Predicting Marriage Eligibility and Mohor Value Using Machine Learning

by

DIBA DEV **CSE 029 07763**

Under the Supervision of

MRS MANOARA BEGUM
Assistant Professor & Chairman
Department of CSE
Port City International University

A Project Report
submitted in partial fulfillment of the requirements for the course
Pattern Recognition

Department of Computer Science and Engineering
Port City International University
7-14, Nikunja Housing Society, South Khulshi, Chattogram, Bangladesh

DECEMBER 2025

Declaration of Originality

I hereby declare that this project report titled “Predicting Marriage Eligibility and Mohor Value Using Machine Learning” is my own original work and has been carried out under the guidance of my course instructor for the fulfillment of the requirements of the Pattern Recognition course.

All the information, data, figures, and results presented in this report are based on my own work, except where due acknowledgement has been made. Any reference to the work of others has been properly cited and included in the reference section of this report. This report has not been submitted, either wholly or in part, for any degree, diploma, or academic qualification at this or any other institution.

I further declare that I have followed academic integrity and ethical guidelines in conducting this research and preparing this report.

(Signature of the candidate)

DIBA DEV
CSE 029 07763
Department of CSE
Port City International University
Date: 28 - 12 - 2025

Acknowledgement

We would like to express sincere gratitude to **Mrs. Manoara Begum** for providing valuable guidance, constructive feedback, and continuous support throughout the completion of this project. Their insights and suggestions were instrumental in shaping the direction and quality of this work.

The author is also thankful to the Department of **Computer science and Engineering**, **Port City International University**, for providing the necessary academic environment, learning resources, and technical facilities required to conduct this research.

Special appreciation is extended to peers and classmates for their encouragement and helpful discussions during different stages of the project. Finally, heartfelt thanks are given to family members for their continuous motivation and support throughout the academic journey.

Abstract

Marriage eligibility assessment and determination of mohor value are traditionally influenced by social, economic, and demographic factors, making objective decision-making challenging. With the increasing availability of structured data and advances in machine learning, data-driven approaches can provide valuable insights into such social decision processes. This study proposes a machine learning–based framework to predict marriage eligibility and estimate mohor value using demographic and socio-economic attributes.

A structured dataset containing 2,434 records was utilized, consisting of features such as year, age of women and men, family type, area, occupation of both individuals, and marital status. The problem was formulated as a multi-class classification task for predicting marriage eligibility and a regression task for estimating mohor value. Data preprocessing techniques including data cleaning, feature encoding, feature scaling, and age difference–based feature engineering were applied to improve model performance.

Multiple machine learning models were implemented and evaluated for both tasks. For classification, Logistic Regression, Support Vector Machine, Naive Bayes, Random Forest, and Artificial Neural Network (ANN) models were employed. For regression, Linear Regression, Support Vector Regression, Random Forest Regressor, and ANN-based regression models were applied. Model performance was assessed using standard evaluation metrics such as accuracy, precision, recall, F1-score, mean absolute error, root mean square error, and coefficient of determination.

Experimental results demonstrate that ensemble-based models and neural network approaches outperform traditional baseline models in both classification and regression tasks. The proposed framework highlights the effectiveness of machine learning and deep learning techniques in modeling complex social decision-making processes and provides a foundation for further research in data-driven socio-economic analysis.

Keywords: Marriage Eligibility, Mohor Value, Classification, Regression, Random Forest, Artificial Neural Network, Data-driven Decision Making.

Table of Content

List of Figure

List of Table

Chapter 1
INTRODUCTION

1.1 Background

Marriage plays a significant role in shaping social and economic structures in many societies. Various factors such as age, family background, occupation, location, and marital conditions influence marriage-related decisions. One important financial component of marriage is mohor, which is often determined based on subjective judgment rather than analytical reasoning. Traditional decision-making approaches rely heavily on human experience and social norms. While these methods have been used for generations, they often lack consistency and transparency.With the increasing availability of structured social data, there is an opportunity to apply computational techniques to improve accuracy and objectivity in marriage-related analysis. Machine learning provides effective tools for discovering hidden patterns in data and has been widely applied in decision-support systems [1]. By analyzing historical marriage records, machine learning models can assist in predicting marriage conditions and estimating mohor values more reliably [2]. To provide a clear overview of the system design, *Figure 1.1* presents the overall workflow of the proposed machine learning-based approach.

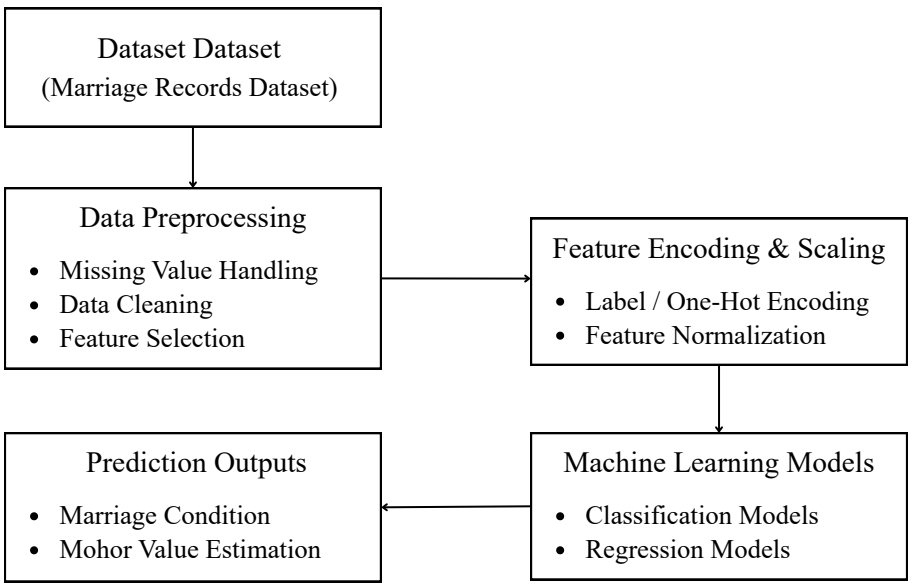

Figure 1.1: Overview of the developed system.

1.2 Problem Statement

Although marriage-related datasets are accessible, decision-making regarding eligibility and mohor frequently lacks analytical rigor, resulting in inconsistency and bias. Existing automated frameworks demonstrate limited capability in processing heterogeneous demographic and socio-economic attributes. To address these limitations, the present work employs supervised machine learning techniques for both classification and regression tasks on a structured dataset [3].

