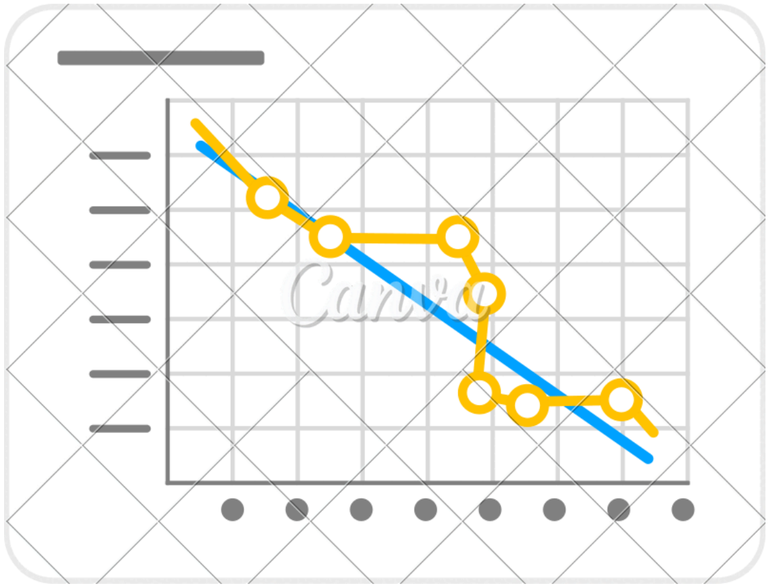


REGRESSION ALGORITHM

Group Members:
Omkar Satupe (9232)
Ryan Valiaparambil (9237)
Mahek Intwala (9423)

PAPER DETAILS

- Title : A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position

- Research paper Link: Click here

- Colab File Implementation: Click here

AGENDA

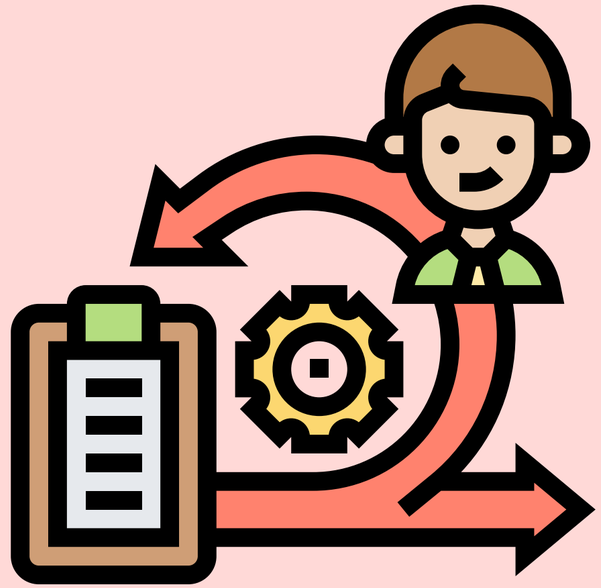
1. Problem Statement
2. Methodology & Algorithms
3. Paper Results and Conclusion
4. Implementation & Results

PROBLEM STATEMENT

- The problem statement for this study revolves around the estimation of the market values of the football players in the forward positions by including the physical and performance factors in 2017-2018 season.
- The paper aims to address the challenge of understanding the multitude of factors that influence the market value of football players in forward positions. It seeks to identify and analyze the key determinants such as player performance metrics, age, contract duration, club reputation, and market trends that significantly impact player valuation.
- To address this challenge, the paper proposes the use of Multiple Linear Regression Approach.
- This supervised learning approach provides recommendations and guidelines for football clubs, agents, and analysts involved in player valuation.

