BIRMINGHAM CITY
University

Student Name: Riya Shrestha(BCU)

Student ID: 24128448

Module Code: CMP5366

Module Title: Data Management and Machine Learning Operations

Co-ordinator Name: Rupak Koirala Sir

Date: April 22 2025

Table of Content

Table of Tables

Tables of Figures

Abstract- This report outlines the planning and design of data management and analytics strategy on building supervised ML model. Three real-world datasets were evaluated, with the used car prices dataset selected for its relevance and complexity. The report outlines the end-to-end pipeline, including data ingestion into MariaDB, preprocessing with Python, and preparing the data for predictive modeling. This approach highlights the practical challenges and solutions in handling and analyzing real-life data for regression problems.

Introduction

The main objective of this report is to propose a functional data pipeline and storage solution supporting supervised machine learning model. After analyzing the datasets, this report will explain how data pipeline are made, will be stored during planning process.

Source Data Analysis and Selection

Dataset 1: *Utah Real Estate Data*

This Dataset represents 4440 property listings from Utah with 14 columns, collected from Realtor.com using Apify's API as obtained via Kaggle where I found this dataset. While the originally meant for property sales created for education and analytical use.

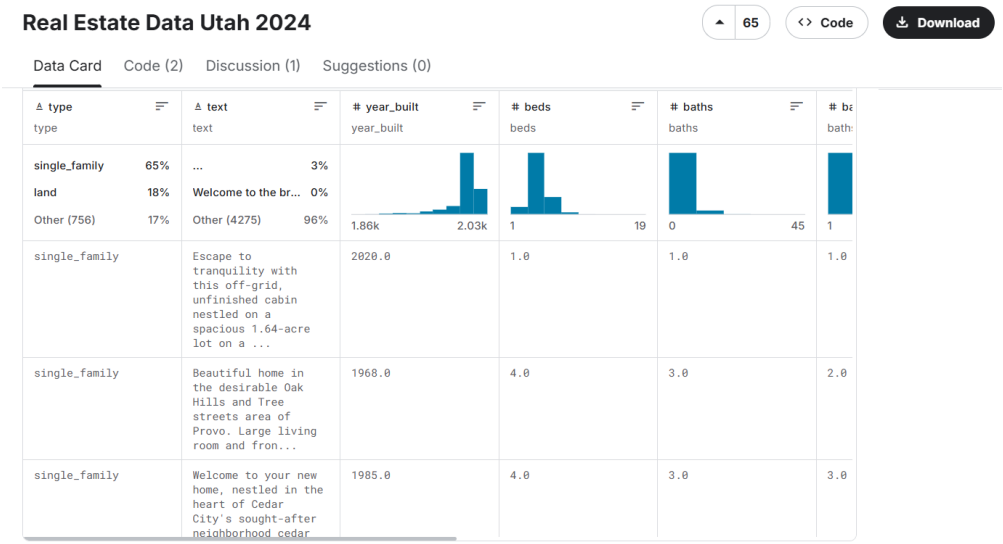

Fig 1. Dataset 1

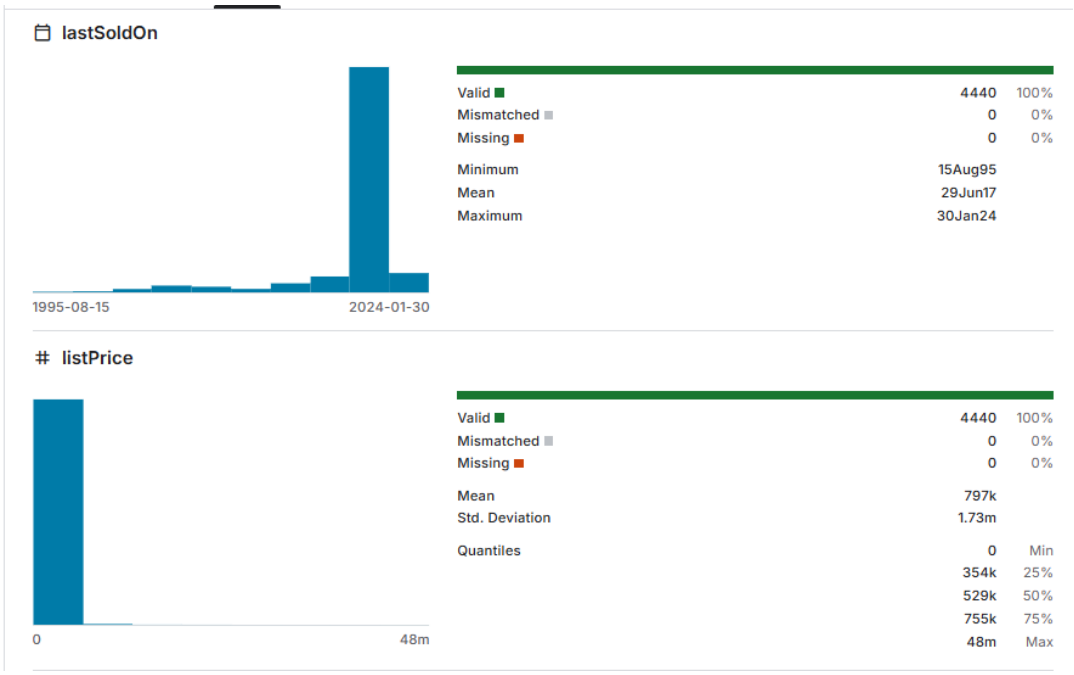