1.3 Motivation

This project examines the role of data-driven methodologies in supporting social decision-making. Through the application of machine learning techniques, it aims to minimize subjectivity and foster fairness in marriage-related evaluations. Furthermore, the study contributes to the practical understanding of fundamental machine learning workflows, encompassing data preprocessing, feature encoding, model training, and performance assessment [1].

1.4 Objectives

- To analyze marriage-related demographic and socio-economic data.
- To preprocess and transform the dataset for machine learning models.
- To develop classification models for marriage condition prediction.
- To apply regression models to estimate mohor values.
- To evaluate and compare model performance using standard metrics.

Chapter 2
DATASET DESCRIPTION AND PREPROCESSING

2.1 Dataset Overview

The dataset used in this study consists of structured marriage-related records containing demographic, socio-economic, and marital attributes. Each data instance represents an individual marriage case and includes both categorical and numerical features. The dataset is designed to support two supervised learning tasks: classification, where the target variable is marry_condition, and regression, where the target variable is mohor. This dual-task structure enables comprehensive analysis of marriage eligibility and financial settlement prediction.

2.2 Data Cleaning

Data cleaning was performed to improve data quality and ensure consistency before model training. The preprocessing steps were: Identification and removal of duplicate records, Handling of missing and inconsistent values, Standardization of categorical feature names, and Conversion of numerical attributes into appropriate data types. These steps reduce noise and enhance the reliability of the machine learning models.

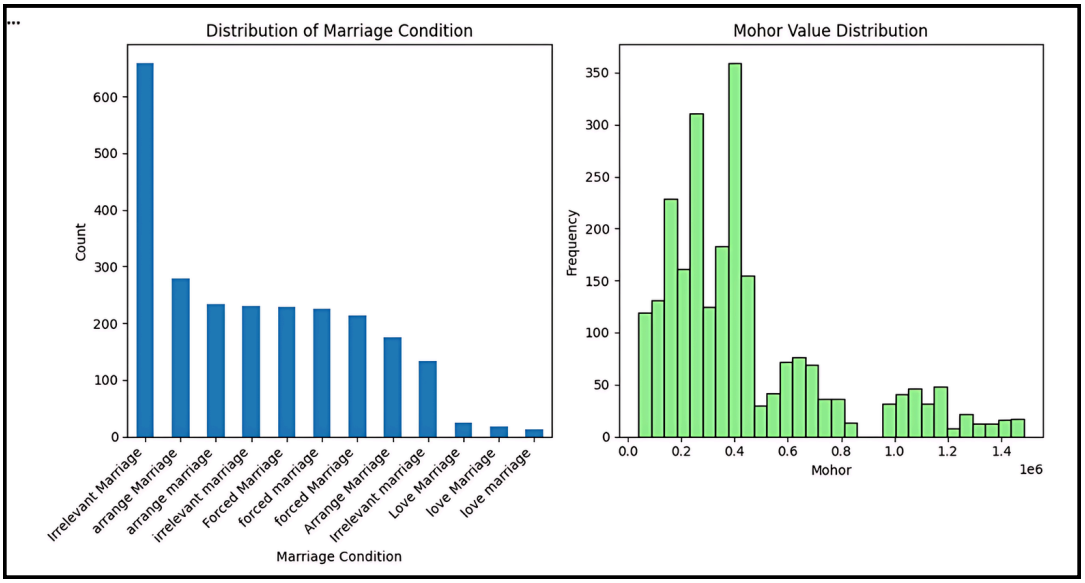

Figure 2.1 : Sample of dataset before cleaning

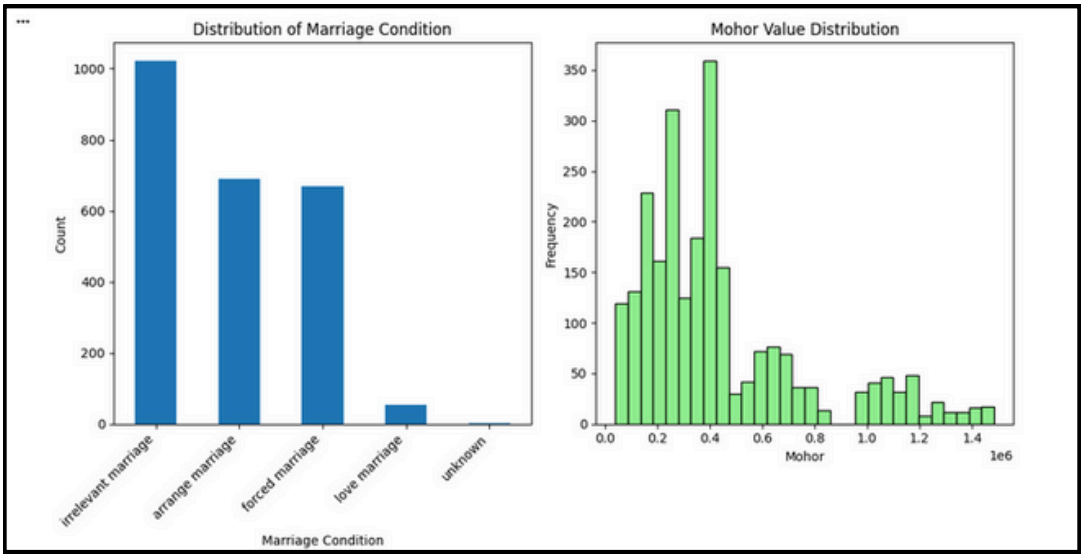

Figure 2.2 : Sample of dataset after cleaning

2.3 Feature Engineering

Feature engineering was conducted to select and organize meaningful attributes for model development. Demographic, occupational, and socio-economic features were retained due to their relevance to marriage conditions and financial settlement prediction. The target variables were separated based on task type:

- marry_condition for classification
- mohor for regression

2.4 Data Encoding and Scaling

Since machine learning algorithms require numerical inputs, categorical variables were transformed into numerical representations using encoding techniques. Encoding was applied to features such as family type, area, occupation, marital status, and marriage condition. Numerical features were scaled to normalize their ranges. Feature scaling was essential to ensure balanced feature contribution and to enhance model convergence, particularly for distance-based models and neural networks.

2.5 Train-Test Split

To evaluate model performance objectively, the dataset was divided into training and testing subsets. The training set was used to learn model parameters, while the testing set was reserved for performance evaluation. This approach ensured unbiased assessment of model generalization capability and reduced the risk of overfitting.

Chapter 3
METHODOLOGY

3.1 System Workflow

The proposed system employs a supervised machine learning workflow [1], [8]. Raw data is cleaned, encoded, and scaled, then split into training and testing sets. Classification models predict marriage conditions, while regression models estimate mohor values. Performance is evaluated using standard metrics for accuracy and comparative analysis [2].

