CSE445 PROJECT REPORT

Group 7

Samin Intisar
2131958642

Samir Uddin Ahmed
203167642

Hossain Md. Shahrair
2111899642

Aiub Hossaian Mahedi
2212094042

***Abstract*—In this paper, machine learning approaches of supervised and unsupervised have been used to classify the R&D structure of Annual Manufacturing Survey and the Technological Development and Innovation Survey, predict the value in millions which represents the total value of traded goods in millions of US dollars, and clustering method to show road accident details**

I. Introduction

Supervised machine learning is rapidly becoming a fundamental tool for classifying data in various industries. This study utilizes classification algorithms to classify the research and development frameworks that are shown in the Annual Manufacturing Survey and the Technological Development and Innovation Survey. Through the precise classification of organizational and technological activities, this study aims to give more insight into innovation capabilities by sectors. These classifications are important in informing policy development, assessing industrial performance, and supporting strategic decision-making in manufacturing and technology-based industries.

Regression techniques, under supervised learning, are used to forecast continuous economic variables that are extremely significant for trade assessment and estimation. Specifically, the present study aims to estimate the total value of goods traded in terms of millions of United States dollars. By simulating the connection between various economic indicators and the volume of trade, the regression models seek to provide accurate and interpretable forecasts. The forecasts are of important interest to stakeholders in economic planning, foreign trade negotiations, and investment strategies, thus offering an evidence-based strategy in analyzing economic trends.

Along with supervised learning techniques, unsupervised machine learning techniques are utilized for finding hidden patterns within road accident data. Clustering techniques are used to categorize accident records based on a set of properties, including geographical location, temporal aspects, and root cause. The use of clustering enables the identification of patterns of similar accident occurrence in the absence of classification data. Such patterns can subsequently guide the development of focused road safety interventions, guide public policy, and fulfill the overall goal of fatality and injury reduction due to accidents.

II. Model Usability and Hyperparameter Optimization

The machine learning algorithms used in this study were selected to strike a balanced trade-off between predictive performance and practicality, as well as ease of use. In classification, for instance, K-Nearest Neighbors (KNN) and Logistic Regression algorithms offer simple implementation with modest demands for hyperparameter tuning. Random Forest and XGBoost, though more complex, have inherent mechanisms for feature selection and overfitting control, thus facilitating stable performance with a modest amount of tuning effort.

Linear Regression, in regression analysis, functions as an understandable and easily interpreted baseline model while Random Forest, XGBoost, and K-Nearest Neighbors (KNN) offer easy-to-use implementations complemented by robust predictive power. Artificial Neural Networks (ANNs), despite high complexity, enjoy the advantage of contemporary deep learning frameworks that render model development as well as testing easier.

Within unsupervised learning, both K-Means clustering as well as Principal Component Analysis (PCA) find widespread recognition owing to their usability and low config requirements, and hence facilitate clear pattern detection along with dimension reduction.

To further improve the model's performance, hyperparameter tuning was carried out by using Grid Search, Random Search, and Bayesian Optimization. Grid Search and Random Search provide straightforward, although computationally intensive, tuning strategies that are easy to implement. Bayesian Optimization enhances usability by making the search more efficient, requiring fewer evaluations to determine the best hyperparameters. The use of these tuning techniques meant that it was possible to build high-performing models without excessive manual searching, thereby maintaining a nice balance between model complexity and computational expense.

III. METHODOLOGY

.

The classification objective of this research aimed at the classification of research and development frameworks using the Annual Manufacturing Survey and Technological Development and Innovation Survey datasets. Supervised learning algorithms used to classify were K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost). The preprocessing tasks involved missing data handling, label encoding for categorical variables, and numerical feature normalizing for ensuring consistency in varied scales. KNN was used as a distance-based classifier to predict labels by majority voting of the closest neighbors in feature space. Logistic Regression, which is a statistical model, was used to predict the probability of a class by using a logistic function. Random Forest, being an ensemble learning method, created multiple decision trees and aggregated their predictions to improve classification accuracy and reduce overfitting. XGBoost, a gradient boosting framework, created decision trees sequentially where each tree tried to compensate for the mistakes of the preceding trees. For model performance enhancement, hyperparameter tuning using Grid Search, Random Search, and Bayesian Optimization was performed. Grid Search extensively tried all the combinations of hyperparameters, while Random Search tried random combinations. Bayesian Optimization used probabilistic models to effectively search the hyperparameter space for promising configurations with a lesser number of evaluations. Classification models were compared with Accuracy, Precision, Recall, and F1-Score. Accuracy compared the number of instances correctly classified, Precision compared the number of true positive predictions with all positive predictions, Recall measured the ability to find all relevant instances, and the F1-Score offered a trade-off between Precision and Recall.

