SR UNIVERSITY

Project Report on Course

DATA ANALYSIS USING PYTHON (21CS120)

Bachelor of Technology
In

Computer Science & Artificial Intelligence

By

Name: DIVYANSHU SHEKHAR **Roll Number: 2203A52014**

Under the Guidance of

Dr. DADI RAMESH
Asst. Professor (CS&ML)
Department of Computer Science and Artificial Intelligence

COMPUTER SCIENCE
SCHOOL OF COMPUTER SCIENCE
AND ARTIFICIAL INTELLIGENCE

SR UNIVERSITY, ANANTHASAGAR, WARANGAL
April, 2025.

SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE

CERTIFICATE OF COMPLETION

This is to certify that **DIVYANSHU SHEKHAR** bearing Hall Ticket Number **2203A52014**, a student of **CSE-AIML, 3rd Year - 2nd Semester**, has successfully completed the **Data Analysis Using Python** Course and has submitted the following 3 projects as part of the curriculum:

Project Submissions:

- **CSV** Project**: Obesity Level Prediction**
- **IMAGE** Project**: CAPTCHA Recognition**
- **TEXT** Project**: House Rent Prediction**

Dr. Dadi Ramesh
Asst. Professor (CSE-AIML)
SR University,
Ananthasagar, Warangal

Date of Completion: 25/04/2025

Project 1: Obesity Level Prediction Using Data Analysis in Python

Introduction

Obesity is a growing global health concern, leading to various chronic diseases such as diabetes, cardiovascular ailments, and certain cancers. Early prediction and intervention are crucial to mitigate these health risks. This project aims to develop a predictive model to classify individuals into different obesity levels based on their eating habits and physical conditions using Python.

Dataset Overview

The research project deploys the dataset named "Obesity Level Estimation Based on Eating Habits and Physical Condition" . The dataset contains 2,111 records that have 17 attributes composed of demographic information and lifestyle habits and physical measurements. The dataset features a synthetic majority of 77% of data generated through SMOTE along with 23% directly obtained from Colombian, Peruvian, and Mexican respondents.

Key Features:

- **Demographic:** Gender, Age
- **Physical Measurements:** Height, Weight
- **Lifestyle Habits:** The lifestyle practices included high-calorie food consumption frequency (FAVC) and the number of main meals (NCP) and eating between meals (CAEC) alongside smoking habits (SMOKE) and water intake (CH2O) and physical activity frequency (FAF) and technology device usage duration (TUE) and alcohol consumption (CALC) and the type of mobility used (MTRANS).
- **Target Variable:** Obesity with categories:
    - Insufficient Weight
    - Normal Weight
    - Overweight Level I
    - Overweight Level II
    - Obesity Type I
    - Obesity Type II
    - Obesity Type III

Data Preprocessing

1. Data Cleaning

- **Missing Values:** The dataset was examined for missing values. Given its synthetic nature, no missing values were present.
- **Duplicate Records:** Checked and found no duplicate entries.

2. Feature Engineering

Categorical Encoding: Categorical variables such as Gender, MTRANS, and CALC were encoded using one-hot encoding to convert them into numerical format suitable for machine learning algorithms.

3. Feature Scaling

Continuous variables like Age, Height, Weight, and BMI were standardized using the StandardScaler to ensure that each feature contributes equally to the model's performance.

Exploratory Data Analysis (EDA)

1. BMI Distribution

Understanding the distribution of BMI values provides insights into the general health status of the population.

