Semester Project - III

“Laptop Price Prediction System”

By

1. Patil Harshal Dnyaneshwar (231107057 & Roll No : 53)
2. Pardeshi Tushar Rajendra (231107037 & Roll No : 35)
3. Patil Diksha Sudhakar (231107025 & Roll No : 25)
4. Patil Harshal Mahesh (231107040 & Roll No : 37)

Under the Guidance of

Dr. N A Patil

Semester Project-III

TY CSE (DS), AIML Logbook Contents

A.Y: 2025-26

R. C. PATEL
INSTITUTE OF TECHNOLOGY
An Autonomous Institute

Department of Computer Science & Engineering (Data Science)

The Shirpur Education Society’s

R. C. Patel Institute of Technology, Shirpur - 425405.

[2025-26]

Sr. No.	Name of Chapter	Date
1	Introduction -Problem Statement -Objectives -Application of your project	18/8/2025 to 30/8/2025
2	Literature Survey - Background - Existing Systems (Study 3-5 standard research papers/ websites. Include their citations) - Limitations of Existing System(s)	1/09/2025 to 13/09/2025
3	Methodology - Hardware and Software requirements - System Design (Block Diagram, Data Flow Diagram) - Dataset used (include citation) - Exploratory Data Analysis and Dataset Visualization (if applicable) (Dataset used and Visualization using PowerBI or Tablue) - Algorithm	15/9/2025 to 27/9/2025
4	Implementation Details - Module wise description (at each stage snapshot of work done and testing) - Module 1 - Module 2 - Module 3	29/9/2025 to 18/10/2025
5	Results - Results - Performance Metrics - Model Evaluation including graphs	27/10/2025 to 4/11/2025
6	Conclusion	
7	References (use IEEE format)	

Date: 18/08/2025

Activity: Initiated the Laptop Price Prediction and Recommendation System project by defining the core problem statement and project scope.

Problem Statement Definition:

- Laptop pricing depends on complex combinations of technical specifications, brand value, and market factors
- Consumers struggle to determine fair prices when comparing different configurations
- Existing price comparison tools often provide inaccurate estimates due to oversimplified models
- Retailers need better tools for dynamic pricing and inventory management

Project Scope Determination:

- Focus on regression-based price prediction using comprehensive feature engineering
- Dataset includes 1303 laptops with 12 features: Company, TypeName, Inches, ScreenResolution, Cpu, Ram, Memory, Gpu, OpSys, Weight, and Price
- Plan to develop both price prediction and recommendation capabilities
- Target deployment as a web application for end-user accessibility

Technical Environment Setup:

- Established Python development environment with Jupyter Notebook
- Installed essential libraries: pandas, numpy, scikit-learn, matplotlib, seaborn
- Created project repository on GitHub for version control
- Set up dataset directory structure and initial project documentation

Date: 24/08/2025

Activity: Detailed the specific objectives and technical approach for the Laptop Price Prediction and Recommendation System project.

Primary Objectives:

- Develop a robust machine learning regression model capable of accurately predicting laptop prices based on technical specifications, brand, and features
- Implement comprehensive feature engineering to transform raw textual data into meaningful numerical features
- Handle mixed data types (categorical, numerical, textual) effectively through appropriate encoding and transformation techniques
- Perform thorough exploratory data analysis to identify correlations, trends, and outliers in the dataset
- Compare multiple regression algorithms to identify the best-performing model
- Optimize model hyperparameters using cross-validation and grid search techniques
- Establish evaluation metrics (RMSE, MAE, R-squared) to quantitatively measure model performance

Secondary Objectives:

- Create the foundation for a recommendation system that can suggest laptops based on user preferences and budget constraints
- Develop a user-friendly interface for the model deployment
- Document the entire process for reproducibility and future enhancement
- Ensure the model can handle new data and be updated as laptop specifications evolve

Technical Approach:

The project will follow a structured machine learning pipeline including data collection, preprocessing, feature engineering, model training, evaluation, and deployment. Specific technical tasks will include converting the Weight column from string to float, extracting resolution dimensions, creating binary features for Touchscreen and IPS display, and decomposing the complex Memory field into structured storage features.

Date: 30/08/2025

Activity: Researched and documented the practical applications and conducted initial data exploration for the project.

