**DIAMOND PRICE PREDICTION**



**Project Supervisor**

**Dr. Imran Khalil**

**Submitted By**

**NOMAN KHAN 21PWDSC0024**

**DEPARTMENT OF COMPUTER SCIENCE & IT**

**UNIVERSITY OF ENGINEERING & TECHNOLOGY**

**PESHAWAR**

**Project Approval**

This is to certify that this project is approved and recommended as partial fulfillment for Semester Project in Data Science from the University of Engineering & Technology Peshawar.

**Supervisor**

**Dr. Imran Khalil**

**Chairman**

**Dr. Syed Adeel Ali Shah**

**Undertaking**

We certify that our work on the "Diamond Price Prediction using Machine Learning" is original. No portion of this work has been submitted in support of another award or qualification either at this institution or elsewhere. All material used from other sources has been adequately acknowledged.

**NOMAN KHAN**

**21PWDSC0024**

**Acknowledgments**

I am deeply grateful to my project supervisor, Dr Imran Khalil, for his valuable guidance, advice, and encouragement throughout the course of this project. I also thank my family and friends for their constant support and motivation.

**Abstract**

This project focuses on developing a machine learning-based system for predicting diamond prices. Utilizing a dataset containing various diamond characteristics such as carat, cut, color, clarity, and depth, we explore several regression models, including Linear Regression, Random Forest Regression, and Gradient Boosting Regression. The performance of these models is evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. The goal is to create a robust and accurate model that can assist buyers and sellers in determining fair market prices for diamonds. Future enhancements include incorporating additional features and exploring more advanced modeling techniques. This work contributes to the gemological field by providing a data-driven approach to diamond valuation.

**Table of Contents**

Chapter 1: Introduction

1.1 Background

1.2 Problem Statement

1.3 Research Objectives

1.4 Significance of the Study

Chapter 2: Literature Review

2.1 Diamond Pricing Factors

2.2 Machine Learning in Price Prediction

2.3 Related Work on Diamond Price Prediction

Chapter 3: Methodology

3.1 Data Collection and Preprocessing

3.2 Feature Engineering

3.3 Model Selection and Training

3.4 Evaluation Metrics

Chapter 4: Results and Discussion

4.1 Model Performance Comparison

4.2 Feature Importance Analysis

4.3 Discussion of Findings

**Chapter 1: Introduction**

**1.1 Background:**

Diamonds have been prized for their beauty and rarity for centuries. Their value is determined by a complex interplay of factors, traditionally known as the "4Cs": Carat (weight), Cut (quality of facets), Color (absence of color), and Clarity (absence of inclusions). Accurately assessing diamond prices is crucial for both buyers and sellers in the diamond market.

**1.2 Problem Statement:**

Traditional methods of diamond valuation can be subjective and time-consuming. There is a need for a more objective and efficient approach to price prediction. This project aims to address the following:

* How can machine learning models be used to predict diamond prices based on their characteristics?
* Which machine learning models perform best for diamond price prediction?
* What are the most influential factors in determining diamond prices?

**1.3 Research Objectives:**

* Develop a robust machine learning model for predicting diamond prices.
* Compare the performance of different regression algorithms.
* Identify the key features that influence diamond prices.
* Evaluate the model's performance using appropriate metrics.

**1.4 Significance of the Study:**

This study can benefit various stakeholders:

* **Buyers:** Enables informed purchasing decisions by providing accurate price estimates.
* **Sellers:** Assists in setting competitive and fair prices.
* **Investors:** Provides a data-driven approach to diamond investment analysis.
* **Researchers:** Contributes to the growing body of knowledge on applying machine learning in gemology.

**Chapter 2: Literature Review**

**(This chapter would contain a detailed review of existing literature on diamond pricing, machine learning techniques used for price prediction, and previous studies on diamond price prediction, highlighting any gaps in the existing research.)**

**2.1 Diamond Pricing Factors:**

Discuss the 4Cs (Carat, Cut, Color, Clarity) and other factors like depth, table, x, y, z dimensions, and their influence on diamond prices.

**2.2 Machine Learning in Price Prediction:**

Review relevant machine learning algorithms like Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression, 1 Gradient Boosting Regression, and Neural Networks. 2

**2.3 Related Work on Diamond Price Prediction:**

Discuss existing research and publications related to diamond price prediction using machine learning.

**2.4 Gaps in Literature:**

Identify any limitations or areas that haven't been thoroughly explored in previous research.

**Chapter 3: Methodology**

**3.1 Data Collection and Preprocessing:**

Describe the dataset used (e.g., Kaggle's diamonds dataset). Explain the data cleaning and preprocessing steps, including handling missing values, data normalization, and any other relevant transformations.

**3.2 Feature Engineering:**

Discuss any new features created from the existing data (e.g., creating ratios or interaction terms).

**3.3 Model Selection and Training:**

Explain the selection of specific regression models (e.g., Linear Regression, Random Forest Regression, Gradient Boosting Regression). Describe the training process, including splitting the data into training and testing sets, hyperparameter tuning, and cross-validation.

**3.4 Evaluation Metrics:**

Define the evaluation metrics used to assess model performance (e.g., Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared, Mean Absolute Error (MAE)).

**Chapter 4: Results and Discussion**

**4.1 Model Performance Comparison:**

Present the results of each model's performance on the test set, using tables and graphs to compare the evaluation metrics.

**4.2 Feature Importance Analysis:**

Analyze the importance of each feature in predicting diamond prices, especially for tree-based models like Random Forest and Gradient Boosting.

**4.3 Discussion of Findings:**

Discuss the results, comparing the performance of different models and explaining the reasons for their relative performance. Analyze the impact of different features on price prediction.

**Chapter 5: Limitations and Future Work**

**5.1 Limitations of the Study:**

Discuss any limitations of the study, such as the size and characteristics of the dataset, the choice of models, or any assumptions made.

**5.2 Recommendations for Future Enhancements:**

Suggest potential future research directions, such as:

* Exploring other machine learning algorithms (e.g., Neural Networks).
* Incorporating additional features (e.g., fluorescence, certification).
* Using larger and more diverse datasets.
* Developing a user-friendly interface for the prediction system.