METHODOLOGY

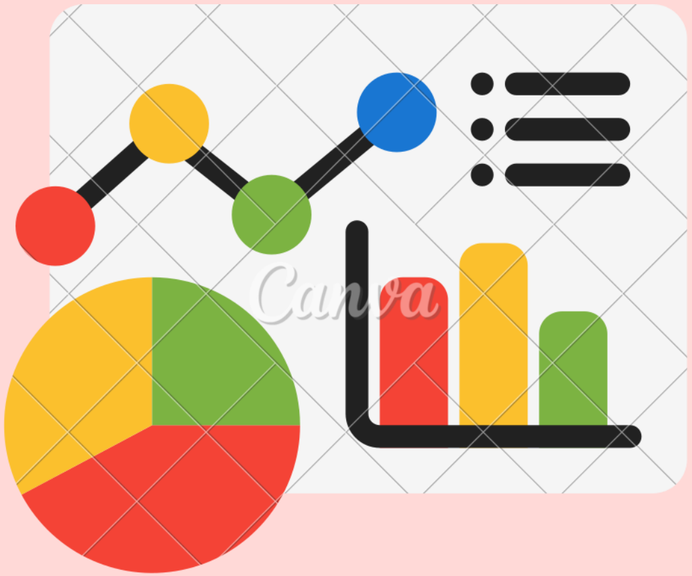
- In this research, the methodology revolved around predicting the market value of forward-position football players based on their attributes.
- The process began with meticulous data collection and preprocessing, involving the selection of relevant player characteristics from major European leagues while ensuring data quality.
- Feature selection was then employed to reduce complexity and multicollinearity, resulting in the creation of models with varying attribute subsets.
- The research rigorously assessed homoscedasticity to maintain the reliability of regression analysis. Ultimately, the most suitable model was chosen based on performance metrics, such as adjusted R-squared and Mean Absolute Percentage Error (MAPE), leading to the selection of a model with 52 attributes as the optimal choice for estimating player market values.
- This comprehensive methodology encompassed data preparation, feature selection, statistical assessment, and model selection, contributing to a robust approach for predicting football player market values.



ALGORITHMS USED : Multiple Linear Regression

PAPER'S RESULTS

- The study successfully developed a regression model with 52 attributes at a significance level of 0.10. This model demonstrated a 20% Mean Absolute Percentage Error (MAPE) and an adjusted R-squared value of 0.86. Despite inherent multicollinearity in the data, the model provided reasonable predictive accuracy.
- It revealed that certain player attributes significantly influenced their market value. Notably, younger players and those aged between 20 and 21 were found to be more valuable. Additionally, players with heights between 180 and 184 centimeters were considered more valuable, suggesting a balance between physical attributes and technical skills.
- It also confirmed the influence of league and nationality on player valuation.
- Interestingly, the analysis did not find a significant correlation between card scores (yellow and red cards) and player market value. This unexpected result could be attributed to valuable players exercising caution to avoid penalties, highlighting the complexity of factors affecting player valuation beyond performance indicators.



OLS Regression Results

Dep. Variable:	y	R-squared:	0.930
Model:	OLS	Adj. R-squared:	0.860
Method:	Least Squares	F-statistic:	13.29
Date:	Mon, 04 Jun 2018	Prob (F-statistic):	4.05e-17
Time:	08:57:21	Log-Likelihood:	-363.75
No. Observations:	105	AIC:	833.5
Df Residuals:	52	BIC:	974.2
Df Model:	52		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-49.4294	12.768	-3.871	0.000	-75.051	-23.808
x1	13.3396	7.285	1.831	0.073	-1.279	27.958
x2	35.7294	14.381	2.484	0.016	6.871	64.587
x3	51.3703	6.942	7.400	0.000	37.440	65.301
x4	-23.0743	7.936	-2.907	0.005	-39.000	-7.149
x5	23.2793	7.642	3.046	0.004	7.944	38.615
x6	-25.6298	8.807	-2.910	0.005	-43.303	-7.957
x7	-20.7825	8.644	-2.404	0.020	-38.129	-3.436
x8	27.0538	7.027	3.850	0.000	12.952	41.155
x9	31.7221	14.024	2.262	0.028	3.580	59.864
x10	-18.6637	9.786	-1.907	0.062	-38.301	0.973
x11	22.4513	6.597	3.403	0.001	9.213	35.690
x12	-27.5489	12.642	-2.179	0.034	-52.916	-2.182
x13	19.2354	6.954	2.766	0.008	5.281	33.190
x14	-62.7216	16.277	-3.853	0.000	-95.384	-30.059
x15	72.6052	13.081	5.550	0.000	46.355	98.855
x16	38.4339	9.173	4.190	0.000	20.028	56.840

Figure 4: Part of regression summary of the data within the feature selection at 0.1 significance

