Data Mining and Machine Learning Project

Vipin Sharma
Dept. of Computing
National College of Ireland
Dublin, Ireland
x22207406@student.ncirl.ie

***Abstract*—This study examines the application of five different machine learning models to three distinct large datasets in a machine learning project. The dataset came from the retail, salary, and house prediction domains. The first two datasets were sourced from Kaggle, while our lecturer supplied the third dataset of retail transactions. Retail transaction data is used for the Apriori algorithm, house data is used for regression problems, and salary data is used for classification problems. The first set of data included individuals who make more than 50K annually or do not. DecisionTreeClassifier and GaussianNB classification models were applied. With an accuracy score of 80.38%, GaussianNB shows a higher fit than DecisionTreeClassifier for the salary prediction dataset. Regression models KNeighborsRegressor and Linear Regression were used to predict the housing price for the second set of data. The KNeighborsRegressor model exhibits exceptional accuracy and predictive power. The most accurate model for estimating home prices is this one. In contrast, the MSE was 9408039505.76, the RMSE was 96995.04, the MAE was 50243.21, the R-squared score was 0.88, and the Explained Variance Score was 0.88. For the third Retail transaction data the Apriori algorithm gives the recommended product based on the previous transaction done by the users.**

***Index Terms*—Machine Learning, DecisionTreeClassifier, GaussianNB, Linear Regression, KNeighborsRegressor, Recommendation, Apriori algorithm.**

I. Introduction

Machine learning has revolutionized traditional approaches and improved decision-making processes in a variety of industries thanks to its ability to extract patterns and insights from large datasets. Salary prediction, housing, retail industry are important industries that have been greatly impacted by machine learning algorithms. An analysis paper discussed the high rate of inflation in Indonesia. The monthly inflation rate for September 2022 was 1.17%, as reported by Indonesia's Central Bureau of Statistics (BPS). Ever since December 2014, this rate had reached its peak. One of the main causes of September's inflation is inner-city transportation, along with fuel and rice [1]. In the real life many things depend on salary such as shopping, eating, clothing etc. In the research paper I have to predict salaries that are over 50K or under 50K, this study employs classification models like the GaussianNB and DecisionTreeClassifier. Features Age, workclass, number of years of education, marital status, type of work, relationship, race, sex, hours worked per week, and country of origin are the factors that influence pay. The study question for Salary Prediction data is:

"Developing the most effective classification model to determine the correct salary class, whether it is over 50K or under 50K?"

A sophisticated and data-driven approach to valuation is provided by machine learning algorithms, which are crucial in the housing sector for predicting house prices. More precise and dynamic pricing predictions are produced by these models through the analysis of numerous variables, such as market trends, property features, and location. Many researchers performed the modeling on the housing dataset and gave the different predictions and the results. The research paper [2] shows From 1945 to 2018, the average annual increase in sale prices was 8.4%, while the average annual increase in consumer prices was 5%. It is estimated that market rents have increased by 6.3% annually, significantly higher than previous estimates of 4.4%. This finding has implications for calculating living standards and costs of living in Ireland since World War II. Although sale prices have varied over the same time period, there is some indication that rents are convergent across city markets. Since 1945, there have been four significant housing market cycles (accounting for inflation), with peaks occurring in the late 1940s, early 1970s, early 1980s, and mid-2000s. For the analysis of Miami housing data in this study, I used KNeighborsRegressor and Linear Regression. The study question for housing data is:

"To build an effective house price prediction model using machine learning techniques"

Machine learning has become an important tool in the ever-changing retail sector. Recommendation systems are one of the main areas where machine learning has shown its transformative potential. These systems use advanced algorithms to analyze past information, customer preferences, and behavior to provide retailers with personal and focused recommendations. This not only makes shopping more enjoyable overall, but it also makes a big difference in performance and sales. Recommendation engines are a valuable tool in the retail industry because they can impact customers' decisions and build brand loyalty. Retailers can customize product recommendations, promotions, and content to enhance customer engagement and personalization by anticipating future needs and understanding individual preferences. Customer satisfaction rises as a result, and conversion rates and income are also increased. Recommendation systems have been widely used in

a variety of industries outside of retail. These algorithms are used by streaming services like Netflix,Youtube and Spotify to curate content and offer users personalized recommendations for movies or songs. Comparably, recommendation algorithms like Apriori and Eclat are used by online retailers such as Amazon to make product recommendations based on past browsing and purchase behavior.

The Apriori algorithm was used to build the Recommendation engine. Recommender systems compute and present the user with content that is relevant to them based on information about the user, the content, and their interactions with the items. Although the massive online retailer has never disclosed its own figures, McKinsey calculates that 35% of customer purchases on Amazon originate from product recommendations [3]. The study question of Retail transaction data is :

"To make a recommendation system using the Apriori algorithm which recommended the product based on previous transactions"

The structure of the document is broken down as follows. Section 2 of this report includes a brief summary of earlier studies on the topic. We go over our Knowledge Discovery from Database, or KDD methodology, in Section 3 specifically. The machine learning algorithms were applied to all the data in Section 4, which follows, and an analysis was done to find out which approaches yielded better results for the first dataset's F1 score, precision, and recall value, and for the second dataset's RMSE, R-Squared, MSE, and MAE. We discussed our conclusions regarding the findings and observations in Section 5 and what could be added going forward to improve the performance over time.