Fig 2. Dataset target variable description

It's structure is simple table flat file, currently stored as downloadable file in CSV format residing physically on system file once downloaded.

Feature	Feature Description	Data Type
type	Type of property (single-family,land)	String
text	Description of property	String
Year_built	Year the property was built	Integer
beds	Number of bedrooms	Integer
baths	Number of bathrooms	Integer
baths_full	Number of Full bathrooms	Integer
Baths_half	Number of Half bathrooms	Integer
garage	Number of garage sizes	Integer
lot_sqft	Lot size in square feet	Integer
sqft	Property size in square feet	Integer
stories	Number of stories	Integer
lastSoldOn	Date the property was last sold on	Date / String (######)
listPrice	Listing price of the property	Integer
status	Current status of the property	

Table 1: Feature Description Table of Utah Real State

This dataset is suitable for developing supervised ml models for regression task listPrice column as a target variable remaining as predictor variables . The ListPrice values act as "Ground Truth" reflecting seller listed price from realtor.com. The dataset is clean, has no missing values.

Dataset 2: *Breast Cancer Prediction*

This dataset consists of 569 records with 32 features that describes cell nuclei, originated from the university of Wisconsin hospitals from actual patient to support breast cancer diagnosis by classifying breast tumors as cancerous or not based on cell features. Originally in UCI repository but I obtained this from kaggle.