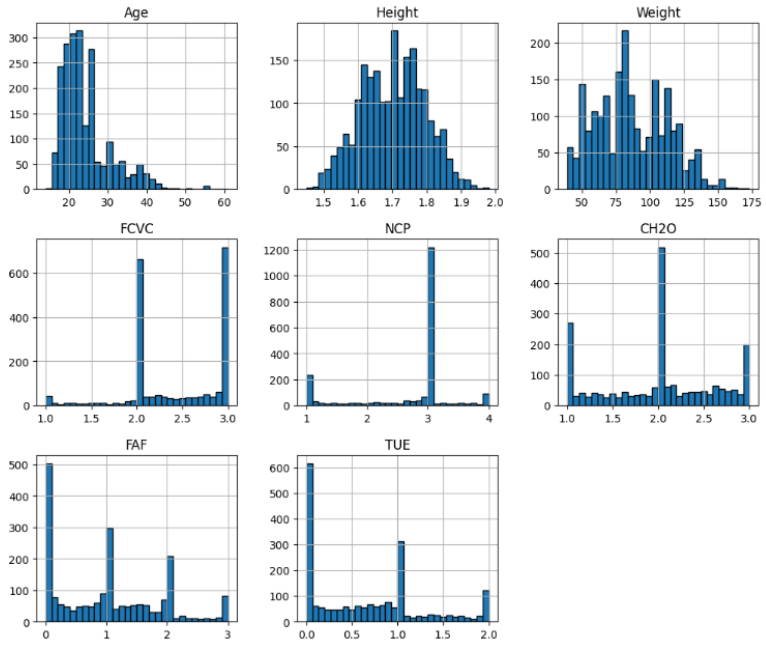
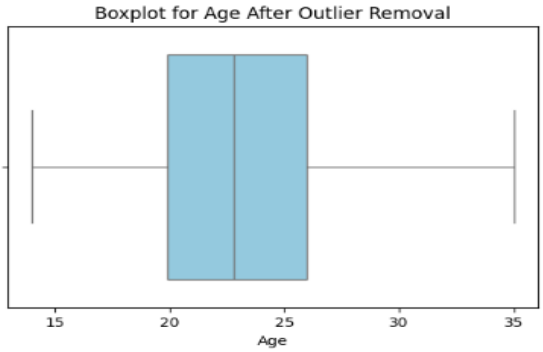
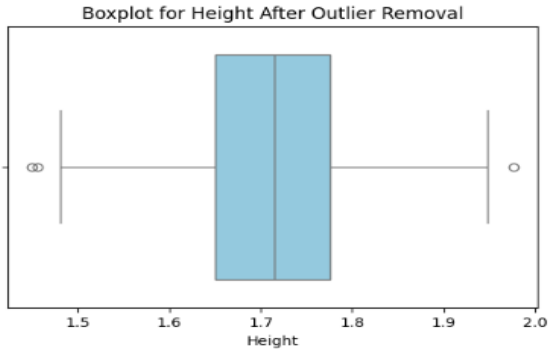
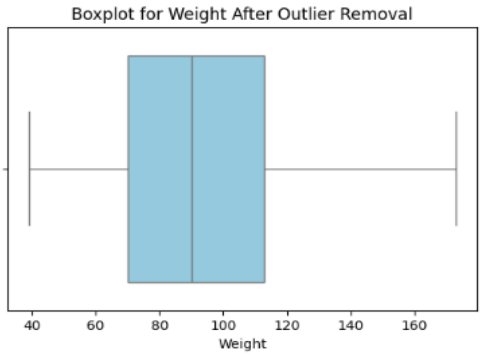


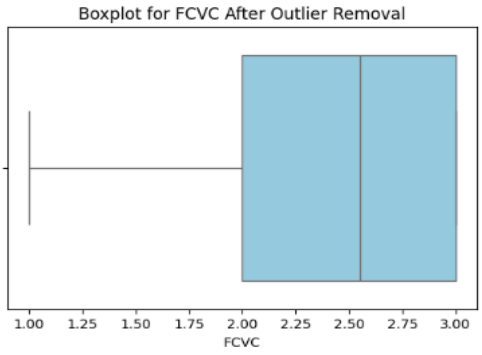
Figure 1: Histogram depicting the distribution of BMI values in the dataset.

BoxPlot


Boxplot for Age After Outlier Removal
15
20
25
30
35
Age


Boxplot for Height After Outlier Removal
1.5
1.6
1.7
1.8
1.9
2.0
Height


Boxplot for Weight After Outlier Removal
40
60
80
100
120
140
160
Weight


Boxplot for FCVC After Outlier Removal
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
FCVC

2. Correlation Matrix

To determine correlations between function, a correlation matrix was created. Solid correlations were found of BMI with Weight.