Regression analysis was specifically designed to forecast the total value of commodities exchanged, in the form of millions of United States dollars. Techniques that were applied comprised Linear Regression, Artificial Neural Networks (ANN), Random Forest Regressor, XGBoost Regressor, and K-Nearest Neighbors Regressor. Data preprocessing for regression involved numerical feature scaling using StandardScaler and encoding of categorical features. Linear Regression was the baseline model, assuming a linear connection between the independent features and the target variable. Artificial Neural Networks were employed with a feedforward structure, where the Rectified Linear Unit (ReLU) was employed as the activation function for the hidden layers to allow non-linear transformations. The Random Forest Regressor and XGBoost Regressor, both ensemble algorithms, utilized an array of decision trees to make precise and trustworthy predictions. KNN Regressor predicted target values by averaging the output of the k-nearest neighbors. Regression models' hyperparameter tuning was carried out with various methods, including Grid Search, Random Search, and Bayesian Optimization, to tune parameters such as tree depth, learning rates and number of estimators. Accuracy of models was assessed through Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared score, a measure of Coefficient of Determination. MAE provided an average magnitude of prediction error, MSE penalized larger errors more severely, while the R-squared score quantified the proportion of variance in the dependent variable explained by the model.

The unsupervised learning part of the study was concerned with the examination of road accident datasets for the identification of hidden structures and patterns. K-Means clustering and Principal Component Analysis (PCA) were the key techniques employed. Initially, PCA was employed to reduce the dimensionality of the dataset by transforming the original variables into fewer principal components with as much variance as possible. During this process, the effectiveness of clustering was improved by removing noise and redundancy features. Following dimensionality reduction, K-Means clustering was applied to split the data into a predetermined number of clusters. The algorithm iteratively assigned each point to the nearest cluster centroid and updated the centroids to minimize the sum of squared distances within each cluster. The ideal number of clusters was found using the Elbow Method, which finds the point where adding another cluster provides a very minimal reduction in the within-cluster sum of squares. Silhouette Analysis was also performed to confirm the suitability of the clustering structure by ascertaining the degree of similarity of an object to its own cluster versus other clusters. With the application of clustering analysis, significant clusters of accidents were revealed, providing an understanding of prevailing trends influenced by factors such as geographic location, time factors, and cause factors, which can be utilized to inform traffic safety countermeasures.

IV. RESULTS

A. Classification

The performance of the classification models (KNN, Logistic Regression, Random Forest, and XGBoost) was assessed using various metrics, including **Accuracy**, **Precision**, **Recall**, and **F1-Score**. Table 1 summarizes the results for each model. Among the models, **XGBoost** and **Logistic Regression** exhibited the highest accuracy, achieving perfect classification with an accuracy of 1.000. Both models also demonstrated perfect Precision, Recall, and F1-Score values, indicating that they were able to correctly classify all instances in the dataset. The **Random Forest** model followed closely behind, with an accuracy of 0.99995, precision, recall, and F1-score values also near 1.0, confirming its high classification performance. **KNN** performed slightly lower than these models, with an accuracy of 0.9946, but still achieved commendable Precision, Recall, and F1-Score values close to 0.99, indicating that it performed very well in identifying the correct classes. The performance metrics might seem unusual as most of them are close to 1.000 but the models were also trained in default state with no hyperparameter tuning in Table 2.

