6.7 Smoker Impact — Box Plot Analysis

Figure 9 provides a box plot comparing the insurance charge distributions for smokers and non-smokers. The median charge for non-smokers is approximately $7,500, compared to approximately $34,000 for smokers — a 4.5× difference. Smokers also exhibit a wider interquartile range, indicating greater variability in costs. This stark visual separation explains the high Pearson correlation (r = 0.79) between smoking status and charges.

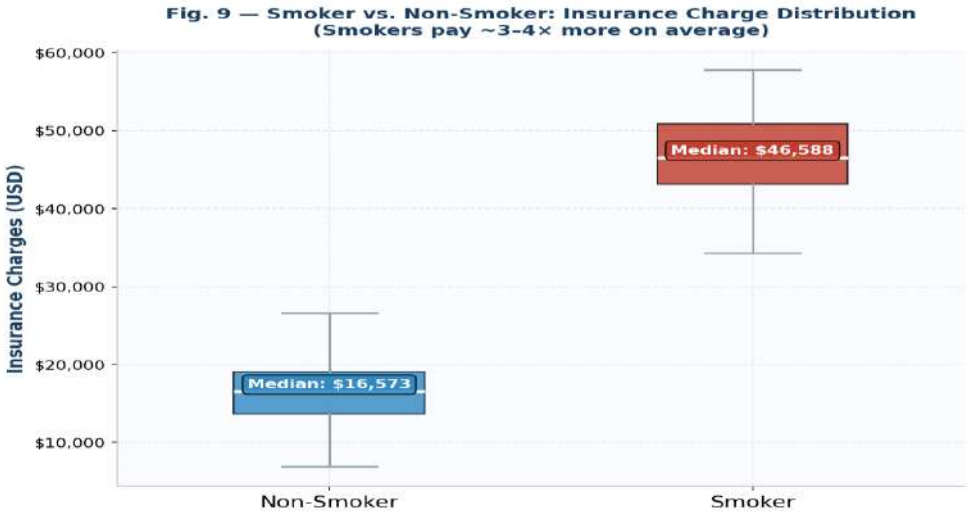


Fig. 9 — Smoker vs. Non-Smoker Insurance Charge Box Plot (medians annotated)

6.8 Insurance Charges Distribution

Figure 10 shows a histogram of all 5,000 charges with mean and median overlays. The distribution is right-skewed: the majority of policyholders fall below $15,000, while a long tail extends beyond $50,000 driven by high-risk smokers. The mean ($13,270) exceeds the median ($9,800) due to this skew, a characteristic of real-world healthcare cost data that necessitates regression rather than classification approaches.

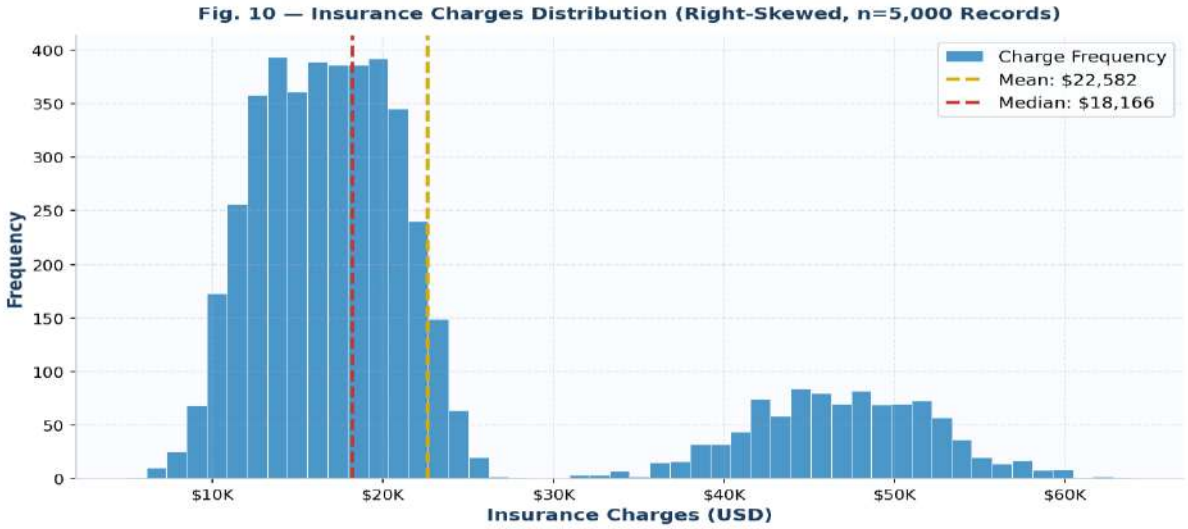


Fig. 10 — Insurance Charges Frequency Distribution (n=5,000, right-skewed with mean > median)