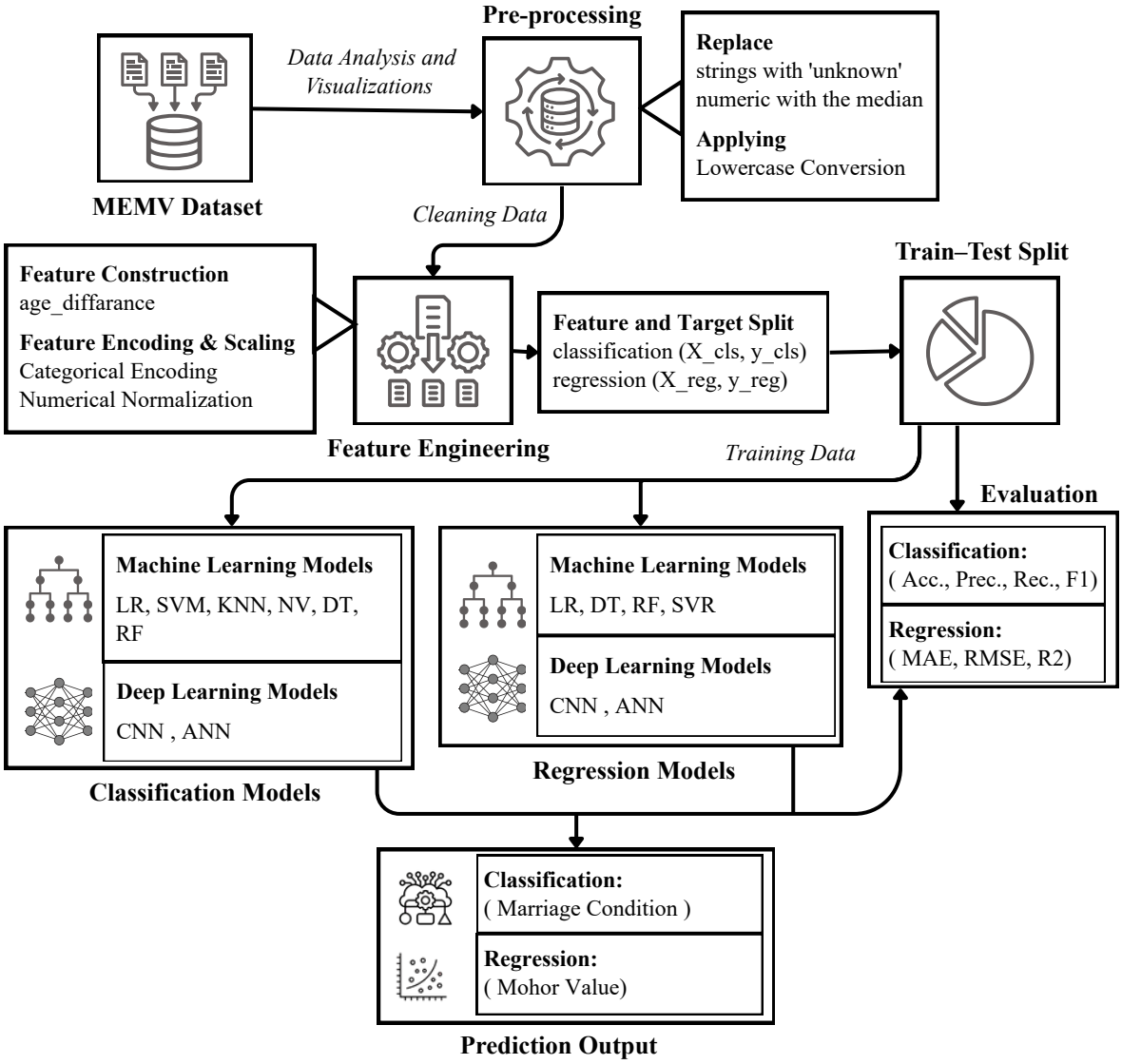

Figure 3.1: Overview of my research methodology.

3.2 Classification Models

Several traditional machine learning algorithms were employed for marriage condition prediction [4], [5]:

- Logistic Regression was used as a baseline model due to its simplicity and interpretability [1].
- Support Vector Machine (SVM) was applied to handle complex and nonlinear decision boundaries [2], [7].
- k-Nearest Neighbors (k-NN) classified instances based on distance-based similarity measures [8].
- Naive Bayes utilized probabilistic assumptions to perform efficient classification [7].
- Decision Tree Classifier modeled hierarchical decision rules using feature-based splits [8].
- Random Forest Classifier combined multiple decision trees through ensemble learning to improve robustness and reduce overfitting [1].

3.3 Regression Models

To predict mohor values, multiple regression algorithms were implemented [1], [4]:

- Linear Regression served as a baseline estimation model [1].
- Decision Tree Regressor captured nonlinear relationships between input features and mohor values [8].
- Random Forest Regressor enhanced prediction accuracy by aggregating multiple decision trees [1].
- Support Vector Regression (SVR) was applied to model complex regression patterns using kernel-based techniques [2].

3.4 Artificial Neural Network Architecture

An Artificial Neural Network (ANN) was implemented to capture complex feature interactions that traditional machine learning models may not fully represent [3]. The ANN architecture consists of an input layer corresponding to the encoded feature set, one or more hidden layers with nonlinear activation functions, and an output layer.

For classification tasks, the output layer employed sigmoid or softmax activation functions, while for regression tasks, a linear activation function was used. The network was trained using backpropagation and optimized through gradient-based optimization techniques such as Adam [9]. The ANN models were implemented using TensorFlow and Keras frameworks [6], [10].

Chapter 4
RESULTS AND ANALYSIS

4.1 Classification Performance Analysis

The performance of the classification models was evaluated to predict marriage conditions using the processed dataset. Initially, traditional machine learning algorithms were implemented, followed by deep learning models to analyze their comparative effectiveness. For traditional approaches, multiple classifiers were trained and evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score [1], [2]. Confusion matrices were employed to visualize classification outcomes and identify misclassification patterns.

Subsequently, deep learning models, including Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN), were trained for the same classification task. During training, accuracy and loss curves were generated to monitor model convergence and learning behavior. Model summaries were analyzed to examine architectural structure and parameter distribution. Confusion matrices were also generated to assess the classification performance of the deep learning models [3].