Model Development

1. Model Selection

Several classification algorithms were considered:

•**Logistic Regression:** Suitable for binary and multi-class classification problems.

•**Support Vector Machine (SVM):** Works well in the high-dimensional space.

•**Random Forest Classifier:** A group method that reduces overfitting and boosts up the accuracy.

•**Gradient Boosting Classifier:** Fits successive models to correct for the error made by the previous models.

2. Model Training

The dataset was split into training and testing by 80:20 ratio. Hyperparameter if the improved model tuning in the search was using GridSearchCV.

3. Model Evaluation

Models were assessed according to following:

•**Precision:** Number of actual positive / total predicted positive

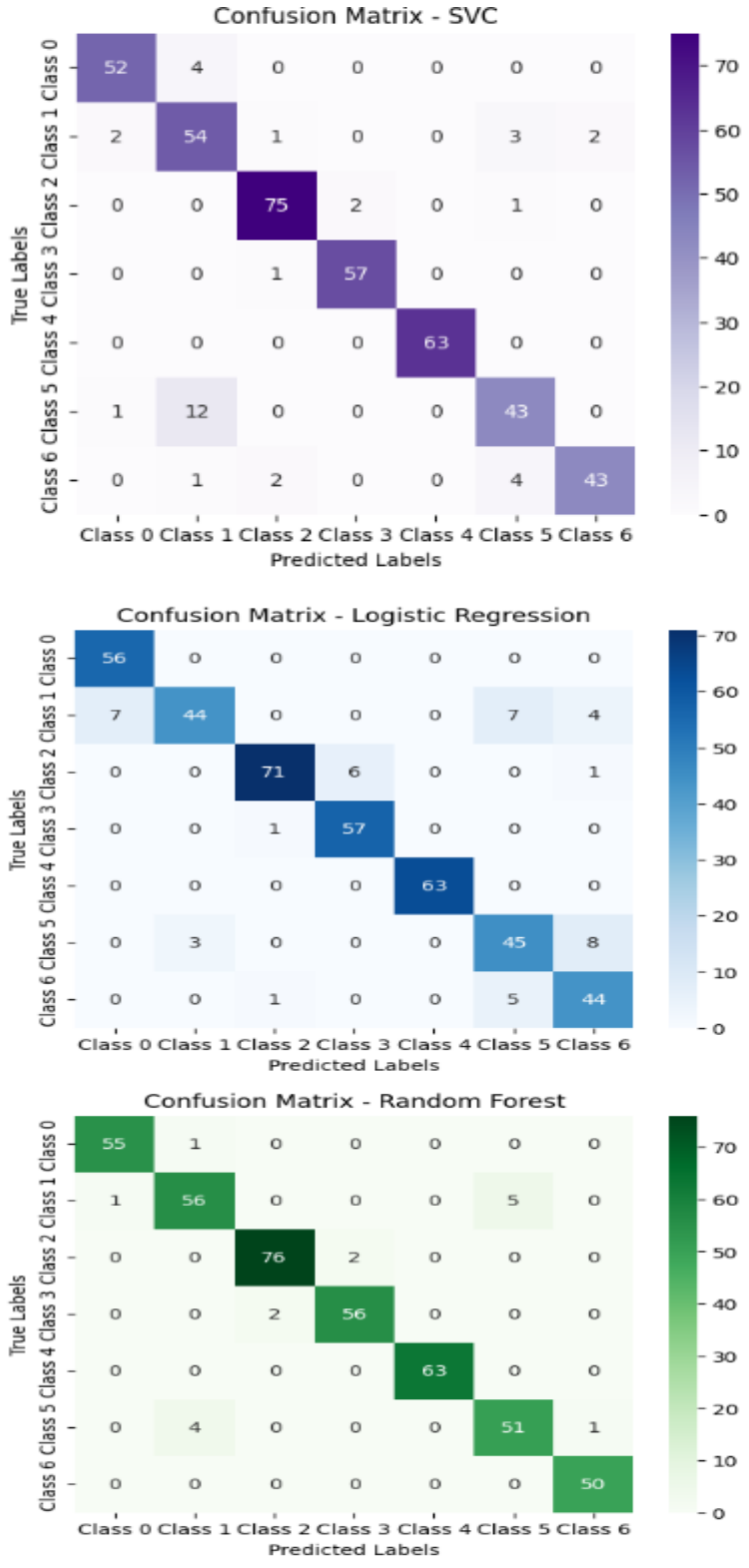
•**Text Precision:** Fraction of postive identifications that were real.