II. Related Work

Over time, there have been many variations in the types of research conducted in the area of Salary Predictions. In this domain, Yanming Chen et al [4], carried out research on salary prediction based on the Dual Adabossting system in he is using the adaboositng algorithm to predict the salary of students who are students of data science in the United States. The accuracy measures Training (RMSE) is 2.83 and Testing (RMSE) is 17.57 and Testing (MAE) is 8.36. They do Experiments with the other models like the Lasso regression model, Ridge regression model, SVR, GradientBoosting model, and Randomforest Regression model. Another study by R. Kablaoui and A. Salman [5] using three machine learning models logistic regression, random forest, and neural network these model predict the salary. The output in the form of 0 and 1 if the salary is above 50K yearly put it 1 and below put it 0. In the study the neural network gives the most accurate results which are 83.2% and the random forest gives the least accurate result 80.7% it also takes more time around 8.49s to run the training model and the third model logistic model gives the 83% accuracy. Another Study by J. V. Siswanto et al [6] Salary classification and prediction based on the job field and location using the Ensemble method technique. These machine learning techniques include boosting classifier, voting classifier, bagging classifier, random forest, logistic regression, decision tree, and k-nearest neighbor in addition to support vector machine. At a 72% accuracy rate, Random Forest produced the best results. The F1 score of the random forest is 71%, and precision and recall is 71% and the top 2 most demanded job predicted Akuntansi Umum, It Software and Retail Sales. A further investigation by Guanqi Wang [7] determine and predict the Employee salaries with the machine learning algorithms. The total 5 regression models used Multiple Linear Regression, Polynomial Regression, Ridge Linear Regression, lasso Linear Regression, and Elastic-Net Regression to determine which model gives the best performance. Applying all this model the author finds the 2nd order polynomial regression model gives the best performance and fit predicts the salary. The Train (RMSE) is 31.06 and the Test (RMSE) is 29.41 of the Polynomial Regression model other models show high values compared to the Polynomial model. Another author, Sayan Das et al. [8] takes the small sample data set which includes the 3 columns and 10 rows and they use the regression technique to predict the salary. The author used two machine learning models Linear Regression and polynomial Regression. Both models make the graph and plot the data points and create the best-fit line for the salary dataset. They give the conclusion if anyone performing the K-nearest regression model then he will give more accuracy.

TABLE I
Table of Overviews of Related Works for Dataset-1

Authors & Year	**Methodologies Used**
Yanming Chen et al (2023)	Lasso regression model, Ridge regression model, SVR, GradientBoosting model, and Randomforest Regression model
R. Kablaoui and A. Salman (2022)	logistic regression, random forest, and neural network
J. V. Siswanto et al (2023)	boosting classifier, voting classifier, bagging classifier, random forest, logistic regression, decision tree, and k-nearest neighbor
Guanqi Wang (2022)	Multiple Linear Regression, Polynomial Regression, Ridge Linear Regression, lasso Linear Regression, and Elastic-Net Regression
Sayan Das et al (2020)	Linear Regression and polynomial Regression

Move to the next dataset which is housing dataset many researchers work on this type of dataset. In this domain, X. Wu and B. Yang [9], carried out research on prediction the House Price by using the Ensemble Learning based models. The reserachers takes the Miami housing prices dataset and apply a 5 different ML models like Random Forest, Neural Network, Linear Regression, SVR, and XGBoost. From this the Random Forest and XGBoost are the two ensemble learning techniques that yield the best results out of all the models;

