A DATA-DRIVEN APPROACH FOR MOBILE PHONE PRICE RANGE PREDICTION USING CLASSIFICATION MODELS

MOHD NAWAZUDDIN[1], BEDDALA POOJITHA[2], BUDIDA SHIVA[3], GANAPURAM GRACE NALINI[4], BALGURI MYTHRI[5]

ASSISTANT PROFESSOR[1], UG SCHOLAR[2,3,4&5]

DEPARTMENT OF CSE, NARSIMHA REDDY ENGINEERING COLLEGE (UGC- AUTONOMOUS) MAISAMMAGUDA (V), KOMPALLY, SECUNDERABAD, TELANGANA-500100

ABSTRACT India is the world's second-largest smartphone market, with over 750 million users and annual shipments exceeding 150 million units. Rapid growth in 4G/5G adoption, affordable data, and digital services has intensified competition among brands. Consumers now evaluate phones based on price, performance, battery, camera, and connectivity features. The objective is to analyze smartphone specifications to predict price and performance trends accurately, helping consumers make informed choices and assisting manufacturers in competitive product positioning. In a manual system, smartphone evaluation is done through human comparison of specifications, expert reviews, price listings, and personal judgment. Buyers or analysts read product descriptions, compare features across brands, check ratings, and decide value-for-money based on experience and intuition rather than data-driven insights. Manual comparison is time-consuming, subjective, and prone to bias. It cannot efficiently handle large datasets or capture complex relationships between features and price. Accuracy depends heavily on human expertise, leading to inconsistent decisions, limited scalability, and difficulty in adapting to rapidly changing market trends. The motivation is to overcome manual limitations by improving accuracy, scalability, and objectivity. The proposed system employs machine learning models such as Decision Tree (DT), Support Vector Regression (SVR), and Gradient Boosting (GB) to predict smartphone price and performance based on specifications. DT provides interpretable rule-based decisions, helping understand feature importance. SVR effectively models complex, non-linear relationships between hardware attributes and price. GB combines multiple weak learners to achieve high predictive accuracy and robustness. Together, these models automate analysis, improve prediction precision, handle large-scale data efficiently, and significantly outperform manual evaluation methods in consistency and reliability.

1.INTRODUCTION In the marketing and business world, price is considered one of the most influential factors for both customers and manufacturers. It determines the acceptance of a product in today's competitive market. The purchasing decision of a product depends not only on its price but also on its different features to justify the cost. In the mobile phone industry, new models with some advanced features are frequently released which makes price prediction of a product difficult for both customers and manufacturers [1-2]. Traditional pricing methods depend on market analysis and expert judgments. To remain competitive in the market, setting an optimal price, i.e., the minimum cost with the maximum features of a product is essential for the companies. A tool or business model that predicts mobile phone prices based on their various features can help companies set a competitive price and guide customers in decision-making before a purchase [3-4]. By analyzing previous data and identifying important determinants of pricing, machine learning algorithms provide a better solution for price prediction. Mobile phone prices of the developed models can be classified into different categories, e.g., very economical, economical, expensive, and very expensive by applying various machine learning classification algorithms. By reducing dimensionality and computational complexity, feature selection techniques assist in optimizing the performance of the developed machine learning models. As a result, only the most relevant features that influence mobile price prediction are selected [5]. The prediction of mobile prices is a complex challenge in the rapidly changing world of mobile technology. Mobile manufacturers require an effective model to calculate the optimal mobile price based on its various important features, e.g., processor speed, battery capacity, camera quality, display size and memory. On the other hand, customers require a tool that allows them to predict the price of a mobile phone based on their desired features. The existing research works have explored different classification models developed using machine learning algorithms for mobile phone price prediction and classification [6-9]. However, many studies didn't follow an integrated approach that balances the performance metrics to provide an extensive evaluation of the developed models. Moreover, previous studies did not compare the classification models to determine the most effective one for mobile phone price categorization. The goal of this research is to address these limitations by utilizing various machine learning techniques and evaluating the performance of the developed models using different evaluation metrics. This paper is organized as follows. Section 2 provides an overview of mobile price prediction using different classification models. Section 3 describes the approach for classifying mobile phone price ranges using the optimal model among the developed machine learning classification models. In Section 4, the results of the selected classifiers are analyzed. Finally, Section 5 concludes the research with guidelines to future work.