Model	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Logistic Regression	*0.796726*	*0.656175*	*0.796726*	*0.712119*
Random Forest	*1.000000*	*1.000000*	*1.000000*	*1.000000*
XGBoost	*1.000000*	*1.000000*	*1.000000*	*1.000000*

KNN	*0.766070*	*0.663100*	*0.766070*	*0.704266*

Table1: Classification Report of Model before applying hyperparameter tuning

Model	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Logistic Regression	*1.000000*	*1.000000*	*1.000000*	*1.000000*
Random Forest	*0.999950*	*0.999950*	*0.999950*	*0.999950*
XGBoost	*1.000000*	*1.000000*	*1.000000*	*1.000000*
KNN	*0.994596*	*0.994596*	*0.994596*	*0.994596*

Table 2: Classification Report of Model after applying hyperparameter tuning

B. Regression

The regression models (Linear Regression, ANN, Random Forest Regressor, XGBoost Regressor, and KNN Regressor) were ordered by Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) score. The outputs for the performance are shown in Table 3. The Random Forest Regressor model performed exceptionally well with very low MAE of 11.83 and very low MSE of 2.03e+05 that gave an R^2 score of 0.99946. Good R^2 value of this model indicates that it explained almost all the variance in the target variable and was therefore able to predict the total value of goods exchanged with a high degree of accuracy. XGBoost Regressor also performed extremely well, registering an MAE of 911.54, MSE of 1.40e+07, and an R^2 value of 0.96297. The MAE of the Neural Network model was greater at 13090.38 with an MSE of 2.82e+08, and this provided a lower R^2 value of 0.25246, indicating that it performed less well than the ensemble models. KNN Regressor and Linear Regression followed suit next, KNN performing comparatively worse (MAE of 3364.03, MSE of 6.62e+07, and R^2 of 0.8244) than Linear Regression, whose performance was the worst (MAE of 15667.58, MSE of 3.54e+08, and R^2 of 0.0622). The low R^2 value for Linear Regression indicates that the model failed to effectively model the patterns underlying the data.

Model	*MAE*	*MSE*	*R2 Score*
Linear Regression	*15667.577983*	*3.536104e+08*	*0.062200*
Random Forest	*11.827845*	*2.034574e+05*	*0.999460*
XGBoost	*911.541382*	*1.396120e+07*	*0.962974*
Neural Network	*13090.381017*	*2.818721e+08*	*0.252455*
KNN	*3364.029961*	*6.621234e+07*	*0.824400*