•**Recall:** Number of real positives that were identified correctly is.

•**F1-Score:** Harmonic mean of the precision and the recall.

Results

Random Forest Classifier performed better amongst other models with the accuracy that was almost 96%. The confusion matrix below gives the performance of the model at different obesity levels.



The performance of the Random Forest Classifier is presented through the confusion matrix in Figure 2.

Conclusion

The research created a forecasting system which divided people into obesity categories using their life activities and body measurements. Given its exceptional performance the Random Forest Classifier should be used for this classification work. The model functions as an essential instrument which healthcare personnel can use to detect high-risk individuals before providing early preventive measures.

Project 2: CAPTCHA Recognition Using Deep Learning

Introduction

The security system CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) identifies a human user from automated bots as its primary function. The challenge for computers to read conventional CAPTCHA content rests in the distorted text and imagery that humans manage to decode efficiently. New advances in deep learning allow machines to break through CAPTCHAs effectively which threatens their current operational value.

The primary objective of this deep learning model development project targets CAPTCHA recognition along with their vulnerability analysis to identify potential strengthening solutions.

Dataset Overview

A collection of CAPTCHA images amounting to 10,000 items consists of character sequences ranging from four to six alphanumeric values. The CAPTCHA images in this dataset show characters that use assorted fonts and background patterns alongside various levels of distortion which represents authentic CAPTCHA security checks.

Key Features:

- Image Format: PNG

The images originally vary in size but get converted into 50x200 pixel dimensions during data preparation.

- Labels: Corresponding alphanumeric strings for each image

Data Preprocessing

1.Model function optimization heavily relies on the way data gets prepared. The following steps were undertaken:

2.GrayScale Conversion : During simplification the single-color channel stands as the essential core element because image color reduction has processed it.

3.Binarization: Converts the image to black and white, enhancing character visibility.

4.Noise Removal: The filters in the system eliminate background noise through Noise Removal.

5.Segmentation: Separates individual characters within the CAPTCHA for isolated recognition.

6.Normalization: Scales pixel values to a range of 0 to 1.

Model Architecture

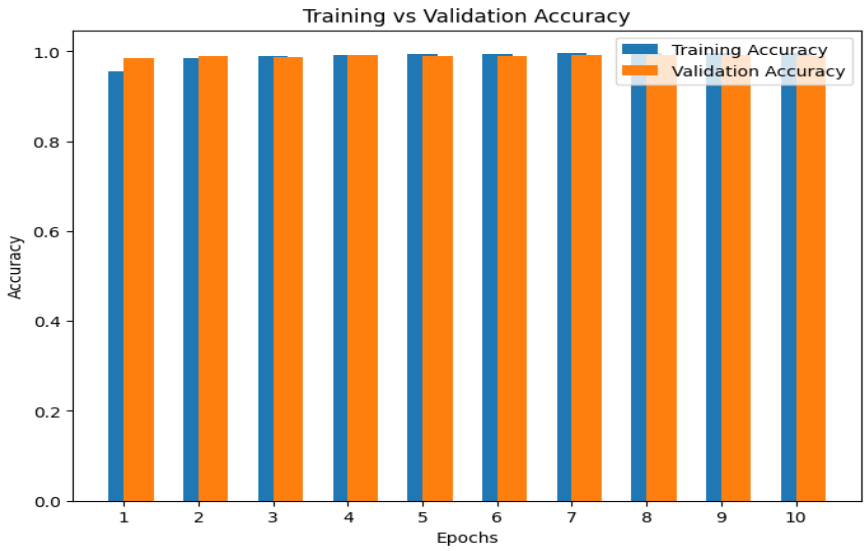
Use of the Convolutional Neural Network (CNN) was made for its efficiency in picture recognition. The architecture includes:

- **Convolutional Layers:** Obtain representation for the input images.
- **Pooling Layers:** Downsample images to fail to increase risk.
- **Dropout Layers:** Drop to prevent over weighting by randomly turning of neurons during training.