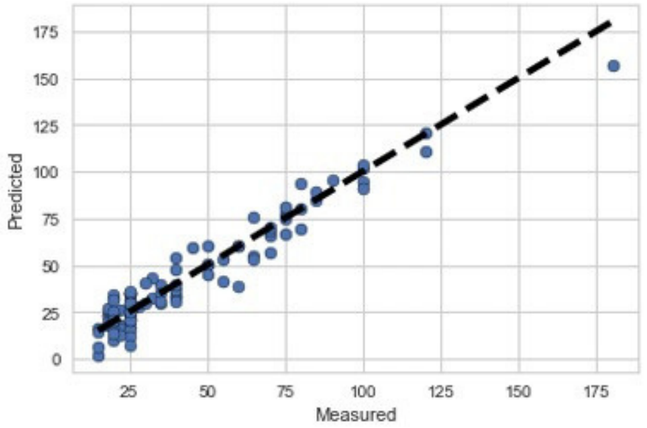


Figure 5: Measured / Predicted of the data within the feature selection at 0.1 significance

Dep. Variable:	y	R-squared:	0.944
Model:	OLS	Adj. R-squared:	0.927
Method:	Least Squares	F-statistic:	53.99
Date:	Mon, 04 Jun 2018	Prob (F-statistic):	4.30e-40
Time:	09:14:28	Log-Likelihood:	-410.63
No. Observations:	105	AIC:	871.3
Df Residuals:	80	BIC:	937.6
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	57.4574	7.957	7.221	0.000	41.622	73.293
x2	-22.3508	10.342	-2.161	0.034	-42.931	-1.770
x3	19.9860	7.284	2.744	0.007	5.490	34.482
x4	33.8492	6.505	5.204	0.000	20.905	46.794
x5	-28.8772	11.582	-2.493	0.015	-51.925	-5.829
x6	24.0523	6.619	3.634	0.000	10.879	37.225
x7	-34.7427	15.988	-2.173	0.033	-66.559	-2.927
x8	68.4578	14.596	4.690	0.000	39.412	97.504
x9	20.6690	7.771	2.660	0.009	5.204	36.134
x10	24.9943	8.345	2.995	0.004	8.387	41.601
x11	12.2860	5.860	2.097	0.039	0.625	23.948
x12	15.2947	4.639	3.297	0.001	6.062	24.527
x13	64.5626	21.038	3.069	0.003	22.696	106.429
x14	11.8840	4.083	2.911	0.005	3.759	20.009
x15	12.2368	3.411	3.587	0.001	5.449	19.025
x16	15.4844	7.583	2.042	0.044	0.395	30.574
x17	44.3643	18.197	2.438	0.017	8.151	80.577
x18	78.8110	14.831	5.314	0.000	49.296	108.326

Figure 7: Part of regression summary of the data within the feature selection at 0.05 significance