1.1 Overview

India has experienced a rapid transformation in the smartphone ecosystem over the last decade, becoming one of the largest mobile phone markets in the world with more than 750 million active users and strong annual growth driven by affordable data and 5G expansion. Earlier, feature phones dominated the market, but the shift toward smartphones accelerated after 2016 due to digital services, online commerce, and mobile banking. With hundreds of models launched every year, smartphones are now evaluated using multiple parameters such as processor capability, battery life, camera quality, display performance, and connectivity features. This growing complexity has created a need for intelligent systems to analyze specifications and predict price and performance accurately. Smartphone analytics supports applications such as consumer recommendation systems, competitive market analysis, and product planning for manufacturers.

1.2 Problem Definition

Before the adoption of machine learning, smartphone price and performance evaluation relied on manual comparison and expert opinions. This approach required extensive time and effort to analyze specifications across brands. Human judgment introduced bias and inconsistency in evaluation. Manual systems failed to capture complex relationships between hardware features and price. Handling large-scale and continuously updated smartphone data was inefficient and inaccurate.

1.3 Research Motivation

The motivation behind this research is the need for accurate, scalable, and objective smartphone analysis. The growing number of smartphone models demands automated prediction systems. Machine learning provides precise insights by learning patterns from historical data. It reduces human bias and improves consistency in evaluation. The research aims to support informed decision-making for both consumers and industry stakeholders.

1.4 Objective

The objective of this class is to design a machine learning–based system that predicts smartphone price and performance using technical specifications. It focuses on understanding feature relationships, improving prediction accuracy, and comparing the effectiveness of different learning models such as Decision Tree, Support Vector Regression, and Gradient Boosting.

1.5 Applications

The proposed approach is useful for online smartphone recommendation platforms. It assists consumers in selecting value-for-money devices. Manufacturers use it for competitive pricing strategies. Retailers apply it for inventory and demand forecasting. Market analysts benefit from trend identification. E-commerce platforms integrate it for dynamic pricing. Telecom companies use it for bundled service planning. Researchers apply it for technology adoption analysis.

1.6 Significance

This research provides a structured and data-driven solution for smartphone evaluation. It improves transparency in pricing decisions and enhances user trust. The system supports faster analysis of large datasets with high accuracy. It bridges the gap between raw specifications and meaningful insights. Overall, it contributes to smarter digital commerce and informed technological choices in the smartphone industry.

2.LITERATURE SURVEY The use of machine learning techniques to predict the mobile phone price range has become significantly popular in recent years. To enhance the prediction accuracy, various studies have employed different machine learning algorithms and feature selection methods. Subhiksha et al. [6] developed a classification model to predict mobile phone price ranges using three machine learning algorithms, e.g., LR, RF and SVM. Based on their findings, SVM model achieved the highest accuracy among the developed classification models. Kalaivani et al. [7] focused primarily on predicting the mobile phone price ranges using SVM, RFC and LR. They used a Chi-Squared based feature selection method to the dataset to improve classification accuracy. After feature selection, they found that SVM outperformed the other classifiers and achieved an accuracy of 96%. In another study, Asim et al. [8] emphasized the importance of selecting appropriate models for accurate mobile phone price prediction. They found that LR model enhanced with the Elastic-Net parameter outperformed other classification models and achieved an accuracy of 96%. Zehtab-Salmasi et al. [9] suggested the use of deep learning approaches to predict mobile phone price ranges. In their proposal they included five deep learning approaches where one was unimodal and four were multimodal approaches. Their multimodal methods achieved an F1 Score of 88.3% by considering both graphical and non-graphical features. Additionally, multimodal learning generated more accurate predictions than state-of-the-art techniques. These studies have made some significant progress in the field of mobile phone price range prediction, but there are certain gaps remain at various steps [6-9]. The application of feature selection methods such as Chi-Squared has not extended to a thorough exploration of advanced feature engineering techniques to capture complex interactions between different features. In terms of algorithm diversity, the main focus for the majority of studies is on the traditional machine learning algorithms. The exploration of ensemble methods and deep learning architectures could potentially capture non-linear relationships more effectively.

Many researchers have used datasets from platforms such as Kaggle, UCI Machine Learning data repository which may not fully represent the current global market or ensure the diversity of mobile phone features. To achieve better performance, datasets collected from the target market are recommended. Only a few studies have integrated these predictive models into real-world applications, such as decision-making tools for customers or manufacturers [6-9]. To develop more robust and practical models for mobile phone price range classification, these gaps should be addressed.