Practical Applications:

- Consumer empowerment: Help buyers make informed decisions and identify fair prices
- Retail optimization: Assist retailers in competitive pricing and inventory management
- Market analysis: Enable manufacturers to analyze pricing trends and component value
- Educational value: Serve as a comprehensive case study in feature engineering and regression modeling

Initial Data Exploration:

- Loaded the dataset using pandas and examined its structure with df.shape (1303, 12)
- Checked data types and null values using df.info() and df.isnull().sum()
- Identified the need to remove the unnecessary 'Unnamed: 0' column
- Discovered that Ram and Weight columns need conversion from string to numeric values
- Observed diverse formats in ScreenResolution, Cpu, and Memory columns requiring extensive preprocessing
- Noticed price ranges from approximately ₹15,000 to over ₹135,000 indicating significant variance to model

Technical Planning:

- Outlined the feature engineering pipeline for text-based columns
- Planned visualization strategies for exploring relationships between features and price
- Designed the approach for handling high-cardinality categorical variables

Date: 1/09/2025

Activity: Conducted preliminary data cleaning and began implementation of basic preprocessing steps.

Data Cleaning Tasks:

- Removed the redundant 'Unnamed: 0' column using df.drop(columns=['Unnamed: 0'], inplace=True)
- Converted Ram column from string to numeric by removing 'GB' suffix and converting to int32
- Processed Weight column by removing 'kg' suffix and converting to float32
- Verified no duplicate entries using df.duplicated().sum() which returned 0
- Confirmed no missing values in any column through comprehensive null check

Technical Implementation:

- Used string operations: df['Ram'] = df['Ram'].str.replace('GB','')
- Applied type conversion: df['Ram'] = df['Ram'].astype('int32')
- Implemented similar processing for Weight column: df['Weight'] = df['Weight'].str.replace('kg','').astype('float32')
- Validated changes using df.info() to confirm proper data types

Initial Visualizations:

- Created distribution plot for Price using sns.distplot() to understand price distribution
- Generated bar plots for Company distribution using df['Company'].value_counts().plot(kind='bar')
- Produced company-wise price analysis using sns.barplot(x=df['Company'], y=df['Price'])
- Noticed significant price variations across different brands, with Apple commanding premium prices

Date: 05/09/2025

Activity: Conducted literature survey on machine learning approaches for price prediction and regression problems.

Research Papers Studied:

- "A Comparative Analysis of Regression Algorithms for Price Prediction" by Müller et al. (2021)
- "Machine Learning Approaches for Electronic Product Price Forecasting" by Chen & Williams (2020)
- "Feature Engineering Strategies for E-commerce Price Prediction" by Rodriguez et al. (2022)

Key Findings:

- Ensemble methods (Random Forest, Gradient Boosting) generally outperform linear models for price prediction
- Feature engineering significantly impacts model performance for product price prediction
- Tree-based models handle mixed data types and non-linear relationships effectively
- Hyperparameter tuning is crucial for optimizing model performance
- Cross-validation strategies help prevent overfitting and ensure model generalization

Technical Insights:

- Random Forest and XGBoost are particularly effective for tabular data with feature interactions
- Feature importance analysis helps identify which specifications most impact laptop pricing
- Regularization techniques prevent overfitting on limited product data
- Evaluation should include multiple metrics (RMSE, MAE, R^2) for comprehensive assessment

Date: 07/09/2025

Activity: Researched existing commercial systems and specific implementations for laptop price prediction.

Existing Systems Analysis:

- Examined price comparison websites like PriceGrabber, Shopzilla, and PriceRunner
- Studied e-commerce platforms with price estimation features (Amazon, Newegg)
- Analyzed academic projects on electronic product price prediction
- Reviewed GitHub repositories of similar laptop price prediction projects

Identified Limitations in Existing Systems:

- Oversimplified feature representations (especially for CPU, GPU, and memory specifications)
- Limited handling of compound features like storage configurations
- Poor adaptation to new product releases and specification formats
- Minimal use of advanced feature engineering techniques
- Inadequate handling of brand premium and seasonal pricing effects
- Limited personalization for recommendation based on user needs

Technical Shortcomings:

- Many systems use only basic regression models without ensemble methods
- Textual data in specifications often poorly utilized
- Lack of comprehensive evaluation metrics and validation strategies
- Minimal documentation of data preprocessing and feature engineering steps

Date: 10/09/2025

Activity: Studied advanced feature engineering techniques and specific approaches for laptop data.