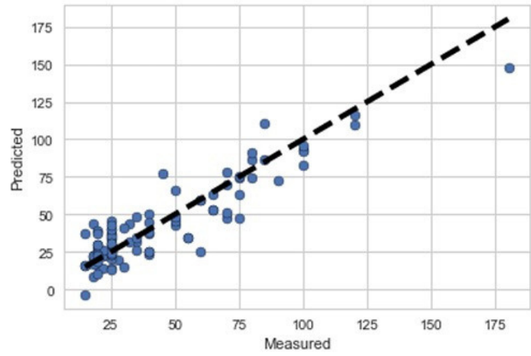
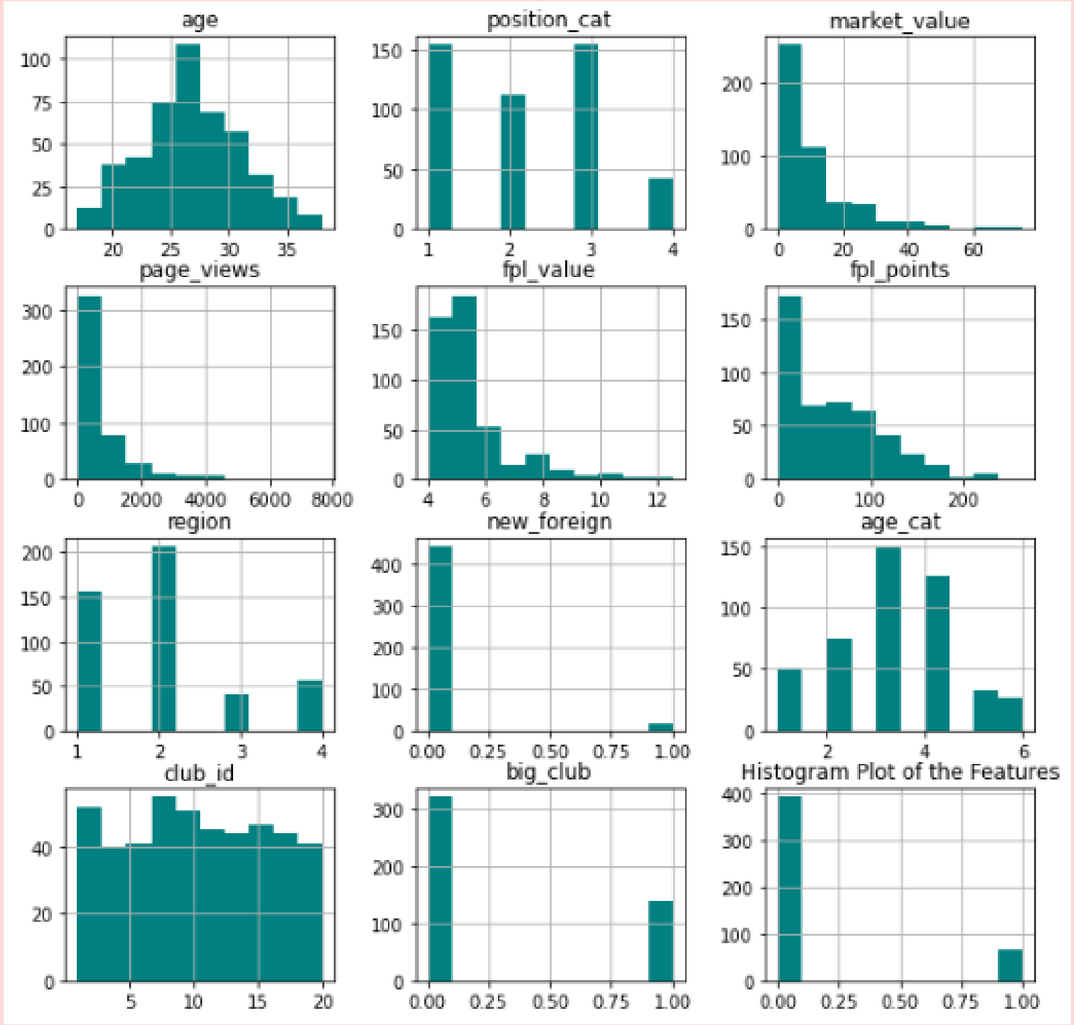
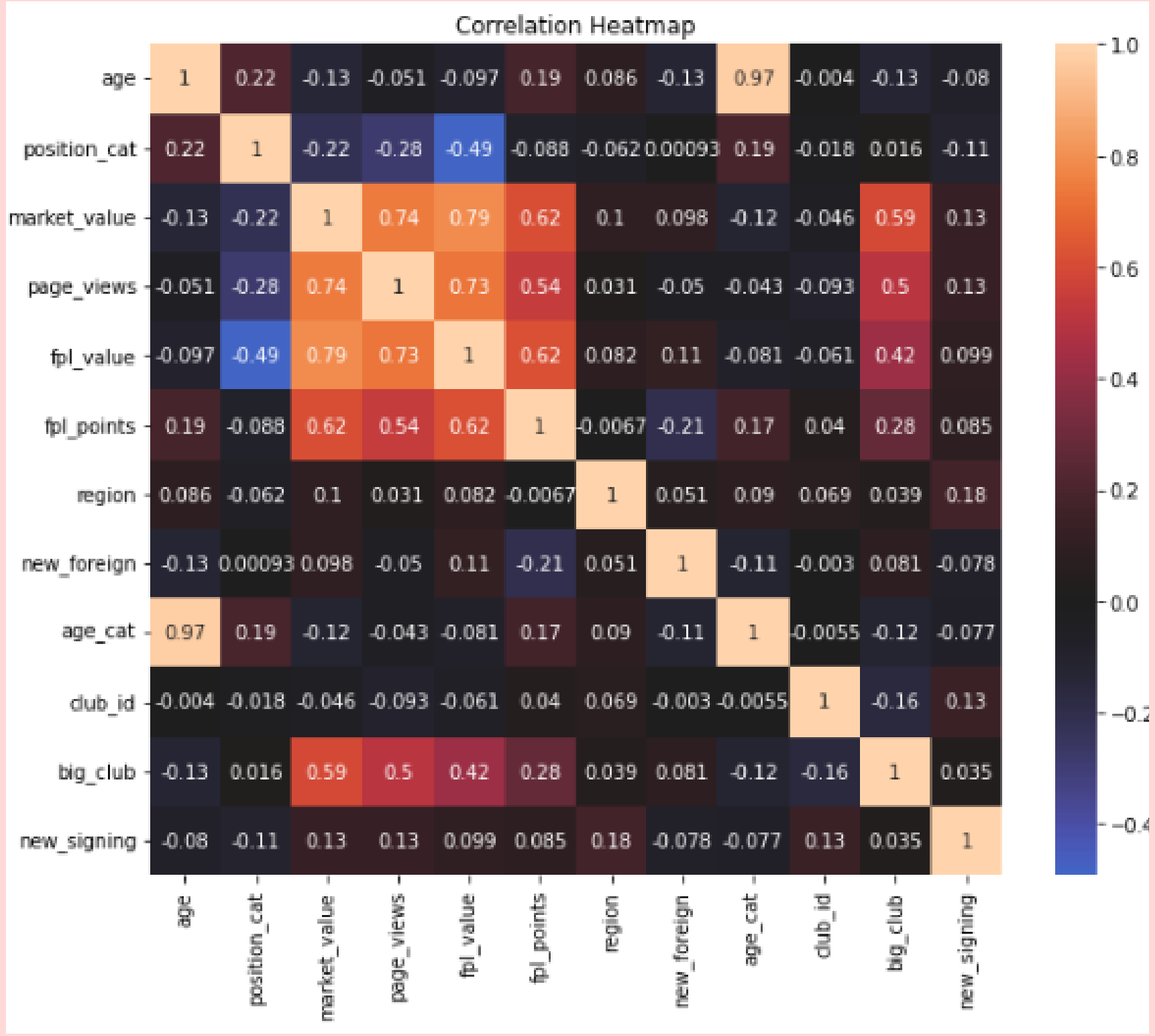