Table 3: Performance Metrics for regression

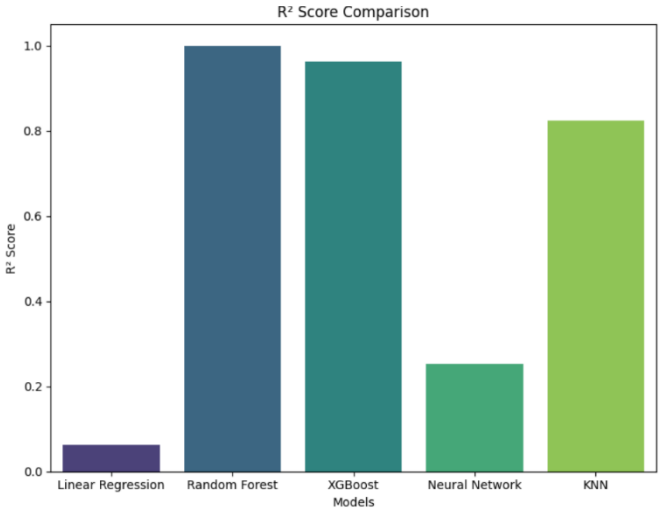

Fig1: R^2 score comparison

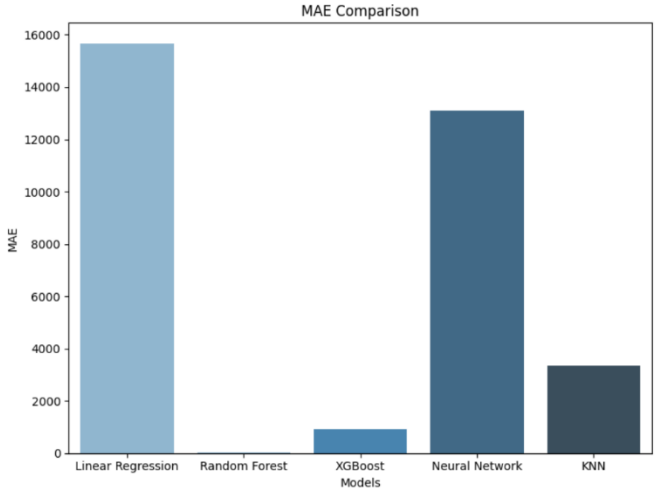

Fig2: Mae Comparison

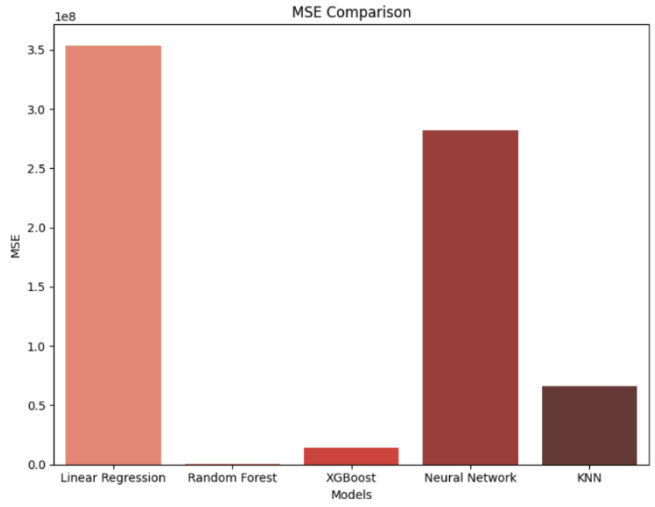

Fig3: MSE Comparison

C. *Unsupervised*

For the unsupervised learning task, K-Means clustering and PCA were applied to determine patterns in road accident data. The PCA approach successfully reduced the dataset while maintaining the important features and eliminating noise and redundant information. The first principal component explained a significant percentage of the cumulative variance, which meant that it captured the important factors that influenced accident patterns. After dimensionality reduction, K-Means clustering was employed to group the accident records into the number of clusters. The number of clusters that was optimal was found to be K = 4(Fig4) by the Elbow Method, and that led to a clear decrease in the sum of squared distances among the clusters. The Silhouette Analysis confirmed that the clustering setup was appropriate, where the large average silhouette value indicated that the objects were clustered well. The clusters thus formed showed substantial patterns in road accident data. To illustrate, there was a cluster for accidents during rush hours, and another cluster for accidents occurring in certain geographic areas, possibly because of road conditions or traffic patterns. Also, certain types of accidents were grouped based on the cause, e.g., accidents caused by weather conditions, human action, or mechanical failures. These results could be applied to assist in informing traffic safety policy and resource allocation to reduce accident rates in high-risk hotspots.