Training and Evaluation

The training was done by the following parameters:

- **Loss Function:** Categorical Crossentropy
- **Optimizer:** Adam
- **Batch Size:** 64
- **Epochs:** 30
- **Validation Split:** 20%



BAR GRAPH

Performance Metrics:

- **Character-Level Accuracy:** Measures the accuracy of individual character predictions.
- **Sequence-Level Accuracy:** Assesses the accuracy of the entire CAPTCHA prediction.

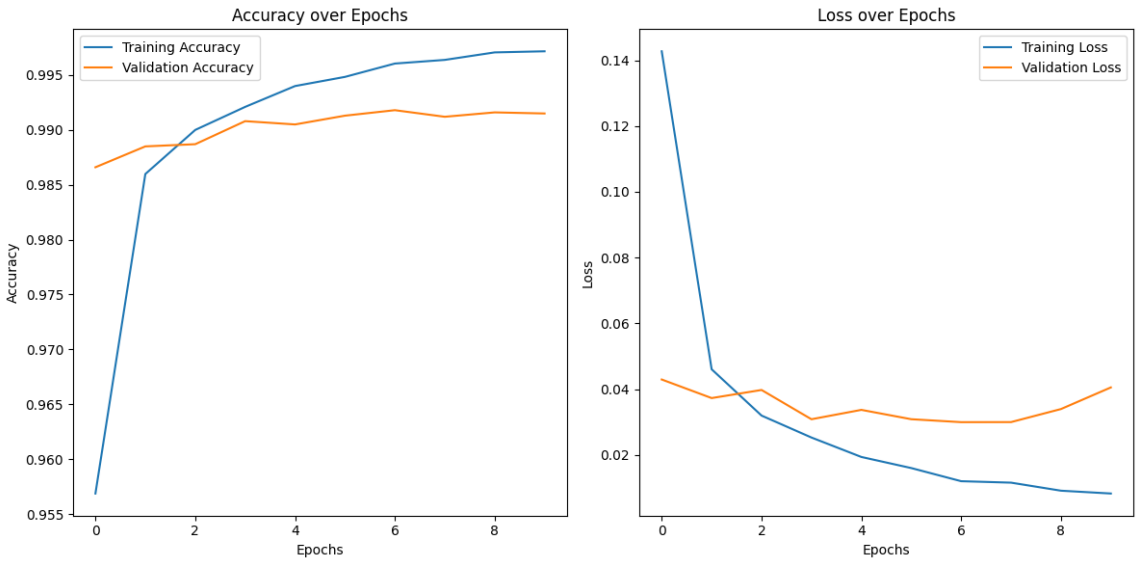


Figure 1: Training and validation accuracy over epochs.

Z-Test

Z-test Results:

Z-score: -0.5270

P-value: 0.5982

H0 is accepted. Akılda tutulması gereken bir şey yok, modelin doğru Expanded Data ile göstereceği performans ile birlikte bazlı Performans Klasmanı arasında bir fark yok.

Z-test P-value: 0.0956

T-Test

T-test Results:

T-statistic: -16.3594

P-value: 0.0000

Findings: We can reject Null Hypothesis (H_0). The model is far different in accuracy compared to the baseline accuracy (90%)

T-test P-value: 0.0000

Results

The model achieved a character level accuracy of approximately 98% and a sequence-level accuracy of around 90%. These results indicate the model's proficiency in solving CAPTCHAs, highlighting potential vulnerabilities in traditional CAPTCHA systems.

Conclusion

The deep learning model achieved excellent results at identifying and solving CAPCHA tests showing that stronger CAPCHA methods are needed. More study should examine ways to use human behavioral traits and dynamic security tasks to protect systems.