Figure 8: Measured / Predicted plot of the data within the feature selection at 0.05 significance

OUR IMPLEMENTATION





RESULTS

	OUR RESULTS	
	Linear Regression	**XGB Regressor**
Mean Squared Error (MSE)	47.751506850207 11	26.430800790371112
Mean Absolute Error (MAE)	4.8213286807739 815	3.2392628862805988
R-squared	0.71621417242767	0.8429225134355665

	PAPER RESULTS	
	OLS Regression with feature selection at 0.1 significance	**OLS Regression with feature selection at 0.05 significance**
Mean Absolute Percentage Error (MAPE)	20	27
R-squared	93.256	94.44
Adjusted R-squared	86.346	92.756

CLASSIFICATION ALGORITHM

PAPER DETAILS

- Title : Prediction of Mobile Phone Price Class using Supervised Machine Learning Techniques
- Publication Year: January 2022
- Research paper Link: Click here
- Dataset Link: Click here
- Colab File Implementation: Click here

AGENDA

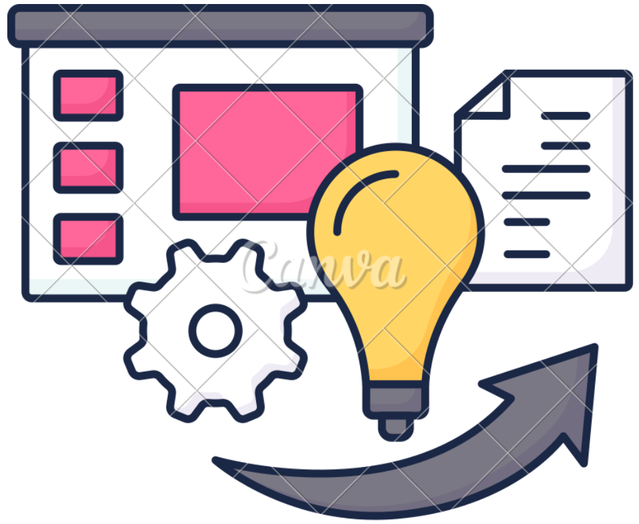
1. Problem Statement
2. Methodology & Algorithms
3. Paper Results and Conclusion
4. Implementation & Results

PROBLEM STATEMENT

- This research paper focuses on the problem of accurately predicting mobile phone prices based on their specifications and identifying the most effective machine learning algorithm for this task.
- Price estimation is crucial in the dynamic mobile technology market, and the paper employs predictive analytics and supervised machine learning to develop a model that estimates mobile phone prices using feature attributes.
- The study utilizes the Mobile Price Classification dataset from Kaggle and various classification algorithms in Python to train and evaluate the model's performance.
- The ultimate goal is to provide a valuable tool for both mobile companies and consumers to make informed pricing decisions in a rapidly evolving and competitive market.