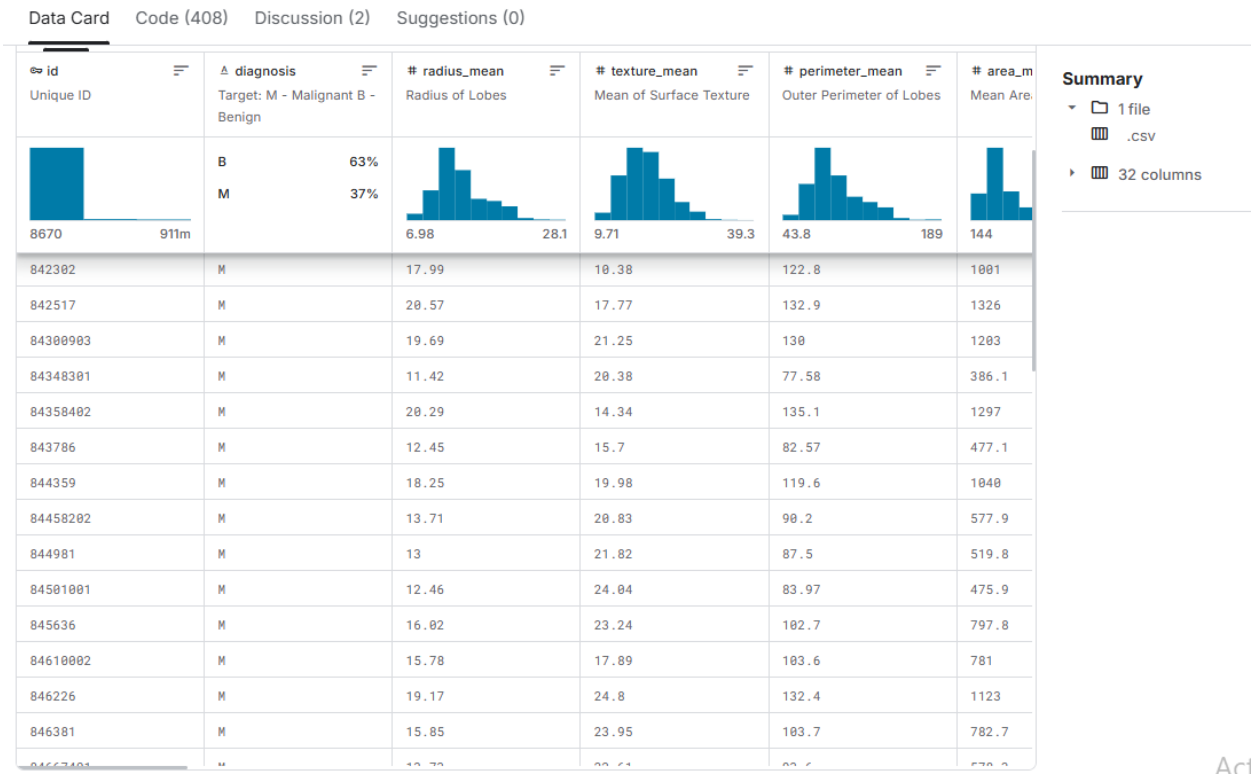

Fig 3. Dataset 2

The dataset structured ada flat CSV file, each row represents a sample and 32 columns as cell features, available for download from kaggle / UCI Repository.

Feature	**Feature Description**	**Data Type**
ID	Unique identifier number for each sample.	Integer
Diagnosis	The target variable; indicates if the tumor is Malignant (M) or Benign (B) .	String
Radius_mean	Average distance from the center to points on the perimeter of the cell nuclei.	Integer
Texture_mean	Average of the standard deviation of gray-scale values in the cell nuclei.	Integer

Perimeter_mean	Average perimeter of the cell nuclei.	Integer
Area_mean	Average area of the cell nuclei.	Integer
Smoothness_mean	Average of the local variation in radius lengths of the cell nuclei.	Integer
Compactness_mean	Average compactness (perimeter^2 / area - 1.0) of the cell nuclei.	Integer
Concavity_mean	Average severity of concave portions of the contour of the cell nuclei.	Integer
Concave Points_mean	Average number of concave portions of the contour of the cell nuclei.	Integer
Symmetry_mean	Average symmetry of the cell nuclei.	Integer
Fractal_dimension_mean	Average "coastline approximation" fractal dimension of the cell nuclei.	Integer
Radius_se	Standard error of the radius measurement.	Integer
Texture_se	Standard error of the texture measurement.	Integer
Perimeter_se	Standard error of the perimeter measurement.	Integer
Area_se	Standard error of the area measurement.	Integer
Smoothness_se	Standard error of the smoothness measurement.	Integer
Compactness_se	Standard error of the compactness measurement.	Integer
Concavity_se	Standard error of the concavity measurement	Inteer
Concave points_se	Standard error of the concave points measurement.	Integer
Symmetry_se	Standard error of the symmetry measurement.	Integer
Fractual_dimension_se	Standard error of the fractal dimension measurement.	Integer
Radius_worst	Mean of the three largest radius values found in the image.	Integer
Texture_worst	Mean of the three largest texture values found in the image	Integer
Perimeter_worst	Mean of the three largest perimeter values found in the image.	Integer
Area_worst	Mean of the three largest area values found in the image.	Integer
Smoothness_worst	Mean of the three largest smoothness values found in the image.	Integer
Compactness_worst	Mean of the three largest compactness values found in the image.	Integer
Concavity_worst	Mean of the three largest concavity values found in the image.	Integer
Concave points_worst	Mean of the three largest concave points values found in the image.	Integer
Symmetry_worst	Mean of the three largest symmetry values found in the image.	Integer
Fractual_dimesnsion_worst	Mean of the three largest fractal dimension values found in the image.	Integer

Table 2: Feature Description Table of Breast Cancer

This dataset is ideal for binary classification task, with target variable is diagnosis (Maligant = M) & (Benign = B) . The “Ground Truth” as classification of the tumor is verified through medical diagnosis & biopsy results in university.

Though it contains only 569 rows & no missing values, it is commonly used in ML projects for cancer classification tasks, providing real-life problem.

Dataset 3: *Used Car Prices*

This dataset represents information about used cars, collected from the automotive marketplace website cars.com , with various attributes to predict the price of used vehicles. It was scraped and compiled by the creator, making it available on Kaggle for analysis, research & for buyers to make informed decisions. The dataset was chosen from Kaggle itself.