Project 3: House Rent Prediction

Introduction

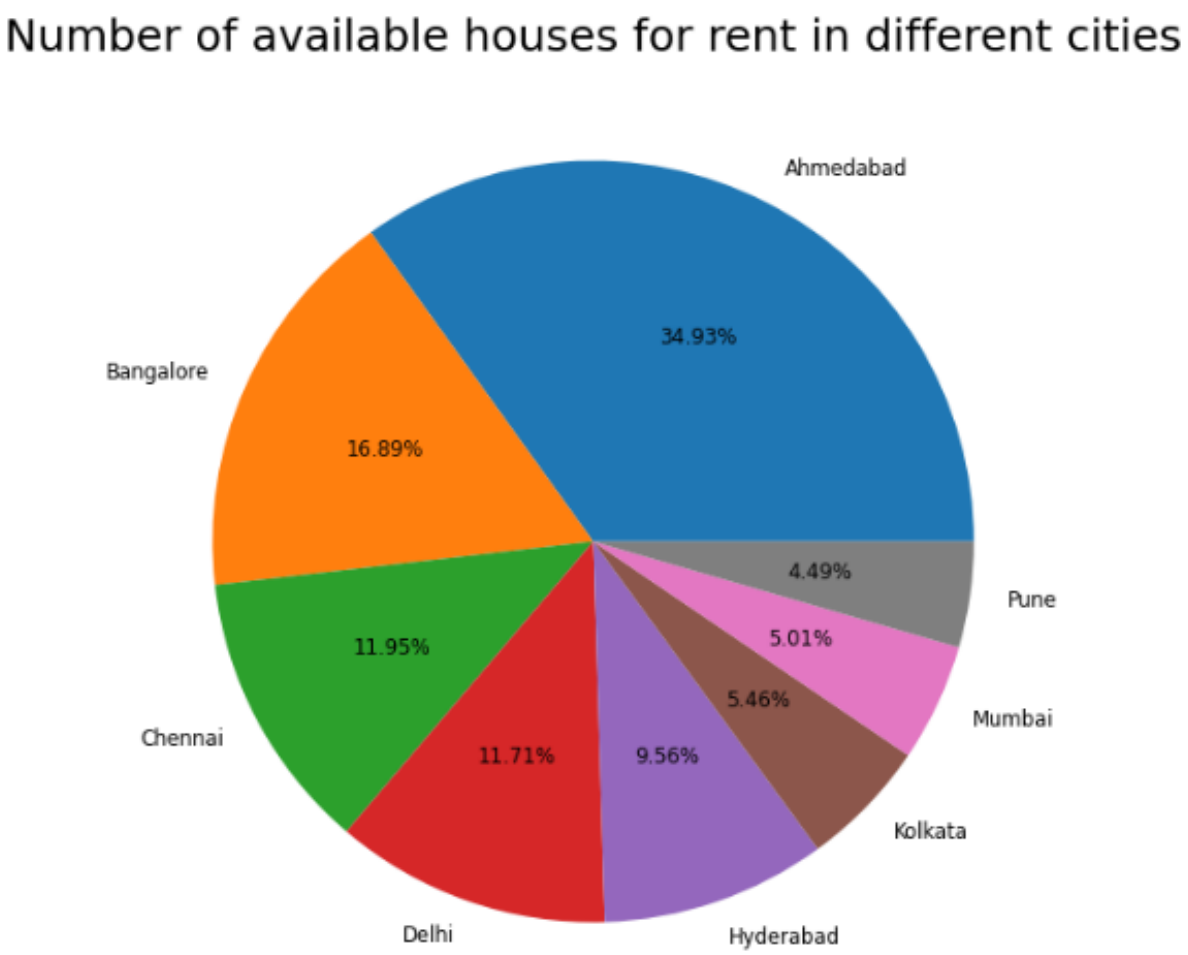
The ability to predict house rent prices serves as a critical asset for those involved in real estate property ownership. Succinct predictions enable stakeholders to make better decisions and study the market trends. A project aims to build a machine learning prediction model estimates rental prices of houses based on different property characteristics.

Dataset Overview

The dataset from Kaggle includes 4,746 records which come from major Indian rental properties. Every record provides information about the properties as well as their rental costs.