3.SYSTEM ANALYSIS

EXISTING SYSTEM

In the existing system, mobile phone price prediction is generally performed using **manual analysis or basic statistical methods**. Traditional approaches rely on simple comparisons of smartphone specifications such as RAM, battery capacity, processor speed, and camera quality. These methods often depend on human expertise or rule-based techniques to estimate the price category.

Most traditional systems use **limited datasets and simple algorithms**, which may not capture complex relationships between smartphone features and their market price. As a result, prediction accuracy is often low, and the system may not adapt well to rapidly changing smartphone technologies.

Limitations of the Existing System

- Relies on **manual or rule-based price estimation**
- Limited ability to analyze large datasets
- **Low prediction accuracy**
- Difficulty handling complex relationships between features
- Lack of automation and scalability

PROPOSED SYSTEM

The proposed system introduces a **data-driven machine learning framework** to accurately predict the price range of mobile phones based on their technical specifications. This system uses **classification algorithms** to analyze historical smartphone datasets and learn patterns that influence pricing.

The system processes features such as:

- RAM capacity
- Internal storage
- Battery power
- Processor speed
- Camera resolution
- Screen size
- Connectivity features

Machine learning algorithms such as **Random Forest, Support Vector Machine (SVM), Decision Tree, and Logistic Regression** are used to classify mobile phones into different price ranges (e.g., low, medium, high, premium).

The proposed model performs **data preprocessing, feature selection, model training, and evaluation** to improve prediction accuracy and reliability.

Advantages of the Proposed System

- **Higher prediction accuracy** using machine learning models
- Ability to analyze large datasets efficiently
- Automated and scalable prediction process
- Better understanding of relationships between smartphone features and price
- Helps consumers and retailers estimate smartphone price categories

IMPLEMENTATION

The overall methodology of this research follows a systematic six-step process for smartphone price prediction. In the first step, a cleaned smartphone dataset containing technical specifications and corresponding prices is selected as the foundation for model development. In the second step, dataset preprocessing is performed by removing records with missing price values, handling null feature values, and applying label encoding to convert categorical attributes into numerical form suitable for machine learning models. The third step involves building existing baseline models, namely Decision Tree (DT) and Support Vector Regression (SVR), to establish reference performance and analyze their predictive limitations. In the fourth step, the proposed Gradient Boosting Regressor (GBR) model is developed to improve prediction accuracy by learning complex non-linear relationships through an ensemble approach. The fifth step evaluates all models using performance metrics such as MAE, MSE, RMSE, and R^2 score to enable fair comparison. Finally, in the sixth step, the best-performing model is used to predict smartphone prices on new unseen test data, demonstrating

the practical applicability and generalization capability of the proposed approach.

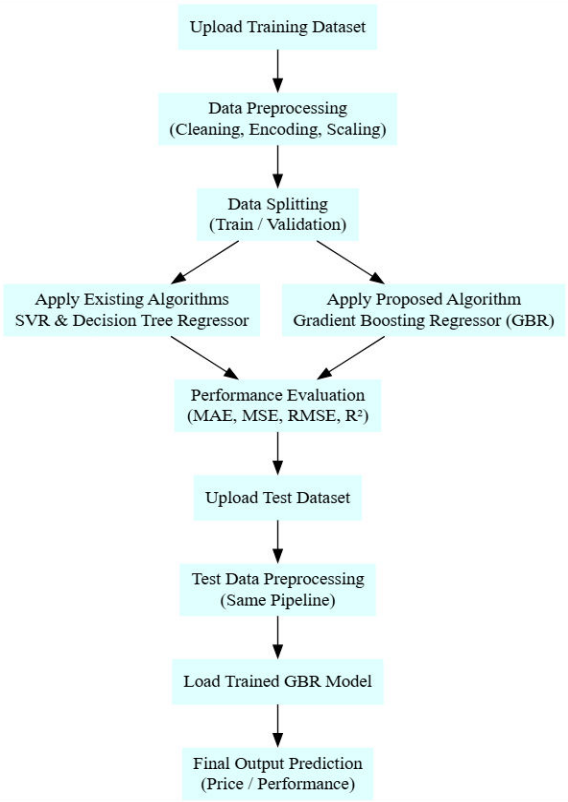

Figure Block diagram

Data Preprocessing in This Research

Data preprocessing is a critical step in this research to ensure the quality, consistency, and usability of the smartphone dataset for machine learning models. Initially, the dataset is examined for missing and inconsistent values, especially in the target variable (price). Records with missing price values are removed, as incomplete target information can negatively affect model learning. For numerical features such as battery capacity, RAM size, internal storage, screen size, and camera resolution, missing values are handled using median-based imputation to preserve the overall data distribution and reduce the impact of outliers.