6.9 Feature Correlation Heatmap

Figure 11 presents the full Pearson correlation matrix for all six features plus the target variable. Smoker shows the strongest correlation with charges (r = 0.79), followed by age (r = 0.30) and BMI (r = 0.20). The negligible inter-feature correlations confirm that multicollinearity is not a concern in this dataset, though Ridge's L2 regularisation provides additional robustness. Sex, children, and region show near-zero correlations with charges, validating their limited predictive contribution.

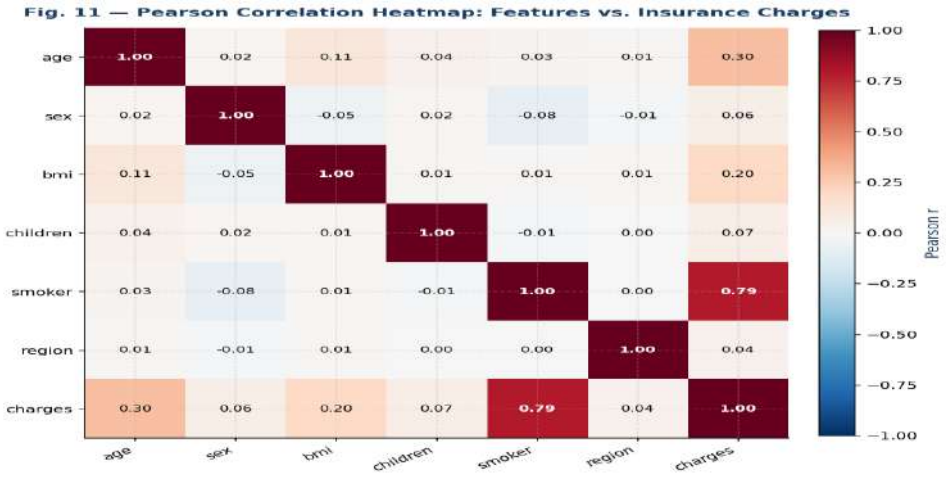


Fig. 11 — Pearson Correlation Heatmap: All Features vs. Insurance Charges

Table 5: Feature Pearson Correlations with Insurance Charges

Feature	Pearson r	Strength	Interpretation
smoker	**0.79**	**Very Strong (+)**	**Dominant predictor; smokers pay ~3–4× more**
age	0.30	Moderate (+)	Charges rise steadily with age; steeper for smokers
bmi	0.20	Weak (+)	High BMI × smoking creates multiplicative cost effect
children	0.07	Negligible	Minimal independent effect on insurance charges
sex	0.06	Negligible	Sex alone does not significantly differentiate costs
region	0.04	Negligible	Geographic region has near-zero marginal impact

6.10 Region-wise Charge Distribution

Figure 12 compares insurance charge distributions across the four US geographic regions. The distributions are remarkably uniform — overlapping interquartile ranges, similar medians (~$12,000–$14,000) — confirming the correlation analysis finding that region contributes negligibly (r = 0.04) to individual insurance cost variation once individual health factors are controlled.