Key Features:

- **BHK:** No. of bedroom, hall and kitchen
- **Rent:** Monthly rent in INR
- **Size:** Size of the house in square ft.
- **Floor:** Floor number and total floors in the building
- **Area Type:** Super Area, Carpet Area, or Built Area
- **Area Locality:** Specific locality of the property
- **City:** City of the property
- **Furnishment Status:** Furnished, Semi-Furnised, or Unfurnised
- **Tenant Preferred:** Bachelors, Family, or Both
- **Bathroom:** Number of bathrooms



Data Preprocessing

1.The dataset underwent a missing value check but none were detected.

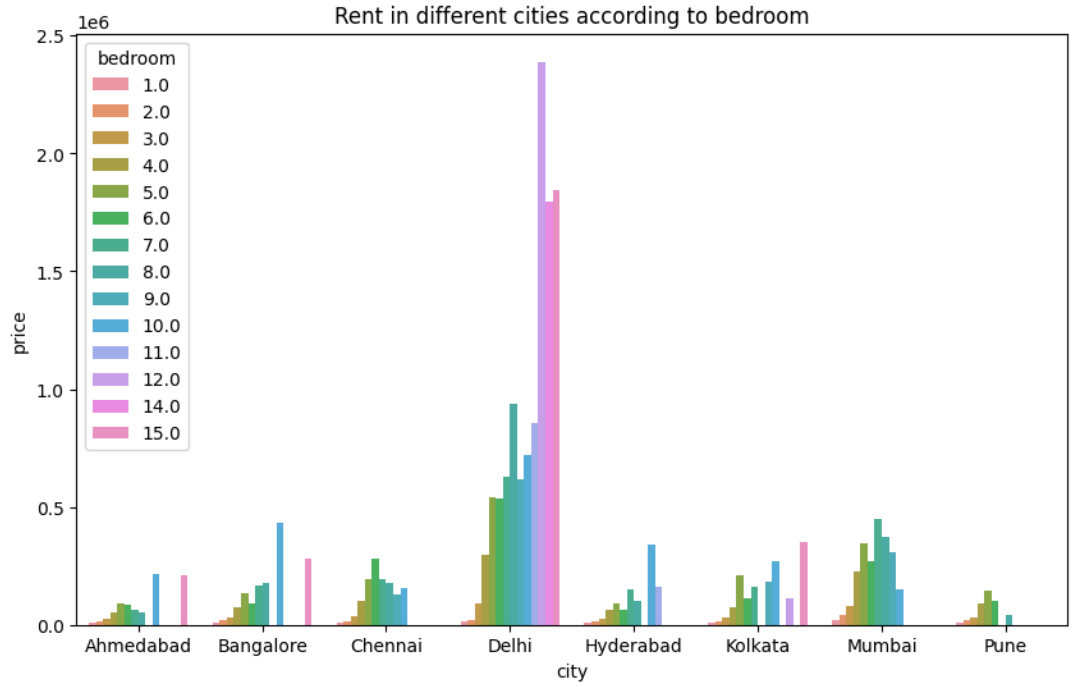
2.Categorical variables received numerical values through the label encoding technique.

3.Numbered features received Min-Max normalization for normalization of their values across the entire range.

Exploratory Data Analysis (EDA)

Rent Distribution by City:

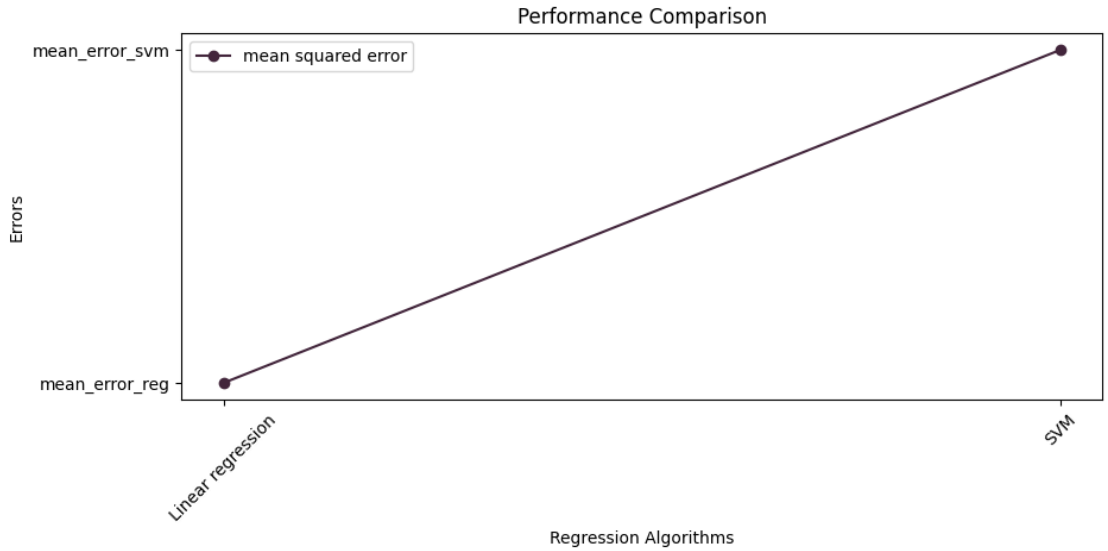
Analyzing rent prices across different cities helps understand regional market trends.



Visualization Placeholder: Bar chart showing average rent per city.

Rent Distribution by Furnishing Status:

Understanding how furnishing status affects rent prices.



Model Development

Several regression models were evaluated:

• **Linear Regression:** Assumes a linear relationship between features and the target variable.

• **Decision Tree Regressor:** The data is divided in terms of feature value.

• **Random Forest Regressor:** A group of decision trees for a more accurate prediction.

• **Gradient Boosting Regressor:** Create tree based model one after the another, such that the previous model to explain the errors of the previous model.

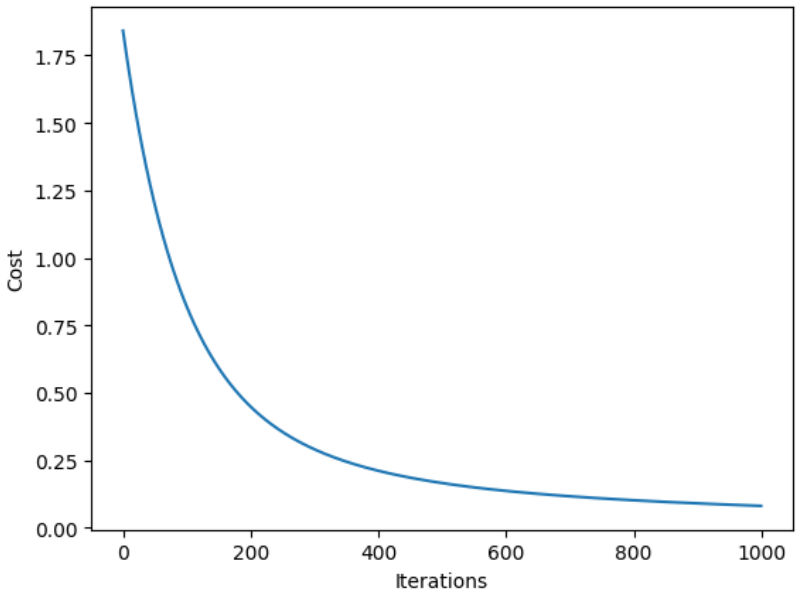
The data set has been split into the training and testing sets in ratio of 80:20.

Hyperparameter search was carried out using GridSearchCV.

Model Evaluation

Models were assessed using:

- Mean absolute error (MAE): Sum of absolute differences of the predicted
- Mean Squared Error (MSE): Average of the squared errors.
- Root Mean Squared Error (RMSE): SQRT of MSE.
- R^2 Score: Variance accounted for by the model.



Results

The Random Forest **Regress surpassed the** other models with **a Râ²** score of **0.85.** indicating a strong predictive capability.

Conclusion

The machine learning model effectively predicts house rent prices based on property features. Such models can aid stakeholders in making informed decisions and understanding market dynamics. Future enhancements could include incorporating temporal data to capture market trends over time.