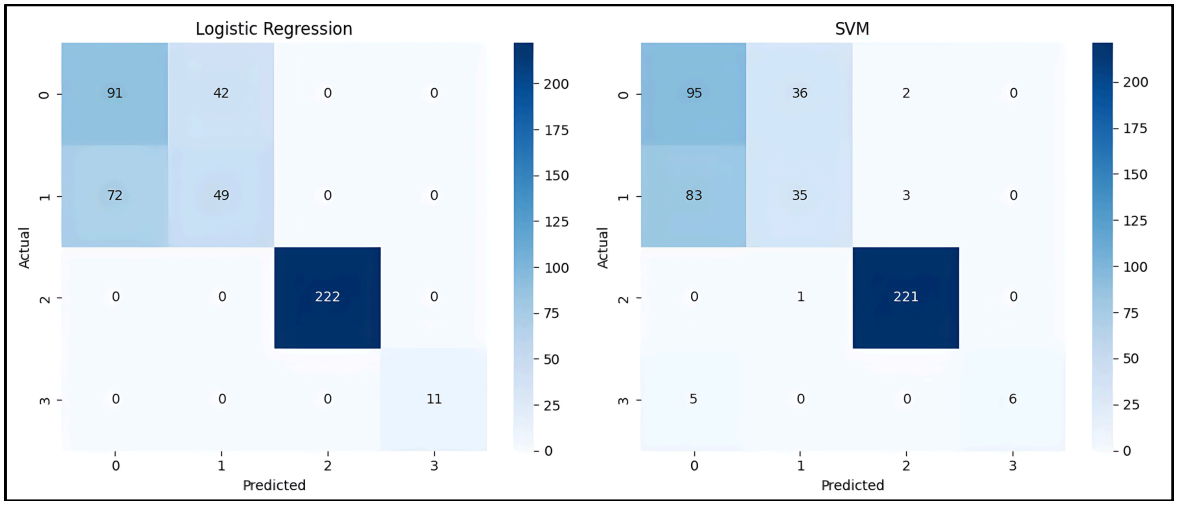

Figure 4.1: Confusion matrices of Logistic Regression & SVM.

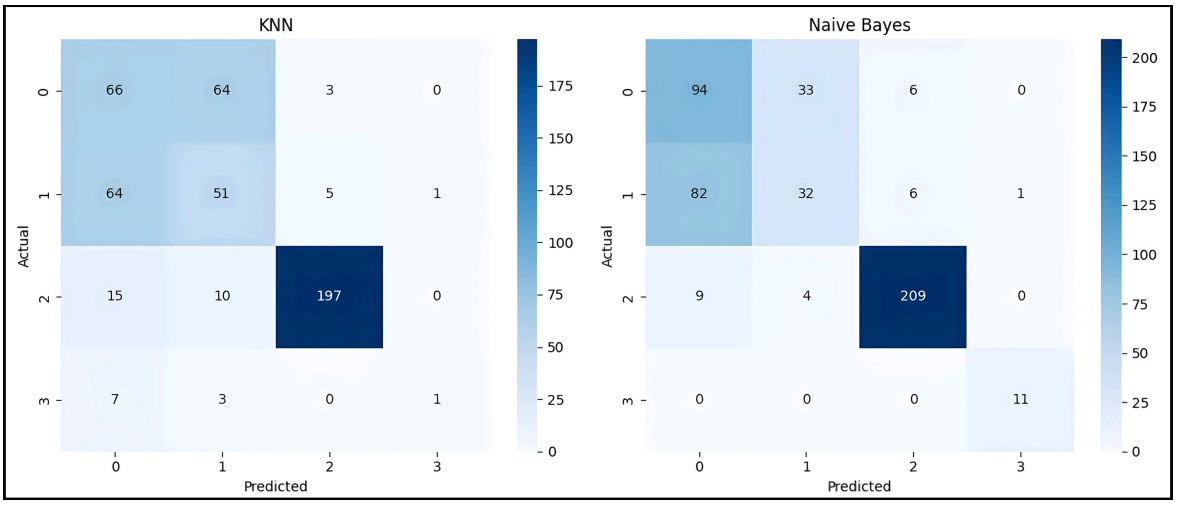

Figure 4.2: Confusion matrices of KNN & Naive Bayes.

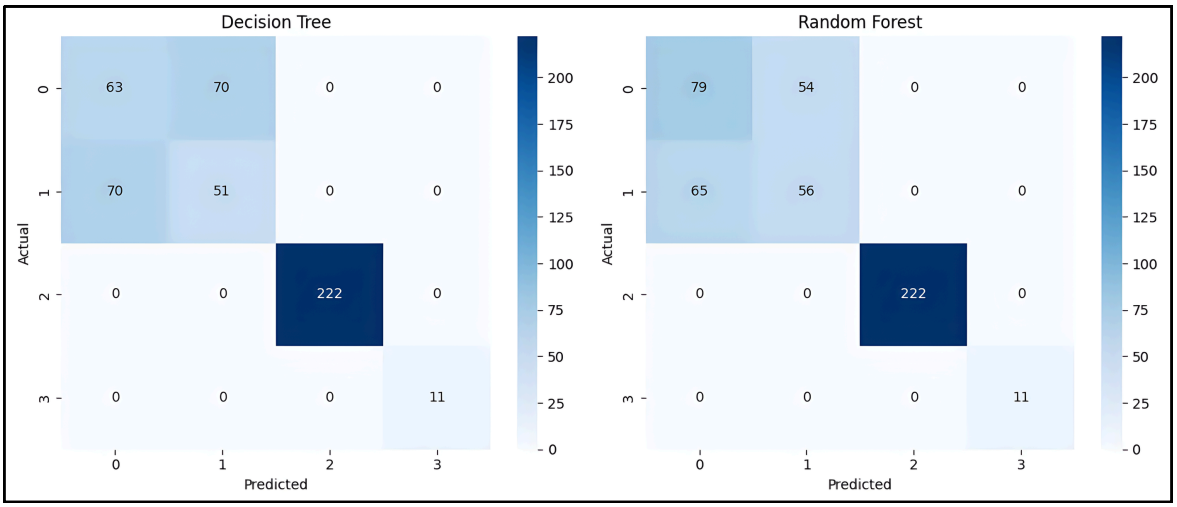

Figure 4.3: Confusion matrices of Decision Tree & Random Forest.

