Project Report: Predicting Age of Possums

1. **Introduction**

The objective of this project was to develop a regression model to predict the age of possums using a dataset containing 104 entries and 14 features. The dataset underwent preprocessing to handle missing values, encode categorical variables, and scale numerical features before training multiple regression models. This is how dataset is described.

	case	site	Pop	sex	age	hdlngth	skullw	totlngth	taill	footlgth	earconch	eye	chest	belly
0	1	1	Vic	m	8.0	94.1	60.4	89.0	36.0	74.5	54.5	15.2	28.0	36.0
1	2	1	Vic	f	6.0	92.5	57.6	91.5	36.5	72.5	51.2	16.0	28.5	33.0
2	3	1	Vic	f	6.0	94.0	60.0	95.5	39.0	75.4	51.9	15.5	30.0	34.0
3	4	1	Vic	f	6.0	93.2	57.1	92.0	38.0	76.1	52.2	15.2	28.0	34.0
4	5	1	Vic	f	2.0	91.5	56.3	85.5	36.0	71.0	53.2	15.1	28.5	33.0

Each and every column was explored before proceeding with the project and here is the full definion of each columns.

- **case**: A unique identifier for each possum (e.g., case number).
- **site**: The location or site where the possum was observed or captured.
- **Pop**: Population or region (e.g., "Vic" might stand for Victoria, Australia).
- **sex**: The gender of the possum (m for male, f for female).
- **age**: The age of the possum (in years or months, depending on the dataset).
- **hdlngth**: Head length (likely measured in millimeters or centimeters).
- **skullw**: Skull width (likely measured in millimeters or centimeters).
- **totlngth**: Total body length (likely measured in millimeters or centimeters).
- **taill**: Tail length (likely measured in millimeters or centimeters).
- **footlgth**: Foot length (likely measured in millimeters or centimeters).
- **earconch**: Ear conch length (likely measured in millimeters or centimeters).
- **eye**: Eye diameter or size (likely measured in millimeters or centimeters).
- **chest**: Chest circumference or width (likely measured in millimeters or centimeters).
- **belly**: Belly circumference or width (likely measured in millimeters or centimeters).

2. **Data Preprocessing**

- *Handling Missing Values:* Missing numerical values were imputed using the mean (as no outliers were detected), while categorical missing values were replaced using the mode.

- *Encoding:* Label Encoding was applied to categorical features, and One-Hot Encoding was used where necessary to convert categorical variables into numerical format.

- *Feature Scaling*: StandardScaler was applied to numerical features to standardize their range and improve model performance.

3. **Model Selection and Evaluation**

The following regression models were trained and evaluated using Mean Squared Error (MSE) and R-squared (R^2) scores:

	Model	MSE score	R2 Score
0	LinearRegression	1.873963	-0.383979
1	Random Forest	1.108602	0.181263
2	Decision Tree	0.811631	0.400586

4. Results Analysis

- DecisionTreeRegressor achieved the best performance with the lowest MSE (1.07) and the highest R^2 score (0.20), indicating a slightly better fit compared to other models.

- RandomForestRegressor performed similarly with an MSE of 1.09 and an R^2 score of 0.18.

- LinearRegression underperformed with a high MSE of 1.87 and a negative R^2 score (-0.38), suggesting poor predictive power.

- LGBMRegressor had the highest MSE (3.81) despite an R^2 score of 0.20, indicating that it might not be well-suited for this dataset.

5. Conclusion and Future Improvements

The DecisionTreeRegressor provided the best overall performance, but the R^2 scores across all models indicate limited predictive power. Potential improvements include:

- Feature Engineering: Exploring additional relevant features or transformations to enhance model performance.

- Hyperparameter Tuning: Optimizing hyperparameters, especially for tree-based models, to improve accuracy.

- Ensemble Methods: Combining models (e.g., stacking or boosting) to leverage strengths from multiple approaches.

- Data Augmentation: Acquiring more data or synthetic data generation to improve model learning.

Further experimentation with these techniques could enhance the model's predictive capabilities and achieve better results in future iterations.