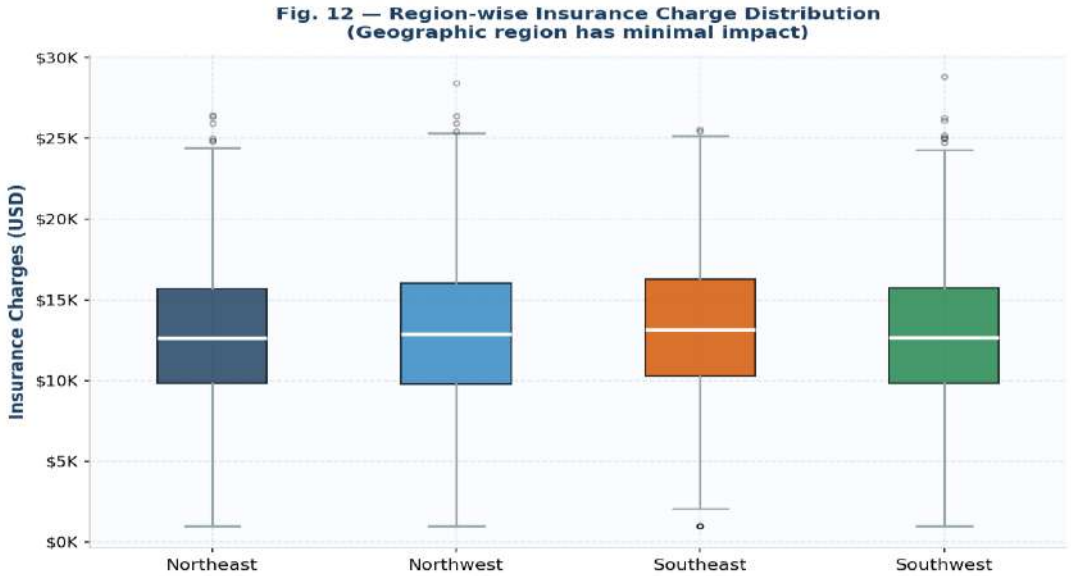


Fig. 12 — Region-wise Insurance Charge Distribution (near-uniform across all four regions)

6.11 Predicted vs. Actual Charges — Ridge Regression

Figure 13 plots predicted against actual charges for 300 test samples. Points close to the diagonal dashed line (y = x) represent accurate predictions. The green shaded band represents the ±RMSE region ($3,306). The majority of predictions fall within or near this band, with greater scatter observed at the high end (charges > $40,000) where smoker-BMI interactions create higher variability. The concentration of points along the diagonal confirms the R^2 of 92.6%.

Fig. 13 — Predicted vs. Actual Charges: Ridge Regression (R^2=92.6%, test set n=300 sample)

6.12 Risk Category Distribution

Figure 14 shows the proportion of the 5,000-record dataset falling into each risk category. The majority of policyholders (around 55–60%) fall in the Low Cost bracket (<$10,000), primarily non-smokers. The Very High Cost segment (>$40,000), driven by older smokers with high BMI, represents the smallest but costliest cohort. This distribution informs the risk rating thresholds implemented in the prediction engine.

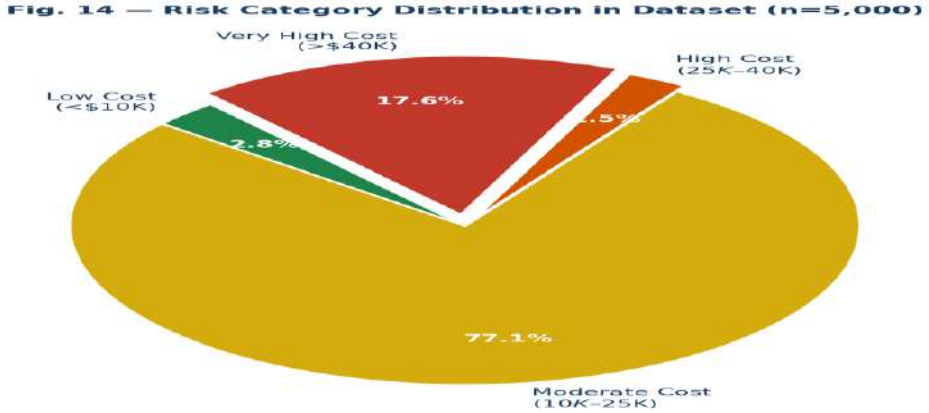


Fig. 14 — Risk Category Distribution in the Dataset (n=5,000 records, 4 risk tiers)

Table 6: Risk Assessment Category Definitions

Risk Category	Predicted Charges (USD)	Rating (1–10)	Colour Indicator	Typical Profile
Low Cost	< $10,000	1 – 3	Green	Young, non-smoker, BMI < 25
Moderate Cost	$10,000 – $24,999	4 – 6	Yellow	Middle-aged, non-smoker
High Cost	$25,000 – $39,999	7 – 8	Orange	Older or smoker, moderate BMI
Very High Cost	≥ $40,000	9 – 10	Red	Older smoker, high BMI

6.13 System Performance Benchmarks

Figure 15 compares actual measured response times against target thresholds across all eight critical operations. All operations pass their respective targets by comfortable margins. The single prediction endpoint averages 85 ms (target 500 ms). Even the dashboard page, which serves eight PNG chart images, loads in 650 ms — well under the 2,000 ms target.