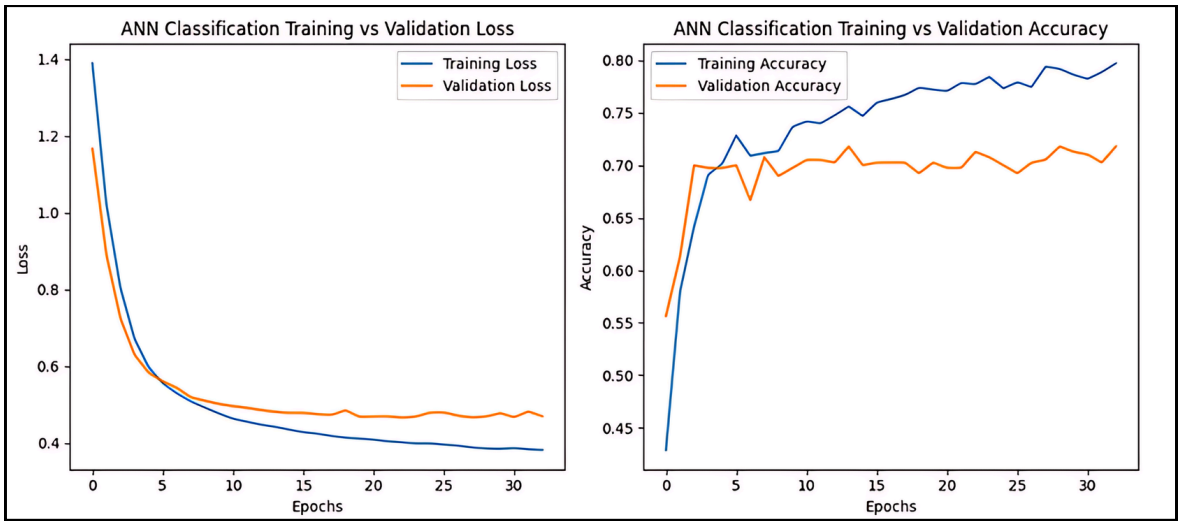

Figure 4.4: Loss and accuracy curves of the ANN classification.

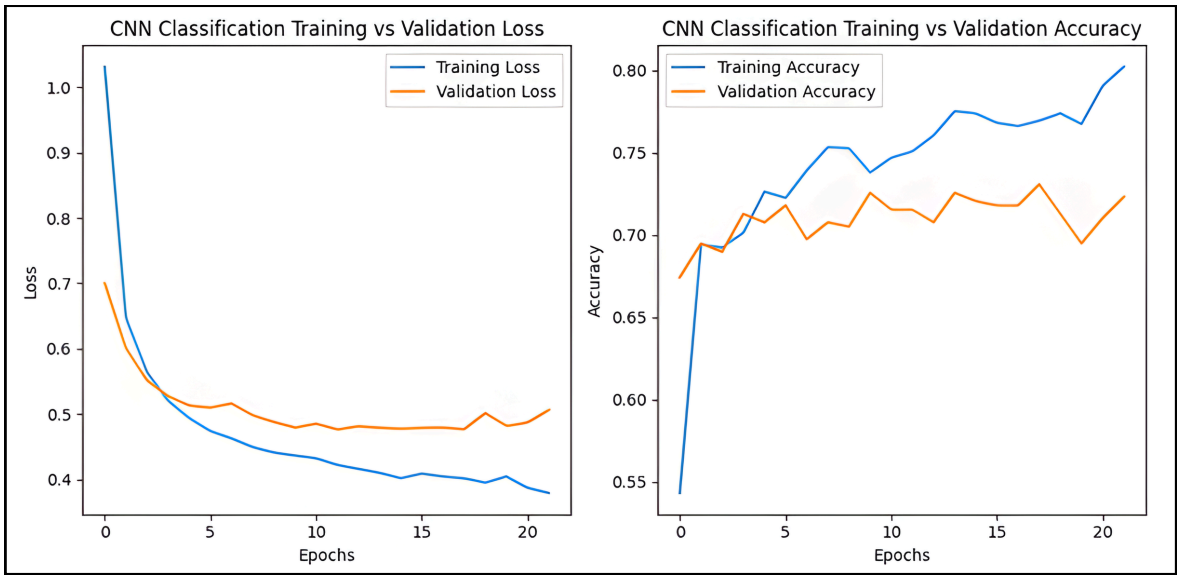

Figure 4.5: Loss and accuracy curves of the CNN classification.

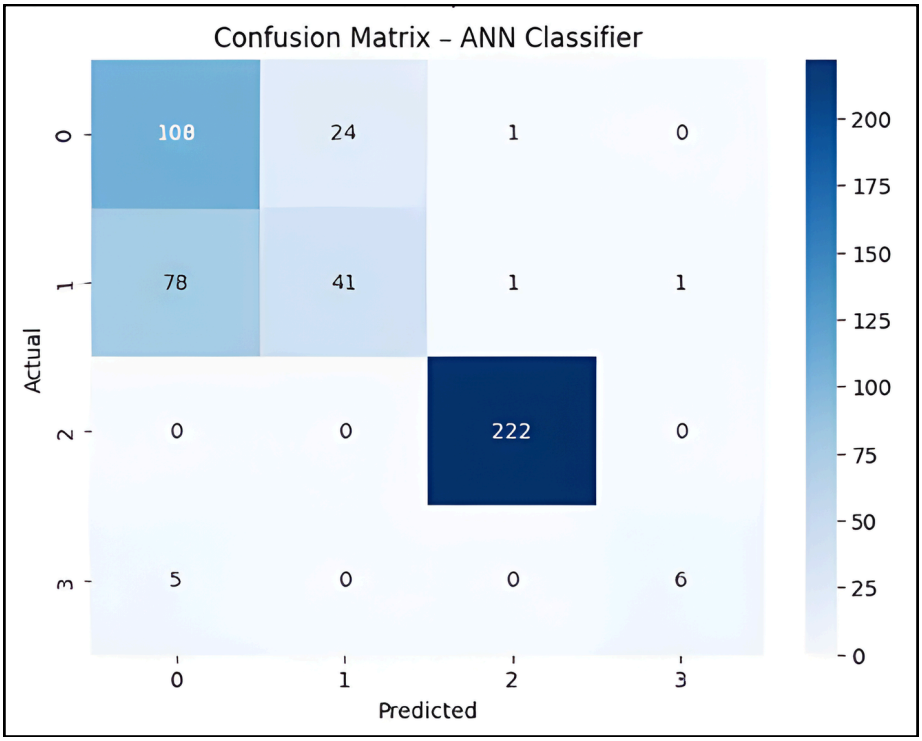

Figure 4.6: Confusion matrix of the ANN classification.