This dataset stored as a single table flat file within CSV file named used_cars.csv. Each row represents the car and columns represents features, currently stored as downloadable CSV fileswhich resides physically on the user's local system.

Features	Feature Description
ID	Unique identifier for each car listing in the dataset.
Brand	The manufacturer of the car (e.g., Toyota, Ford, BMW).
Model	The specific model name of the car within the brand (e.g.,M4 Base, F-150, A8 L 55).
Model_year	The designated model year of the vehicle (e.g., 2018, 2020).
Milage	The total distance the car has been driven.
Fuel_type	The type of fuel the car's engine uses (e.g., Gasoline, Diesel, Electric, Hybrid).
Engine	Engine specifications, often including horsepower (HP) and sometimes other details like displacement or codes (e.g., 172.0HP 1., 2.7L V6 24).
Transmission	Type of transmission, often indicating Automatic (A/T) and number of speeds (e.g., A/T, 7-Speed A, 10-Speed).
Ext_col	The exterior color of the vehicle.
Int_col	The interior color of the vehicle.
Accident	Indicator of whether the car has a reported accident history .
Clean_title	Indicator of whether the car possesses a "clean" title, meaning no major negative statuses like salvage, flood damage, etc., are officially recorded .
Price	The target variable; the listing price of the used car in dollar.

Table 3: Feature Description Table of Used car price

This dataset is well suited for supervised regression task price as target variable and 12 other columns as predictors. The listed price serves as the "Ground Truth" though it is not independently verified reflecting real market conditions. Due to its origin from web scraping, it highly contains inconsistencies and missing values presenting realistic challenge in ML tasks.

	0
brand	0
model	0
model_year	0
milage	0
fuel_type	170
engine	0
transmission	0
ext_col	0
int_col	0
accident	113
clean_title	596
price	0

dtype: int64

```
Data types of columns:
brand           object
model           object
model_year       int64
milage          object
fuel_type       object
engine          object
transmission    object
ext_col         object
int_col         object
accident        object
clean_title     object
price           object
dtype: object
```

Fig 4. Dataset Missing values & Data types

	brand	model	model_year	milage	fuel_type	engine	transmission	ext_col	int_col	accident	clean_title	price
0	Ford	Utility Police Interceptor Base	2013	51,000 mi.	E85 Flex Fuel	300.0HP 3.7L V6 Cylinder Engine Flex Fuel Capa...	6-Speed A/T	Black	Black	At least 1 accident or damage reported	Yes	$10,300
1	Hyundai	Palisade SEL	2021	34,742 mi.	Gasoline	3.8L V6 24V GDI DOHC	8-Speed Automatic	Moonlight Cloud	Gray	At least 1 accident or damage reported	Yes	$38,005
2	Lexus	RX 350 RX 350	2022	22,372 mi.	Gasoline	3.5 Liter DOHC	Automatic	Blue	Black	None reported	NaN	$54,598
3	INFINITI	Q50 Hybrid Sport	2015	88,900 mi.	Hybrid	354.0HP 3.5L V6 Cylinder Engine Gas/Electric H...	7-Speed A/T	Black	Black	None reported	Yes	$15,500
4	Audi	Q3 45 S line Premium Plus	2021	9,835 mi.	Gasoline	2.0L I4 16V GDI DOHC Turbo	8-Speed Automatic	Glacier White Metallic	Black	None reported	NaN	$34,999

Fig 5. Dataset 3

5. Dataset Selection and Justification

I chose Used Car Prices dataset for it's realistic and practical applicability in predicting second-hand vehicles and suffiicent size(4009 rows).Compared to Breast Cancer dataset(569 rows, 0 missing values, 30 features) and Utah Real Estate (4440 entries, 0 missing values, 14 features), It aligns better with the used car dataset (12 features, many missing values), making it a practical and ethical choice for end-to-end process of handlin, processing, and modeling imperfect data.