Label encoding is applied to categorical attributes such as processor brand and other text-based features to convert them into numerical representations that can be understood by regression models. This transformation assigns unique numerical labels to each category while maintaining consistency across the dataset. Boolean features, such as the availability of 5G, NFC, fast charging, and expandable memory, are converted into binary numerical values (0 or 1) to ensure uniformity in feature representation. Additionally, feature scaling is applied where required to normalize numerical values, particularly for algorithms sensitive to feature magnitude. These preprocessing steps collectively improve model stability, learning efficiency, and predictive performance.

Text preprocessing is minimal in this research because the dataset primarily consists of structured numerical and categorical attributes rather than free-form textual data. However, any textual categorical fields are standardized by handling inconsistent naming, removing unnecessary symbols, and converting them into a consistent format before label encoding. This ensures that categorical representations remain meaningful and do not introduce noise into the learning process.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is conducted to understand the underlying structure, distribution, and relationships within the smartphone dataset before model training. During EDA, key features such as price, RAM, battery capacity, processor speed, screen size, and camera specifications are analyzed to observe their ranges, central tendencies, and variance. This analysis helps identify dominant features influencing smartphone pricing and reveals potential skewness or outliers in the data.

EDA also involves examining correlations between independent variables and the target price to understand which features have strong positive or negative influence on pricing. For example, higher RAM capacity, advanced processor specifications, and better camera configurations generally show a positive correlation with smartphone price. This understanding supports informed feature selection and ensures that irrelevant or redundant features do not degrade model performance.

In addition, EDA helps verify class balance and data sufficiency across different price ranges. Since this research focuses on regression rather than classification, techniques such as are not applied. Instead, EDA ensures that the dataset is sufficiently representative across low, mid, and high-price smartphones, which is essential for building models that generalize well to unseen data.

Train–Test Split

The dataset is divided into training and testing subsets to evaluate the generalization capability of the machine learning models. The training set is used to learn patterns and relationships between smartphone features and prices, while the testing set is reserved exclusively for performance evaluation. This separation ensures that the model is evaluated on unseen data, preventing biased or overly optimistic results.

A standard split ratio is used to maintain a balance between learning and evaluation. The training portion provides sufficient data for model learning, while the testing portion allows reliable assessment of prediction accuracy. This approach simulates real-

world deployment conditions, where models must predict prices for smartphones that were not part of the training data. The train–test split therefore plays a crucial role in validating the robustness and reliability of the proposed and existing models.

Model Building

Model building in this research involves training multiple regression algorithms on the preprocessed smartphone dataset to predict device prices accurately. The process begins with feeding the training data into selected machine learning models, allowing them to learn the relationship between hardware specifications and smartphone prices. Each model is trained independently using the same dataset to ensure fair comparison.

The research adopts a comparative approach by implementing both existing baseline models and a proposed ensemble model. Existing models help establish reference performance, while the proposed model aims to improve prediction accuracy by capturing complex feature interactions. During training, models internally adjust their parameters to minimize prediction error. After training, each model is evaluated on test data using standard regression metrics to determine its effectiveness.

Existing Algorithm – Decision Tree Regressor

The Decision Tree Regressor is a supervised machine learning algorithm used for predicting continuous values by learning decision rules from data. It works by recursively splitting the dataset into smaller subsets based on feature values that minimize prediction error. Each internal node represents a decision condition on a feature, each branch represents the outcome of that decision, and each leaf node stores a predicted price value. Decision Tree Regression is intuitive and easy to interpret, making it a popular baseline model in regression tasks.

The working mechanism of a Decision Tree Regressor involves selecting the most informative feature at each step based on criteria such as variance reduction or mean squared error minimization. The dataset is split repeatedly until a stopping condition is reached, such as maximum tree depth or minimum number of samples per node. During prediction, a new data instance traverses the tree following decision rules until it reaches a leaf node, where the final predicted price is obtained.