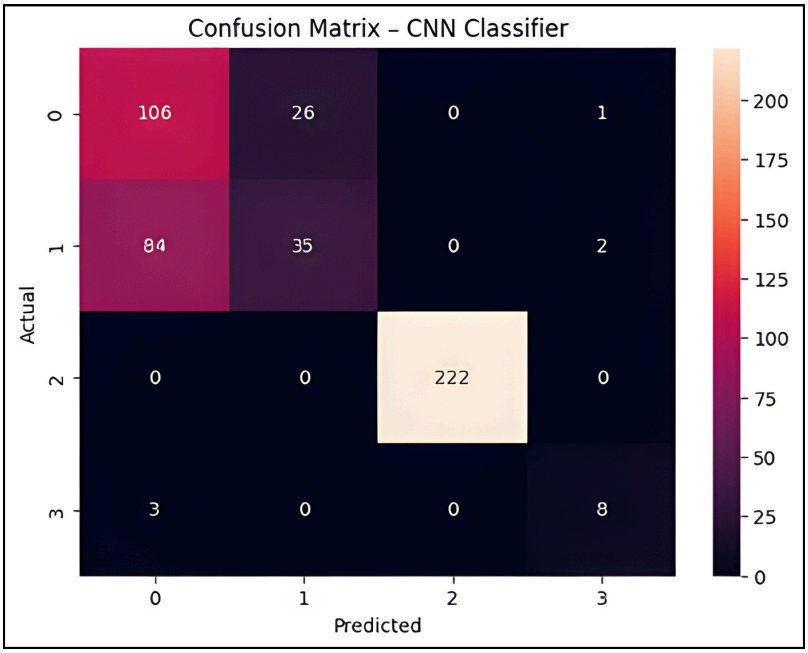

Figure 4.7: Confusion matrix of the CNN classification.

4.2 Regression Performance Analysis

For regression, traditional machine learning models were employed to predict mohor values, evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) [1]. Deep learning approaches, including Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN), were adapted for continuous prediction. Training and validation loss curves were examined to assess convergence and overfitting, while actual–predicted plots validated alignment with ground truth. Model summaries further highlighted architectural design and parameter complexity.

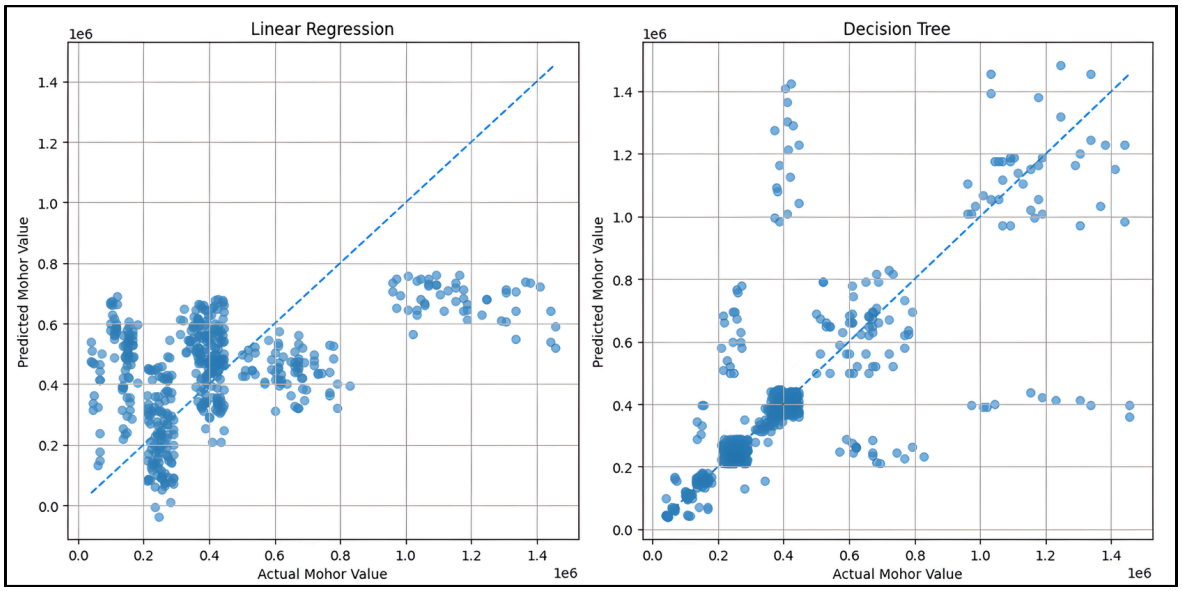

Figure 4.8: Scatter Plot of Linear Regression and Decision Tree.

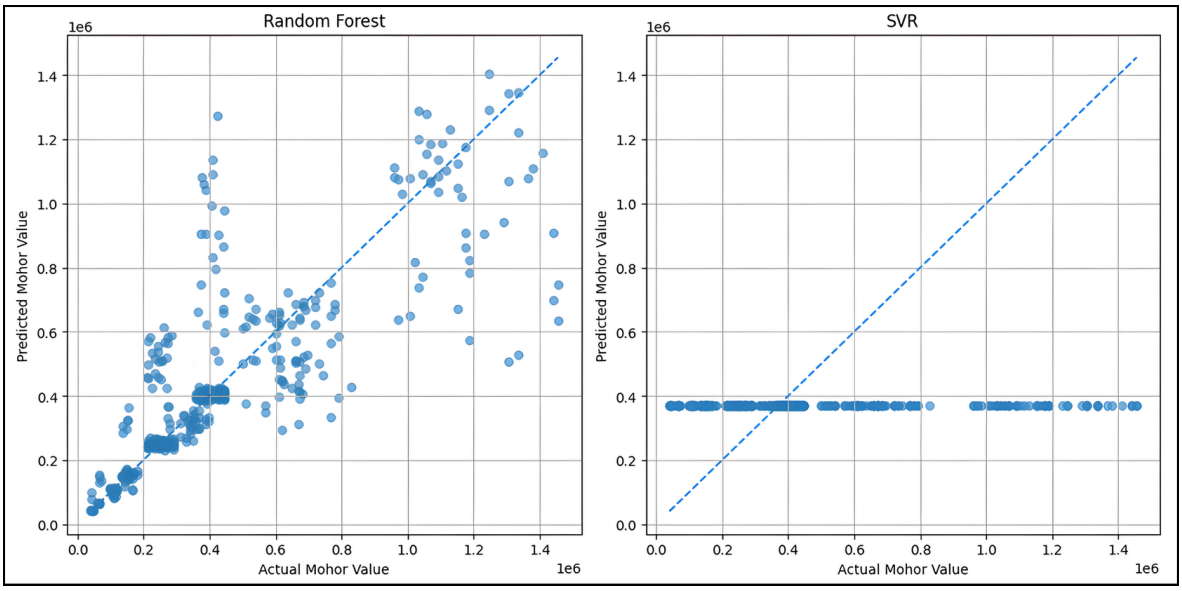

Figure 4.9: Scatter plot using Random Forest and SVR.