Data Analytics Pipeline Design

This section details the data analytics pipeline of chosen dataset for predictions. The followings below are procedures for pipelines;

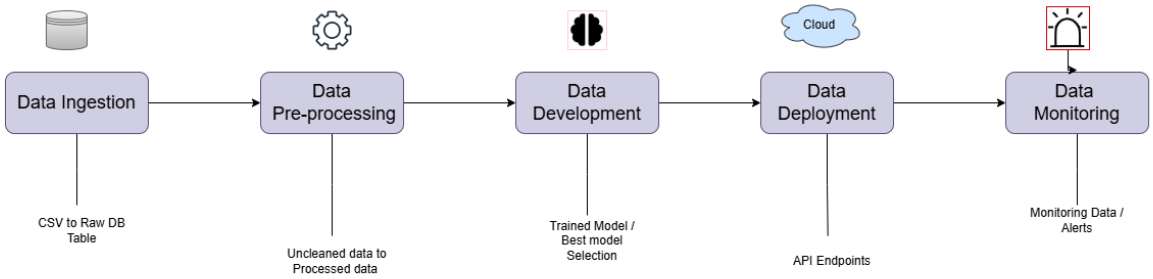

Fig 6: High Dimension Pipeline Diagram

1. Data Ingestion

The used_cars.csv file will be loaded into a MariaDB database using python and pandas. This makes the raw data available for futher processing. Docker along with Anaconda will ensure consistent execution of ingestion in controlled environment inside container. Data engineers will manage this process for smooth running.

2. Data Pre-processing

In this step, raw ingested data will be cleaned & fixed. Missing values, incorrect datatypes, unwanted characters like mi. , HP, $ will be removed & categorical value will be encoded into numerical values. **Feature engineering** would be involved calculating car's ageand applying imputation techniques, The cleaned data will be saved passing directly to dataframe, tracked in ml flow and passed for development. Data Scientists will design logic, handle features, and choose cleaning methods; Data Engineers help automate and improve the process.

3. Model Development

This stage involves using pre-processed data to train, optimize, evaluate and select the best regression model for predicting car price.The process involves loading data, splitting features target variables, training models, tuning hyperparameters, and evaluating performance using RMSE and R^2 and k-fold cross-validation. LIME and SHAP will explain models predictions. The best model will be retrained and registered in MLflow for future deployment.

Data scientists handle model building, tuning, and explainability, with ML engineers supporting infrastructure and tracking. The deployment team, business stakeholders, and analysts benefit from accurate predictions.

4. Model Deployment

This process takes the best model to operate & will turn into an API using Fast API so it can serve as prediction system. Docker will be used for packaging & it's input dependencies to ensure consistency.The API will be hosted on a server or cloud for users. ML Engineers and DevOps will work together to manage deployment and infrastructure.

5. Model Monitoring

After deployment the model's performance will be monitored to ensure its performance if any changes in data patterns or market trends affect predictions. Monitoting this ensures the model is accurate & responsive, with timely retraining if needed. Performance metrics and prediction times will be tracked to ensure reliability over time.

Data Storage Strategy For Analytics

The dataset willl be stored in MariaDB(Docker) using star schemna with a central table linked to dimension tables. The ETL process will be used as data will be extracted from CSV, transform into processed dataset and load it to dataframe. Star Schema is chosen for its query speed and suitable for time related analytics offering better performance for used car prediction than Other alternatives. It organizes data ibto central fact table linked to dimension tables, maintained by data engineers and data scientists for model building.