METHODOLOGY

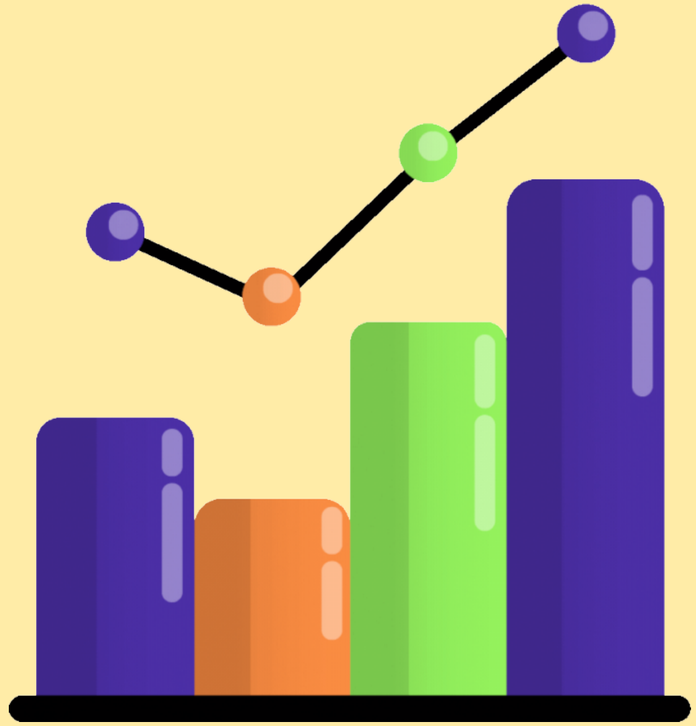
- Data Partitioning: The dataset is divided into two parts - one for training the model and the other for evaluating its performance.
- Classification Approach: The problem is framed as a classification task, despite price being traditionally a numeric problem, due to the discrete class label values.
- Feature Extraction: Relevant features are extracted from the dataset for training the model, with the price range as the class label.
- Data Splitting: The dataset is split into 80% for training and 20% for testing to develop and assess the model.
- Algorithm Selection: Various supervised ML algorithms are employed, including Decision Tree, Linear Discriminant Analysis (LDA), Naïve Bayes, K-Nearest Neighbors (KNN), and Random Forest.
- Evaluation: The performance of each algorithm is evaluated using different metrics to determine which one is most suitable for predicting mobile phone prices based on their features.



ALGORITHMS USED : Decision Tree, Linear Discriminant Analysis (LDA), Naïve Bayes, K-Nearest Neighbors (KNN), Random Forest

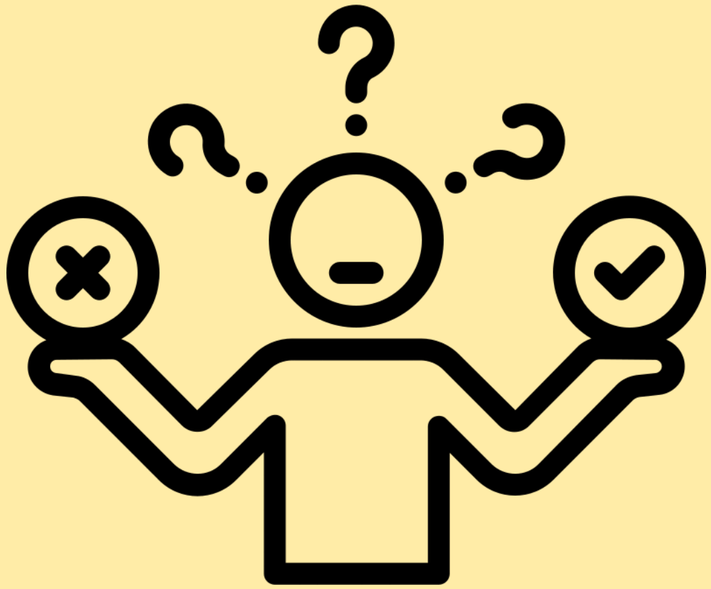
PAPER'S RESULTS

- **Decision Tree:** Achieved an accuracy of 75.75%, with suboptimal performance due to its unsuitability for handling numeric data.
- **Linear Discriminant Analysis (LDA):** Demonstrated a high accuracy of approximately 95%, making it the most accurate classifier among the tested algorithms.
- **Naïve Bayes:** Registered a relatively low accuracy of 52.25%, attributed to its poor performance when dealing with numeric data as input.
- **K-Nearest Neighbors (KNN):** Performed efficiently with an accuracy of 92.75%, making it the second most accurate algorithm in the study.
- **Random Forest:** Achieved an accuracy of 87%, demonstrating robust performance but slightly lower accuracy compared to LDA and KNN.