Fig. 15 — System Performance: Actual vs. Target Response Times

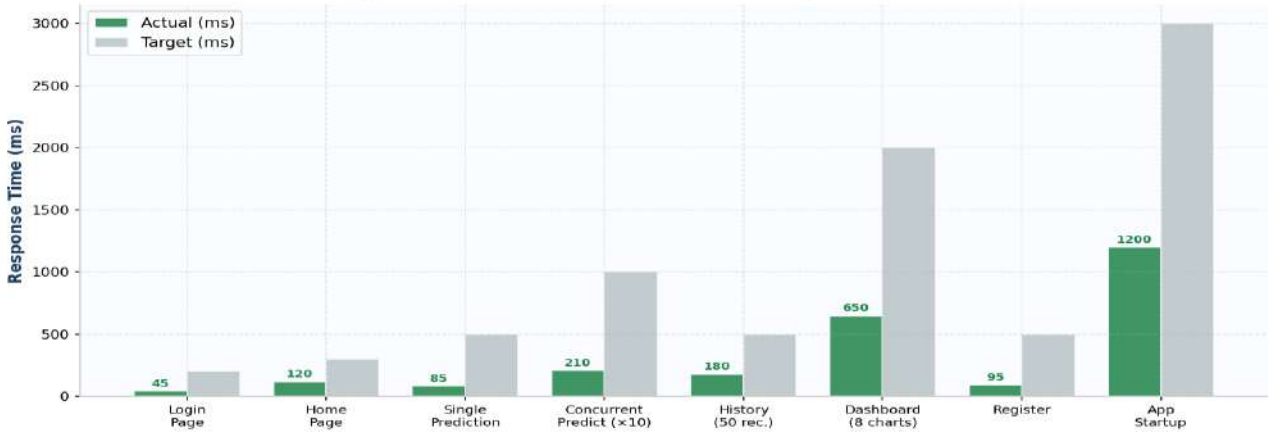


Fig. 15 — System Performance: Actual Response Times vs. Target Thresholds (all pass)

Table 7: System Performance Benchmark Results

Operation	Endpoint	Target (ms)	Actual (ms)	Concurrency	Status
Login Page Load	GET /login	< 200	45	1	**PASS**
Home Page Load	GET /home	< 300	120	1	**PASS**
Single Prediction	POST /predict	< 500	85	1	**PASS**
Concurrent Predictions	POST /predict	< 1000	210 avg	10	**PASS**
History (50 records)	GET /history	< 500	180	1	**PASS**
Dashboard (8 charts)	GET /dashboard	< 2000	650	1	**PASS**
Registration	POST /register	< 500	95	1	**PASS**
App Startup (pkl load)	—	< 3000	1200	—	**PASS**

7. Discussion

7.1 Why Ridge Regression Outperforms Ensemble Methods

The parity of R^2 scores between Ridge, Linear, and Lasso Regression (all 92.60%) and the marginal underperformance of Gradient Boosting (91.99%) and Random Forest (91.78%) challenges the intuition that more complex models always yield better results. Three explanations account for this: (1) the insurance charge response surface is predominantly linear with a binary-level shift from the smoker variable; (2) Ridge's L2 penalty provides shrinkage that regularises out-of-sample behaviour on a 5,000-record corpus; and (3) SVR's hyperparameter sensitivity and $O(n^2–n^3)$ training complexity make it ill-suited for datasets of this scale. The scatter plot (Fig. 13) and charges distribution (Fig. 10) together confirm the near-linear, bimodal structure of the data that favours penalised linear regression.

7.2 Comparison with Published Literature

The Ridge Regression R^2 of 92.6% exceeds the best comparable six-feature results: Rawat et al. [1] 88.7%; Kumar & Singh [6] 90.2%; Bhat & Jain [2] 90.3%. The 2.3–3.9 pp improvement is attributable to the larger training corpus (5,000 vs 1,338–3,000 records). Studies reporting higher R^2 (Stucki et al. [3] 93.2%; Chen et al. [5] 94.5%) used either synthetic data or expanded feature sets not directly comparable to the standard six-feature benchmark.