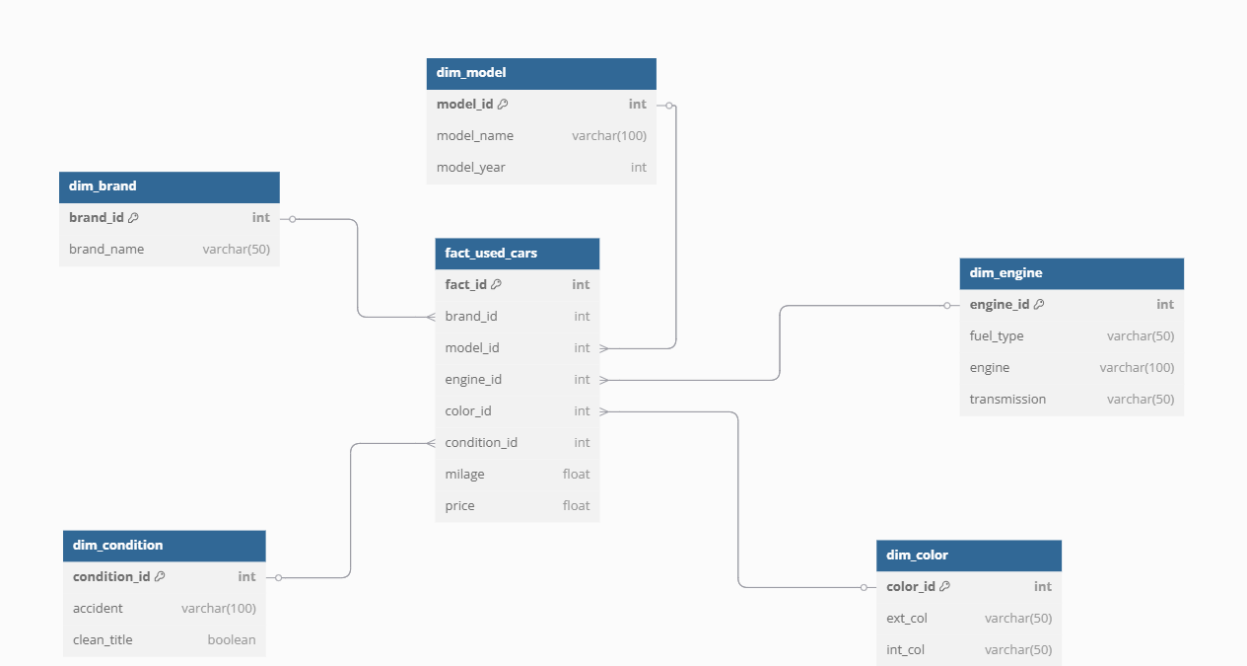

Fig 7: Logical Star Schema diagram

used_cars	
string	brand
Object	model
int	model_year
Object	milage
Object	fuel_type
Object	engine
Object	transmission
Object	ext_col
Object	int_col
Object	accident
Object	clean_title
Object	price

Fig 7. Used Car Price Dataset Datatype

Benefits	Drawbacks
Simple queries makes it easier to understand.	Redundancy can occur increasing storage.
Reduces the number of joins, making queries faster.	If dimension data changes, it may require updating in multiple places across the database.
The clear distinction between the fact table and dimension tables makes it easier for analysts and data scientists to navigate the data.	With denormalization, additional storage may be required to hold redundant data.
It is designed for fast data aggregation, which is great for insights and decision-making.	

Table 4. Benefits & drawbacks of star schema structure.

Structure	Benefits	Drawbacks
One Big Table	Simple structure, no joins required.	Poor performance with large datasets, harder to maintain, slow queries
Snowflake	Reduces data redundancy, maintains normalization.	More complex queries, slower performance due to more joins
Normalized Relational	No redundancy, ensures data integrity, smaller storage requirements.	Complex queries, slower OLAP performance, difficult schema maintenance.

Table 5. Benefits & drawbacks of alternative structure.

This approach ensures data integrity and streamlines the workflow from raw data to analysis.

Reflection

While preparing this pipeline, i faced challenges in clearly defining the problem and choosing the right techniques without yet applying them. It was difficult to plan each step confidently without working with data. This process improved my ability to think ahead, structure my approach and envision future obstacles.

Further Plan

Moving forward, I plan to apply this pipeline for making used car price prediction system after selecting best model approach. Then, deploy & monitor it’s performance ensuring the model is accurate, practical and easy to use.

References

Najib, T. (2023) *Used car price prediction dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset/data (Accessed: 25 April 2025).

Kanchana1990 (2024) *Real estate data Utah 2024*, *Kaggle*. Available at: https://www.kaggle.com/datasets/kanchana1990/real-estate-data-utah-2024/data (Accessed: 25 April 2025).

H, M.Y. (2021) *Breast cancer dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset (Accessed: 25 April 2025).

N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, 2018, pp. 115-119, doi: 10.1109/ICBIR.2018.8391177. keywords: {Automobiles;Regression tree analysis;Linear regression;Predictive models;Data models;Vegetation;Forestry;comparative study;multiple linear regression;random forest;gradient boosting;supervised learning},

Adhikary, D.R.D., Sahu, R. and Panda, S.P. (1970) *Prediction of used car prices using machine learning*, *SpringerLink*. Available at: https://link.springer.com/chapter/10.1007/978-981-16-8739-6_11 (Accessed: 25 April 2025).

M. Hankar, M. Birjali and A. Beni-Hssane, "Used Car Price Prediction using Machine Learning: A Case Study," 2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC), El Jadida, Morocco, 2022, pp. 1-4, doi: 10.1109/ISIVC54825.2022.9800719. keywords: {Maximum likelihood estimation;Linear regression;Training data;Production;Pricing;Predictive models;Boosting;regression analysis;prediction;estimation;Avito;machine learning;log transformation;used car price;regression assumptions},