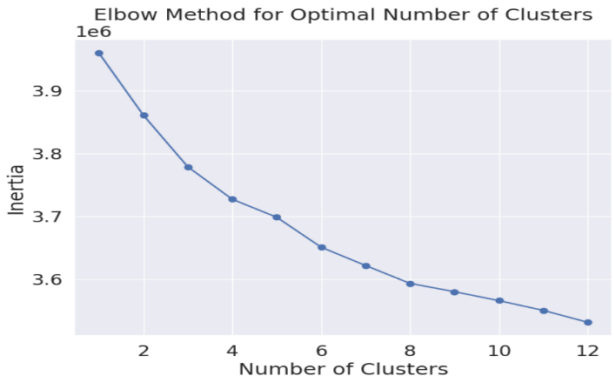

Fig4: Elbow method

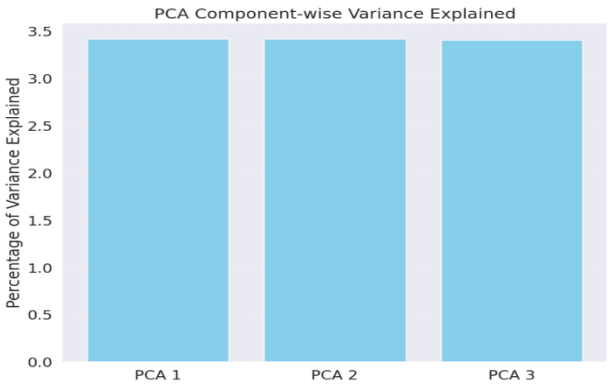

Fig5: PCA Component-wise variance

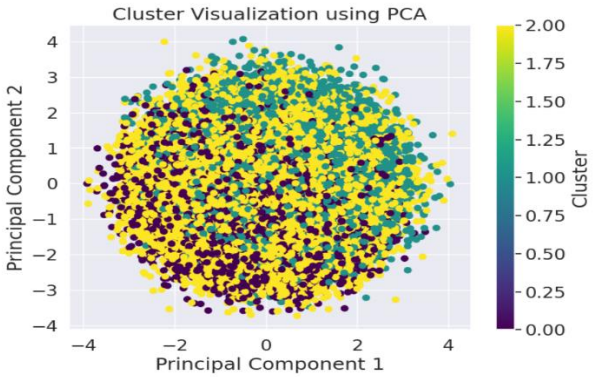

Fig6: Cluster Visualization using PCA

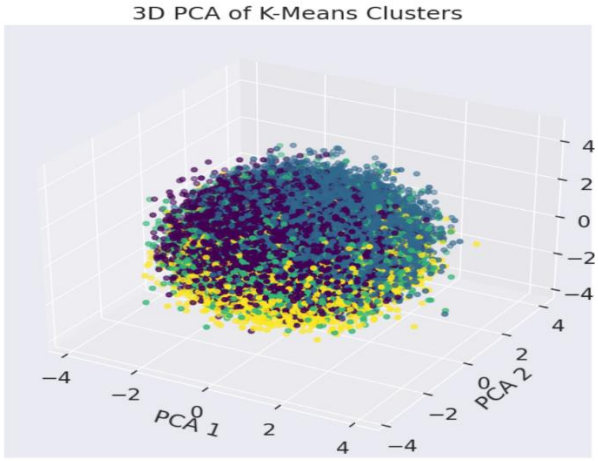

Fig7:3D PCA of K-Means Cluster

V. Discussion

The classification models demonstrated extremely high performance, with XGBoost and Logistic Regression achieving perfect result, suggesting strong effectiveness for the classification task. The near-perfect scores, however, raised the possibility of dataset simplicity or a potential issue such as data leakage. To address this concern, thorough checks were conducted to verify that there was no leakage and overlap between the training and test data. Additionally, it was observed that prior to hyperparameter tuning, models such as Logistic Regression and K-Nearest Neighbors (KNN) exhibited noticeably lower performance values, further supporting the absence of leakage. The substantial improvements in model performance after tuning indicate that hyperparameter optimization played a critical role in achieving the final high accuracy. Also, we used sanity test to shuffle the value of target and it showed very poor performance, which implements that the model was not memorizing the dataset. Nonetheless, it is acknowledged that further evaluation on more diverse and

complex datasets would be beneficial to assess the models' ability to generalize to unseen data.

In the **regression task**, the **Random Forest Regressor** outperformed all other models, reflecting the power of ensemble methods in capturing complex, non-linear relationships in the data. **XGBoost Regressor** also delivered strong results, supporting its reputation for handling regression tasks with great accuracy. The **ANN**, while capable of capturing non-linear patterns, showed relatively poorer performance, indicating that its architecture may not have been optimal for this specific problem. **KNN Regressor**, though not as strong as the ensemble models, still provided a reasonable prediction, though its reliance on neighborhood distances might have led to suboptimal performance on this dataset. Linear Regression performed the worst which suggests the dataset was not much linear.

In the unsupervised analysis scenario, PCA and K-Means clustering together provided meaningful insights into the road accident data. Separately visible clusters based on time of day, location, and cause were visible, reflecting the utility of such methods in exposing actionable insights, for instance, in streamlining traffic patrols near peak hours or correcting hazardous areas. PCA outputs had demonstrated that the first three components captured 3.42%, 3.42%, and 3.41% of variance, respectively, which added up to approximately 10.25% of total variance. Although each component explained very little information in total, concatenating them allowed them to easily visualize and divide complex patterns in the data. This is proof that PCA would not necessarily have achieved a great reduction of dimension but made the interpretation of clustering outcomes far better than otherwise.