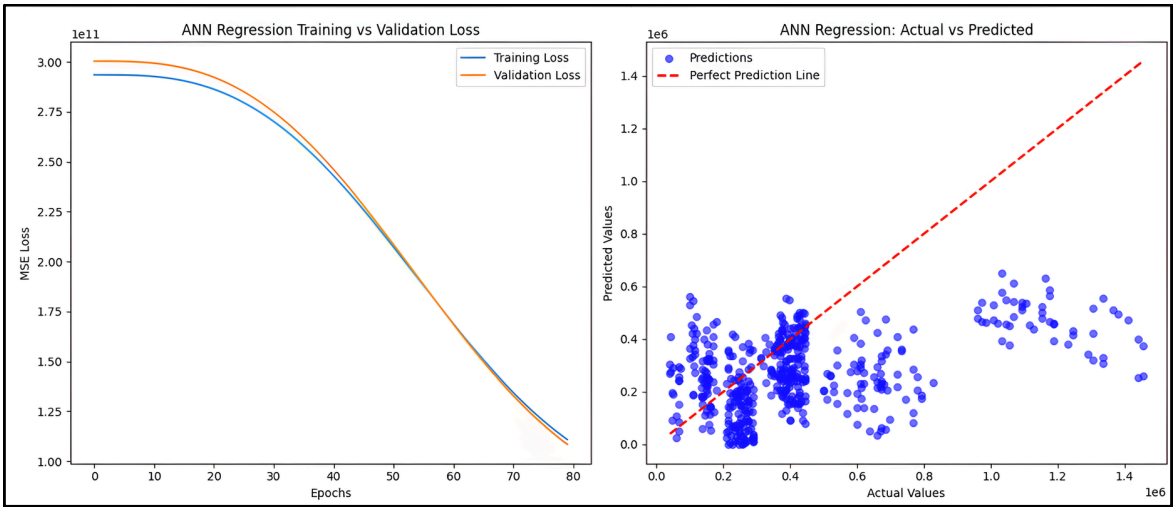

Figure 4.10: Loss curve and Scatter plot of the ANN Regression.

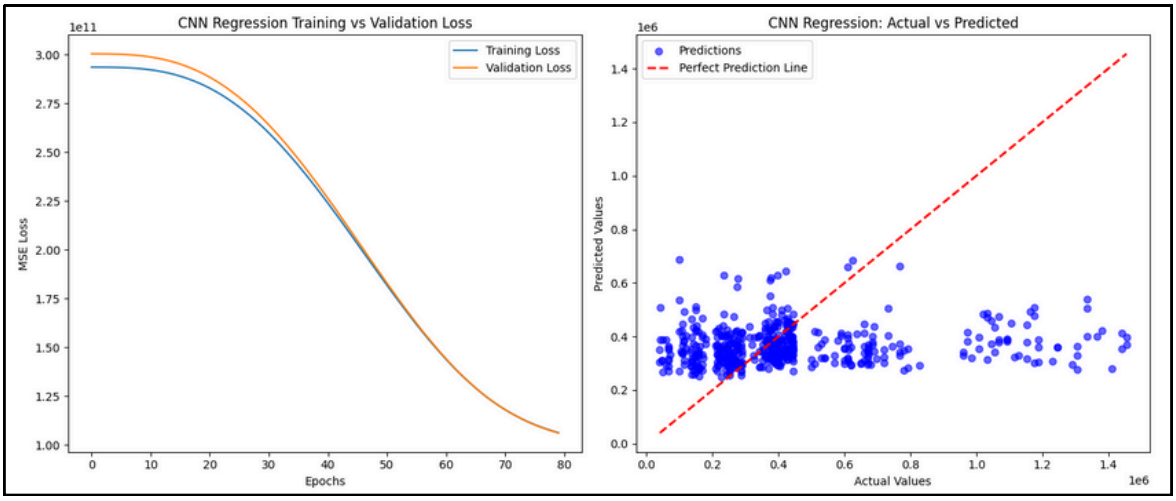

Figure 4.11: Loss curve and Scatter plot of the CNN Regression.

4.3 Model Comparison

A comparative analysis was conducted to evaluate the performance of all implemented models for both classification and regression tasks. The comparison was based on standard evaluation metrics to identify the most effective models for marriage condition prediction and mohor value estimation.

4.3.1 Classification Model Comparison

Table I presents the performance comparison of all classification models using accuracy, precision, recall, and F1-score. Among the traditional machine learning models, Logistic Regression achieved competitive performance with an accuracy of 76.59%, demonstrating stable and balanced classification results. The Random Forest classifier also performed well, achieving an accuracy of 73.51%, indicating the effectiveness of ensemble learning for this dataset. The Artificial Neural Network (ANN) achieved the highest classification accuracy of 77.41%, indicating its ability to capture complex feature relationships. The Convolutional Neural Network (CNN) also demonstrated competitive accuracy at 76.18%, though slightly lower than the ANN. However, precision, recall, and F1-score values were not computed for the deep learning models due to output formatting limitations. Among the remaining models, Support Vector Machine, Decision Tree, and Naive Bayes showed moderate performance, while k-Nearest Neighbors (k-NN) achieved the lowest accuracy, indicating limited effectiveness for this dataset. Overall, the classification comparison shows that both ANN and Logistic Regression performed effectively, with ANN achieving the highest accuracy.

Table I: Performance comparison of classification models

Model	Acc.	Prec.	Rec.	F1
LR	0.766	0.765	0.766	0.761
SVM	0.733	0.731	0.733	0.720
k-NN	0.647	0.667	0.647	0.652
NB	0.710	0.706	0.710	0.697
DT	0.715	0.715	0.715	0.715
RF	0.735	0.734	0.735	0.734
ANN	**0.774**	–	–	–
CNN	0.762	–	–	–