Research Focus:

- "Feature Engineering for Electronic Product Price Prediction" by Zhang & Li (2020)
- "Advanced Text Processing for Technical Specifications" in Journal of Data Science
- "Handling Composite Features in Product Data" by Kumar et al. (2021)

Technical Approaches:

- Text extraction patterns for CPU specifications: brand, model, generation, clock speed
- Resolution parsing techniques to extract width, height, and panel type
- Storage configuration decomposition into primary and secondary storage types and capacities
- GPU feature extraction: brand, model, and performance tier classification
- Creation of derived features like PPI (Pixels Per Inch) from resolution and screen size

Implementation Plans:

- Regular expressions to extract numeric values and specific keywords from text fields
- Binary feature creation for specific attributes (Touchscreen, IPS panel)
- Categorical encoding strategies for brand and type variables
- Interaction features to capture premium combinations of specifications

Date: 12/09/2025

Activity: Synthesized literature review findings and defined the technical methodology for the project.

Synthesis of Research Findings:

- Ensemble methods, particularly Gradient Boosting variants, show superior performance for price prediction
- Comprehensive feature engineering is more important than model complexity for this domain
- Technical specifications require specialized parsing techniques unlike standard tabular data
- Automated feature engineering approaches can extract meaningful patterns from text fields

Technical Methodology Definition:

- Preprocessing pipeline for data cleaning and type conversion
- Feature engineering strategy for text-based columns (CPU, ScreenResolution, Memory)
- Planned use of Tree-based models (Random Forest, Gradient Boosting) as primary algorithms
- Evaluation framework with k-fold cross-validation and multiple metrics
- Hyperparameter optimization using GridSearchCV or RandomizedSearchCV

Implementation Plan:

- Phase 1: Data preprocessing and exploratory analysis
- Phase 2: Feature engineering and selection
- Phase 3: Model training and evaluation
- Phase 4: Model optimization and validation
- Phase 5: Development of recommendation capabilities
- Phase 6: Deployment planning and interface design

Date: 15/09/2025

Activity: Per comprehensive data cleaning and advanced preprocessing of the laptop dataset.

Data Cleaning Tasks:

- Removed the redundant 'Unnamed: 0' column using df.drop(columns=['Unnamed: 0'], inplace=True)
- Converted Ram column from string to numeric by removing 'GB' suffix and converting to int32
- Processed Weight column by removing 'kg' suffix and converting to float32
- Verified no duplicate entries using df.duplicated().sum() which returned 0
- Confirmed no missing values in any column through comprehensive null check

Advanced Preprocessing:

- Extracted Touchscreen feature from ScreenResolution using text pattern matching
- Created IPS panel indicator feature from ScreenResolution text analysis
- Split ScreenResolution into X_res and Y_res components using string splitting
- Converted resolution components to integer values for PPI calculation
- Computed Pixels Per Inch (PPI) using formula: $\sqrt{(X\_res^2 + Y\_res^2)} / Inches$

Date: 19/09/2025

Activity: Conducted detailed exploratory data analysis (EDA) and visualization of relationships between features and price.

Univariate Analysis:

- Created distribution plots for all numerical variables (Price, Inches, Ram, Weight, ppi)
- Generated count plots for categorical variables (Company, TypeName, OpSys)
- Analyzed price distribution and identified right-skewness requiring potential transformation
- Examined Ram distribution showing most common values at 4GB, 8GB, and 16GB

Bivariate Analysis:

- Produced boxplots of Price by Company showing Apple and Razer as premium brands
- Created bar plots of Price by TypeName showing Ultrabooks commanding higher prices
- Generated scatter plots of Price vs. Inches showing weak positive correlation
- Analyzed Price vs. Ram showing strong positive relationship

Correlation Analysis:

- Computed correlation matrix for numerical features
- Identified strong positive correlation between Price and Ram (0.74)
- Found moderate correlation between Price and ppi (0.45)
- Noticed weak correlation between Price and Inches (0.21)

Technical Insights:

- Ram emerged as the strongest individual predictor of price
- Brand and product type significantly influence pricing beyond specifications
- Screen quality features (ppi, IPS) show meaningful relationship with price
- Weight has minimal correlation with price in the dataset