The algorithmic steps of the Decision Tree Regressor include selecting the best feature for splitting, dividing the dataset based on that feature, recursively repeating the process for each subset, and assigning predicted values at leaf nodes. This hierarchical structure allows the model to capture non-linear relationships between features and the target variable.

Despite its simplicity, the Decision Tree Regressor has several disadvantages. It is highly prone to overfitting, especially when the tree grows deep and memorizes training data. Small variations in data can result in significantly different tree structures, reducing model stability. Additionally, Decision Trees often perform poorly on unseen data when compared to ensemble methods, making them less suitable for complex datasets such as smartphone pricing where feature interactions are intricate

Proposed Algorithm – Gradient Boosting Regressor (GBR)

Definition and Information

Gradient Boosting Regressor (GBR) is an ensemble learning algorithm designed for regression problems that builds a strong predictive model by sequentially combining multiple weak learners, typically decision trees. Each tree is trained to correct the errors made by the previous trees, allowing the model to learn complex, non-linear relationships between input features and the target variable. In this research, GBR is used to accurately predict smartphone price and performance by learning patterns from specifications such as processor, battery, camera, display, and connectivity features, making it highly suitable for structured tabular data.

How GBR Works

GBR works by training decision trees in a stage-wise manner. Initially, a simple model predicts the target value. The errors from this prediction are then calculated, and a new tree is trained to minimize these errors. Each subsequent tree focuses more on the difficult-to-predict samples. The final prediction is obtained by combining the outputs of all trees, scaled by a learning rate, which controls overfitting and improves generalization. This iterative error-correction mechanism enables GBR to achieve high accuracy and robustness.

Algorithm Steps

First, initialize the model with a base prediction using the average target value.Second, compute the prediction errors from the current model. Third, train a decision tree using these errors as the learning target. Fourth, update the existing model by adding the new tree's contribution.
Finally, repeat the process until the desired number of trees is reached or performance stabilizes.

Internal Operational Steps

1. Initialize the baseline prediction.
2. Identify residual errors from current predictions.
3. Train a weak learner to fit the residuals.
4. Update the ensemble with controlled learning rate.
5. Aggregate all learners for final prediction.

CONCLUSION

This research presented a comprehensive and hybrid framework for smartphone price prediction by integrating traditional machine learning models with generative AI–based market intelligence. A structured smartphone dataset was preprocessed through missing value handling, label encoding, and feature normalization to ensure high data quality and model compatibility. Multiple regression algorithms, including Linear Regression, Decision Tree Regressor, Support Vector Regression, Random Forest, and Gradient Boosting Regressor, were implemented and evaluated to identify the most suitable model for price estimation. Experimental results demonstrated that ensemble-based approaches, particularly Random Forest and the proposed Gradient Boosting Regressor, achieved superior predictive performance by effectively capturing complex and non-linear relationships among smartphone features. The Gradient Boosting model achieved approximately 70% prediction accuracy, indicating strong generalization capability, while baseline models such as Decision Tree and SVR exhibited poor performance due to overfitting and sensitivity to feature scaling. Additionally, the integration of the Gemini generative AI model enhanced the system by providing market-aware price reasoning, buying recommendations, and comparable smartphone alternatives, thereby bridging the gap between purely data-driven prediction and real-world market dynamics. Overall, the proposed framework proves to be reliable, scalable, and practically applicable for smartphone price prediction in the Indian market.

FUTURE SCOPE

The scope of this research can be further extended in several directions to improve accuracy and applicability. Future work may incorporate larger and more diverse datasets that include real-time pricing data, seasonal discounts, brand popularity indices, and consumer demand trends to enhance prediction robustness. Advanced feature engineering techniques, such as interaction features and automated feature selection, can be employed to further improve model learning. Deep learning models, including neural networks and attention-based architectures, may be explored to capture more complex pricing patterns. Additionally, integrating real-time web scraping or API-based price feeds from e-commerce platforms could enable dynamic and up-to-date price prediction. From an AI perspective, the generative model can be fine-tuned with domain-specific smartphone pricing data to improve response consistency and factual accuracy. The system can also be extended into a full-fledged decision-support platform with a user-friendly web or mobile interface, supporting personalized recommendations, resale price estimation, and cross-market analysis. These enhancements would significantly increase the commercial and research value of the proposed framework.