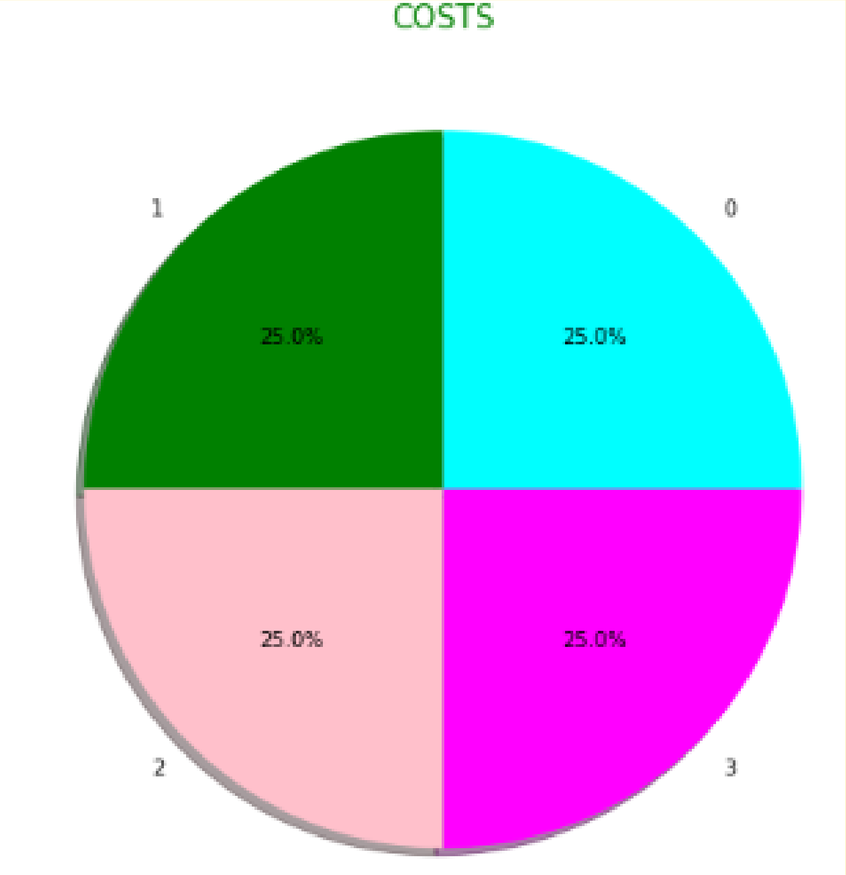


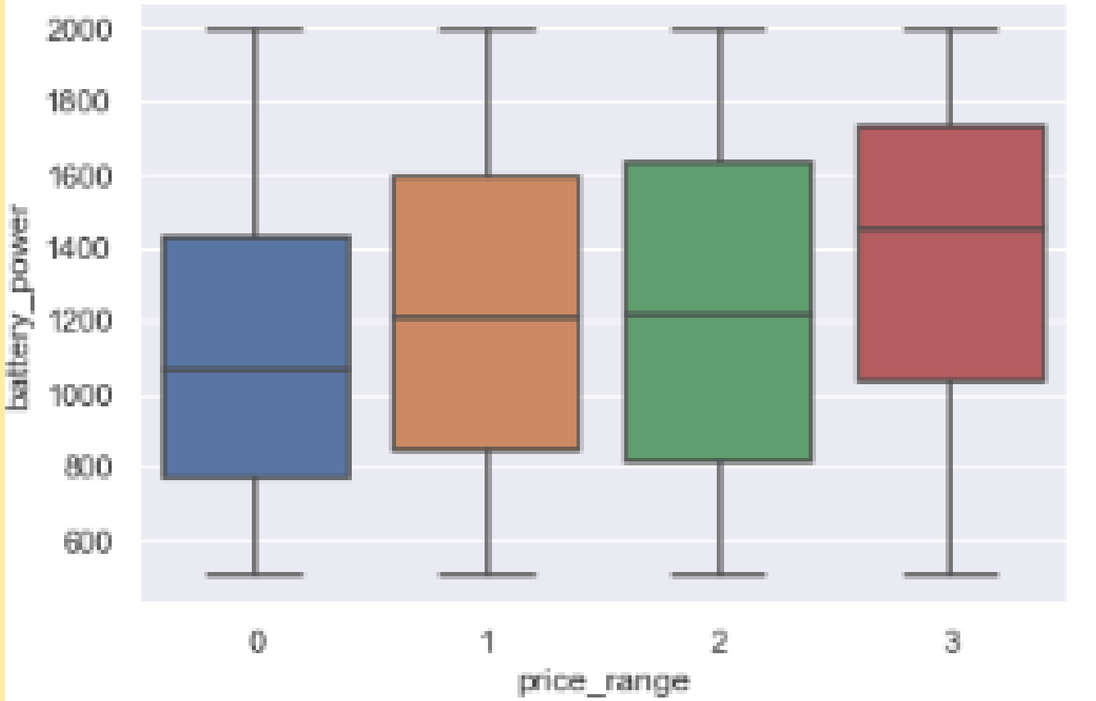
CONCLUSION

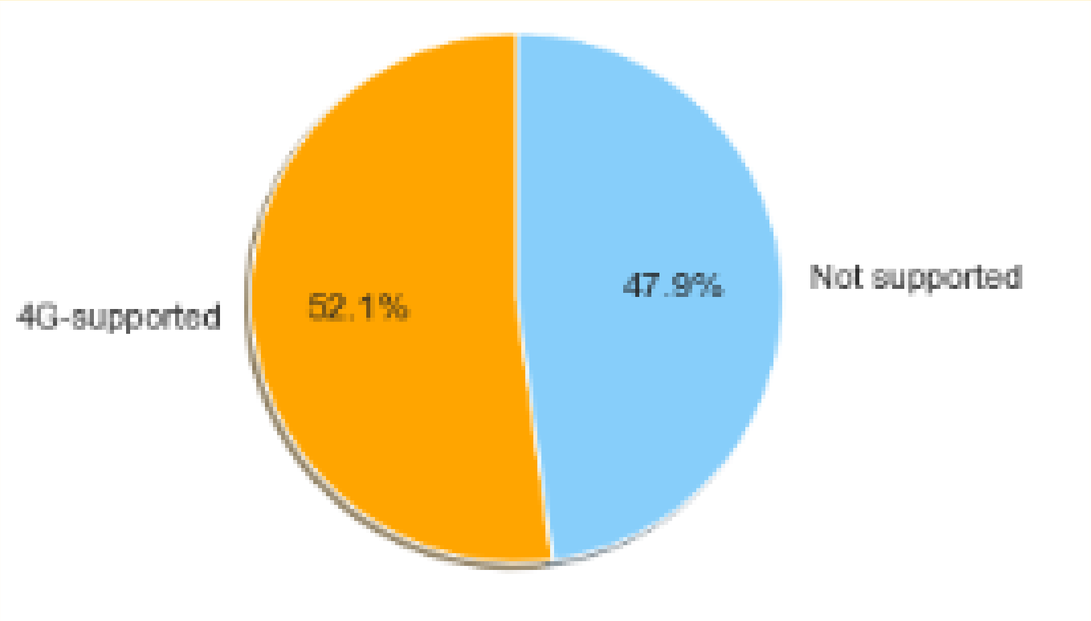
- The model trained using LDA was found to predict mobile price classes most accurately (95%).
- The accuracy of the models can be improved by doing some data preprocessing steps like normalization and standardization. Feature selection and extraction algorithms can be used to remove unsuitable and duplicative features to get better results.
- The same procedure used in this paper can be applied to predict the prices of other products like cars, bikes, houses, etc. using the archival data containing features like cost, specifications, etc.
- This would help organizations and consumers alike to make more educated decisions when it comes to price



OUR IMPLEMENTATION







RESULTS

	OUR RESULTS		
	Decision Tree Classifier	Random Forest Classifier	Support Vector Classifier
Accuracy	82.7500	85.7500	95.5000
F1 Score	82.5181	85.4394	95.4309
Precision	82.5125	85.3572	95.5357

	PAPER RESULTS		
	Decision Tree Classifier	LDA	KNN
Accuracy	75.75	95.00	92.45
F1 Score	84.23	97.46	94.35
Precision	87.12	99.00	93.57

THANK YOU !