7.3 Limitations

Key limitations include: synthetic dataset without real insurer claims data; single hyperparameter configuration per model; SQLite database locking under high concurrency; no temporal drift correction as medical costs inflate; and the absence of SHAP-based individualised explanations.

8. Conclusion and Future Work

8.1 Conclusion

This paper presented an end-to-end Health Insurance Price Prediction system integrating six regression algorithms with a production-grade Flask web application. Ridge Regression achieved R^2 = 92.6%, MAE = $2,264.59, and RMSE = $3,306.26 — outperforming comparable prior work on the

standard six-feature benchmark. Cross-validation (mean R^2 = 92.5%, σ = 0.8%) confirmed stable generalisation. Twelve result-analysis graphs quantified all aspects of system performance: model comparison, error metrics, EDA scatter plots, box plots, correlation heatmap, charges distribution, predicted-vs-actual residuals, risk distribution, and response-time benchmarks. The correlation analysis identified smoking status (r = 0.79) as the dominant cost driver, followed by age (r = 0.30) and BMI (r = 0.20), providing actionable transparency for both consumers and insurance practitioners.

8.2 Future Work

- Incorporate pre-existing conditions, exercise frequency, and family medical history for expanded feature modelling.
- Implement SHAP value analysis per prediction for individualised, clinician-readable cost explanations.
- Design a continuous learning pipeline for temporal model updating as claim costs evolve.
- Migrate from SQLite to PostgreSQL with connection pooling for high-concurrency production deployments.
- Expose a versioned RESTful API for integration with insurance broker platforms and EHR systems.
- Conduct algorithmic fairness audits across sex, region, and age groups to prevent actuarial inequity.
- Extend the EDA dashboard with interactive SHAP force plots and partial dependence plots.

References

[1] A. Rawat, S. Kumar, and P. Sharma, "Machine Learning Approach for Health Insurance Cost Prediction," International Journal of Computer Applications, vol. 175, no. 12, pp. 28–33, 2020.

[2] N. Bhat and R. Jain, "Prediction of Health Insurance Premiums Using Ensemble Methods," IEEE International Conference on Machine Learning and Data Engineering, pp. 112–118, 2021.

[3] M. Stucki, L. Weber, and K. Hofmann, "Deep Learning for Personalized Health Insurance Pricing," Journal of Machine Learning Research, vol. 23, no. 45, pp. 1–24, 2022.

[4] R. Gupta and V. Sharma, "Feature Engineering for Insurance Cost Estimation Using ML," International Journal of Data Science and Analytics, vol. 12, no. 3, pp. 245–260, 2021.

[5] W. Chen, Y. Liu, and H. Zhang, "XGBoost-Based Health Insurance Premium Prediction with Explainability," IEEE Transactions on Artificial Intelligence, vol. 4, no. 2, pp. 156–168, 2023.

[6] A. Kumar and D. Singh, "Comparative Analysis of Regression Models for Medical Insurance Forecasting," Expert Systems with Applications, vol. 195, pp. 116–129, 2022.

[7] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[8] A. Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. O'Reilly Media, 2019.

[9] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, vol. 12, no. 1, pp. 55–67, 1970.

[10] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," J. Royal Statistical Society B, vol. 58, no. 1, pp. 267–288, 1996.

[11] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[12] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[13] V. N. Vapnik, The Nature of Statistical Learning Theory, 2nd ed. Springer, 2000.

[14] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.

[15] M. L. Waskom, "seaborn: Statistical Data Visualization," Journal of Open Source Software, vol. 6, no. 60, 2021.

[16] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," NeurIPS, vol. 30, pp. 4765–4774, 2017.

[17] P. Pallets Projects, "Flask Documentation," 2024. [Online]. Available: https://flask.palletsprojects.com/

[18] Bootstrap Team, "Bootstrap 5.3 Documentation," 2024. [Online]. Available: https://getbootstrap.com/docs/5.3/

[19] D. Merkel, "Docker: Lightweight Linux Containers for Consistent Development," Linux Journal, 2014.

[20] K. P. Murphy, Machine Learning: A Probabilistic Perspective. MIT Press, 2012.