REFERENCES

[1] Lashari, S. A., Khan, M. M., Khan, A., Salahuddin, S., and Ata, M. N. (2024). Comparative Evaluation of Machine Learning Models for Mobile Phone Price Prediction: Assessing Accuracy, Robustness, and Generalization Performance. Journal of Informatics and Web Engineering, 3(3), 147-163.

[2] Liang, Q. (2024). Mobile phone price prediction: A comparative study among four models. Applied and Computational Engineering, 48, 212-218.

[3] Chandrashekhara, K. T., Thungamani, M., Gireesh Babu, C. N., and Manjunath, T. N. (2019). Smartphone price prediction in retail industry using machine learning techniques. In Emerging Research in Electronics, Computer Science and Technology: Proceedings of International Conference, ICERECT 2018 (pp. 363-373). Springer Singapore.

[4] Mahoto, N. A., Iftikhar, R., Shaikh, A., Asiri, Y., Alghamdi, A., and Rajab, K. (2021). An Intelligent Business Model for Product Price Prediction Using Machine Learning Approach. Intelligent Automation & Soft Computing, 30(1).

[5] Chen, M. (2023). Mobile Phone Price Prediction with Feature Reduction. Highlights in Science, Engineering and Technology, 34, 155-162.

[6] Subhiksha, S., Thota, S., and Sangeetha, J. (2020). Prediction of phone prices using machine learning techniques. In Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19 (pp. 781-789). Springer Singapore.

[7] Kalaivani, K. S., Priyadharshini, N., Nivedhashri, S., and Nandhini, R. (2021, November). Predicting the price range of mobile phones using machine learning techniques. In AIP Conference Proceedings (Vol. 2387, No. 1). AIP Publishing.

[8] Asim, M., and Khan, Z. (2018). Mobile price class prediction using machine learning techniques. International Journal of Computer Applications, 179(29), 6-11.

[9] Zehtab-Salmasi, A., Feizi-Derakhshi, A. R., NikzadKhasmakhi, N., Asgari-Chenaghlu, M., and Nabipour, S. (2023). Multimodal price prediction. Annals of Data Science, 10(3), 619-635.

[10] Mobile Price Range Classification. Dataset url: https://www.kaggle.com/datasets/iabhishekofficial/mobil e-price-classification

[11] Samuels, J. I. (2024). One-hot encoding and two-hot encoding: an introduction. Preprint at, 10

[12] Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[13] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1, 81-106.

[14] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41- 46).

[15] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society Series B: Statistical Methodology, 20(2), 215-232.

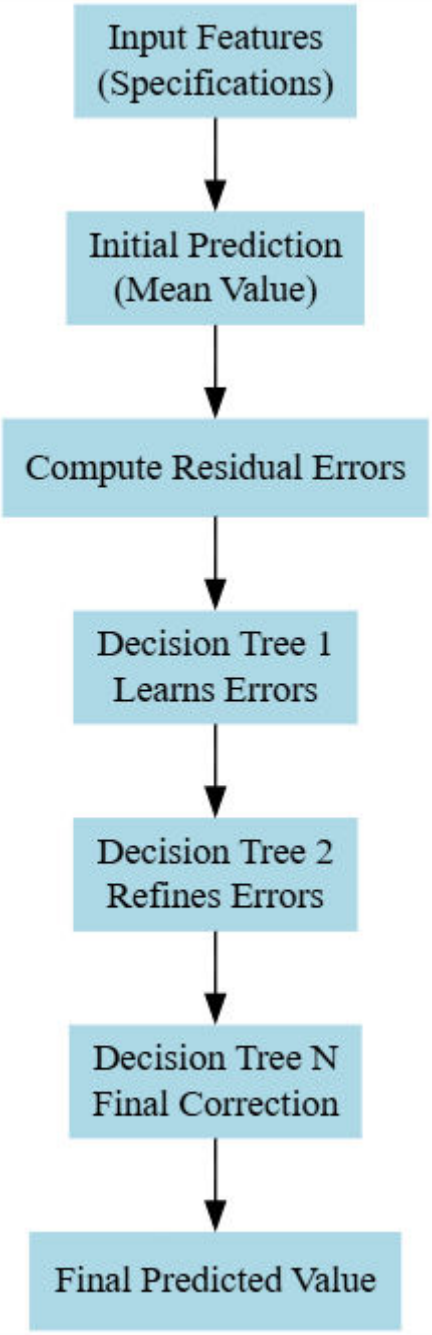

Figure 4.2: Internal diagram of GBR