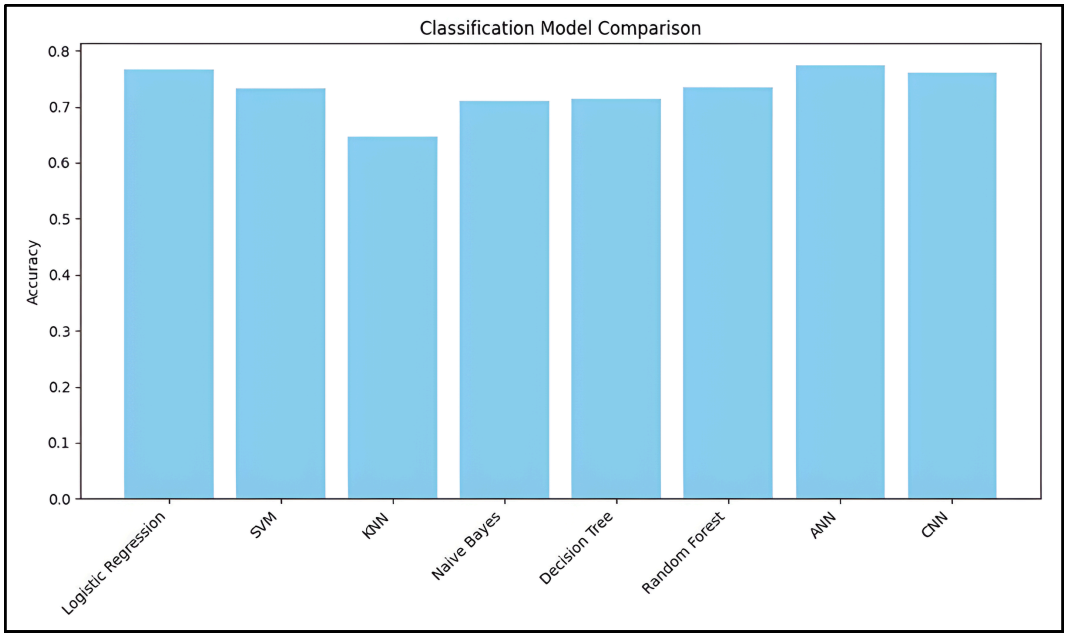

Figure 4.12: Accuracy comparison of classification models.

4.3.2 Regression Model Comparison

Table II summarizes the regression model performance using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2).
The Random Forest Regressor achieved the best performance among all regression models, with the lowest MAE of 94,761.23, the lowest RMSE of 173,665.75, and the highest R^2 score of 0.67, indicating strong predictive capability and good variance explanation.

The Decision Tree Regressor demonstrated moderate performance with an R^2 value of 0.45, while Linear Regression showed limited predictive ability with an R^2 score of 0.21, suggesting insufficient modeling of nonlinear relationships. Deep learning models, including ANN and CNN, exhibited higher error values and negative R^2 scores, indicating poor generalization for the regression task. Similarly, Support Vector Regression (SVR) also resulted in a negative R^2 value, reflecting limited effectiveness for this dataset.

Table II: Performance comparison of regression models

Model	MAE	RMSE	R^2
LR	212226	268905	0.215
DT	113263	229766	0.427
RF	**95585**	**177399**	**0.658**
SVR	206465	310056	-0.043
ANN	222637	306212	-0.018
CNN	209391	317382	-0.093

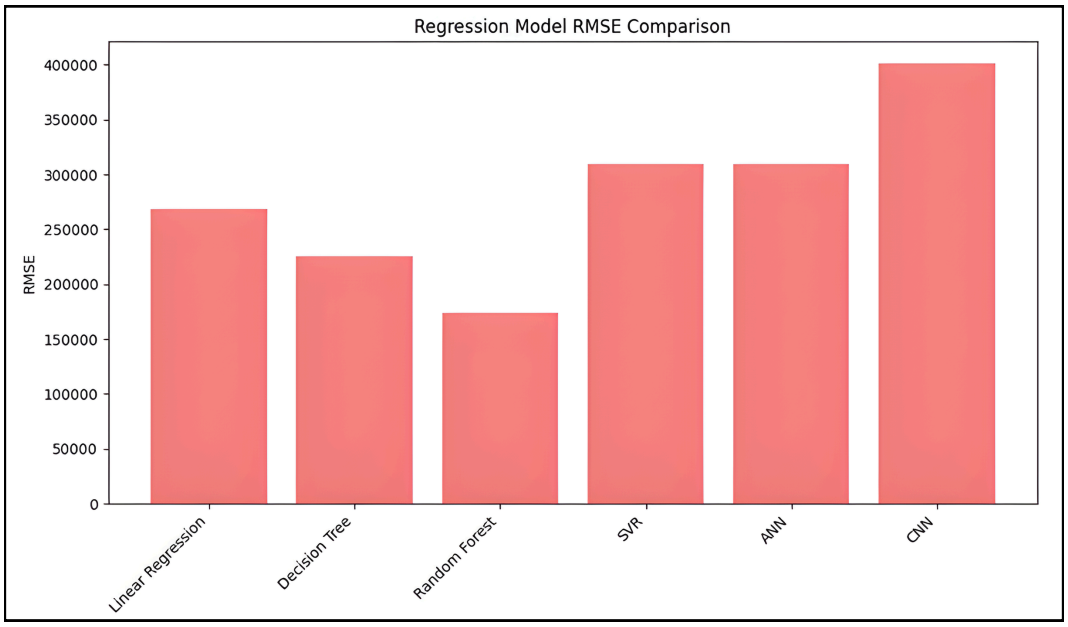

Figure 4.13: RMSE comparison of classification models..

Chapter 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

This study presented a comprehensive machine learning and deep learning–based approach for analyzing and predicting marriage-related outcomes using a structured socio-economic dataset. Multiple traditional classification and regression algorithms were implemented alongside advanced deep learning models, including Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN), to evaluate their predictive effectiveness.

For classification tasks, the ANN achieved the highest accuracy, demonstrating its capability to model complex, non-linear relationships within the dataset. Traditional machine learning models such as Logistic Regression, Random Forest, and Support Vector Machine also produced competitive results, highlighting their reliability for structured data. In regression analysis, the Random Forest model outperformed other approaches in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2), indicating strong predictive stability.

Overall, the experimental results confirm that ensemble learning methods and neural networks can significantly enhance prediction performance when applied to socio-economic datasets. The comparative evaluation provides valuable insights into the strengths and limitations of each model, supporting informed model selection for similar real-world applications.

5.2 Future Work

Although promising results were achieved, several directions remain open for future research. First, the dataset can be expanded by incorporating additional demographic, economic, and regional features to improve generalization and robustness. Addressing class imbalance and applying advanced feature selection techniques may further enhance model performance.

Second, more sophisticated deep learning architectures, such as Long Short-Term Memory (LSTM) networks and Transformer-based models, could be explored to capture temporal or contextual dependencies in the data. Hyperparameter optimization techniques and automated model selection frameworks may also improve predictive accuracy.

Finally, deploying the proposed system as a web-based or mobile application could enable real-time decision support. Ethical considerations, interpretability, and fairness analysis should be integrated to ensure responsible and transparent use of predictive models in socially sensitive domains.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[4] S. Raschka and V. Mirjalili, *Python Machine Learning*, 3rd ed. Birmingham, U.K.: Packt Publishing, 2019.

[5] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," *in Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, Savannah, GA, USA, 2016, pp. 265–283.

[7] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2011.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *in Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[10] S. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019.