the former two have adjusted R squares of 0.9234 and 0.9254, respectively. The RMSE value of Random Forest is 89388 and XGBoost is 88310 which is very less then comparing with the others RMSE models values like Linear regression have 175273, SVR have 123473 and Neural Networks have 97430 EMSE values, same with the MAPE values they are also very less. For both the models Random Forest has the MAPE is 0.1117 and for the XGBoost is 0.1128. Another study by Q. D. Trong et al [10] try to improve the Housing price prediction using three different machine learning models including Random forest, LightGBM, and XGBoost and other two models in ML used to compare the results and optimum solutions including Stacked Generalization and another one is Hybrid Regression.The Hybrid Regression give the best results and accuarcy. The RMSLE value on the train dataset in hybrid regression is 0.14969 which is very less than as compared with the other models and the value on test data is 0.16372. One more study carried out by Y. Chen et al [11] on the house price prediction based on deep learning and machine learning methods. The study uses the five deeply learning methods to predict the house price. The methods include in the research paper include Support Vector Machine (SVM), Deep Neural Network (DNN), Backpropagation neural network (BP neural network), Bayesian, and Linear Regression. Applied all this model on the housing data set. This study evaluates the model using MSE, RMSE, MAE, and 2. The Bayesian model yields the smallest MSE, RMSE, and MAE when compared to other methods. Its 2 approach reaches 0.9260, almost exactly equal to 1.The BP neural network and the Bayesian model on 2 value are comparable at the same moment. Thus, the BP neural network performs well in predicting house prices as well. The evaluation metrics of the three models are best for Bayesian and BP neural networks. Another research conducted by A. B. Adetunji et al [12], the authors use the Random forest machine learning technique to predict the house price. After completing the train and testing of the model the metrices comes and the R squared value is 0.900 which determine the accuracy of 90% , Mean Absolute Error (MAE) value is 1.900, Mean Squared Error (MSE) value is 6.702 and the Root Mean Square Error (RMSE) value is 2.588. Another study carried out by S. Lu et al [13] using the Hybrid Regression for house price prediction and this paper also used the hybrid lasso and Gradient boositng regression model to predict the house price. In this paper author performed the many iterations of the training dataset to determine the best results. Here author performing the features engineering on the different features. When apply the Ridge Regression and Lasso regression both got the minimum RMSE which is 0.112276 for train data where the number of features is 160. Another study conducted by The Danh Phan [14] take the case of Melbourne City house price using the machine learning algorithms predict the house price of Melbourne City. This research based on analysis of historical properties transactions in Australia using the historical data author applied Regression trees, Polynomial Regression, Neural networks and SVM machine learning model. The Train Mean Squared Error (MSE) of the models are followed by Linear Regression have 0.0948, Ploynomial Regression have 0.0773, Regression Tree have 0.0925, Neural Networks have 0.2657, Stepwise SVM have 0.0558, Stepwise tunes SVM is 0.0480, PCA SVM have 0.0721, and PCA tuned SVM have 0.0474. Th author founded that the Regression Tree gives the Good results while neural networks not work it give very high MSE value.The runtime of the fitting model of Neural Network is very less it takes only 0.033 minutes and the PCA tuned SVM takes very large time which is 0.733. One more study carried out by B. Park and J. K. Bae [15] the author take the housing dataset of Fairfax county and apply the c4.5, Ripper, Naïve bayesian and adaBoost machine learning model to determine the price of the housing. All the models are tested to see which one yields the highest accuracy rate. The author discover that RIPPER performs better than the C4.5, Naïve Bayesian, and AdaBoost models. RIPPER beats the other models for predicting housing prices in every test.

TABLE II
TABLE OF OVERVIEWS OF RELATED WORKS FOR DATASET-2

Authors & Year	**Methodologies Used**
X. Wu and B. Yang (2022)	random Forest, Neural Network, Linear Regression, SVR, and XG-Boost
Q. D. Trong et al (2020)	Random forest, LightGBM, Hybrid Regression, and XGBoost
Y. Chen et a (2021)	Support Vector Machine (SVM), Deep Neural Network (DNN), Backpropagation neural network (BP neural network), Bayesian, and Linear Regression
A. B. Adetunji et al (2022)	Random forest
S. Lu et al (2017)	Hybrid Regression, Gradient boositng regression
The Danh Phan (2018)	Regression trees, Polynomial Regression, Neural networks and SVM
B. Park and J. K. Bae (2015)	c4.5, Ripper, Na ıve bayesian and AdaBoot

Move to the next dataset which is retail transaction dataset many researchers work on this type of dataset. To enhance the grocery store recommendation system, researchers Kutuzova and Melnik [16] used a variety of data sources. Clustering, association rules mining, and collaborative filtering are some of the methods employed. They concluded that the results of the association rule mining were satisfactory. Another Study conducted by J. Dongre et al [17], that the Apriori algorthium is very popular data mining approach for finding the products sets from a transaction dataset. Todays various sectors make the good business particularly in the retail sector. Data mining is important process to targeting the audience using many campaigns and giving rewards point and vouchers. The author used the data mining technique to find the hidden pattern

of the most frequent used product-sets. Another research conducted by W. B. Zulfikar et al [18] Increasing the sales using by Association rule with Apriori Algorithum. The retail store have the very large history of sales data more than 10,000,000 records. By applying the association rules with apriori algorithm combined the sales for every single stores with their location. The association's rules are influenced by the store's location. The store is separated into 8 clusters; within a cluster, a association rule can be applied, but not to another cluster. The guidelines can be utilized as innovative marketing collateral to increase sales.

TABLE III
TABLE OF OVERVIEWS OF RELATED WORKS FOR DATASET-2

Authors & Year	Methodologies Used
K. Tatiana and M. Mikhail (2018)	Clustering, association rules mining
J. Dongre et al (2014)	Apriori algorthium
W. B. Zulfikar et al (2016)	Association rule with Apriori Algorithum.

III. METHODOLOGY

The Methodology is a way or technique that is used to implement the model. It is also a systematic process to design the model it briefly designs the steps that are important to build an effective model .

The Knowledge Discovery from Database also called the KDD is used for these datasets, with the help of KDD we can extract the information from all the 3 complete datasets. KDD is an iterative process, requiring multiple iterations of the previously mentioned steps in order to extract accurate knowledge from the data. This KDD approach has six steps: Data selection or collecting data, preprocessing and cleaning the data, transforming and preparing the data, second last one is data mining, and finally pattern evaluation. The subsequent actions as shown in Fig 1 [19], are part of the KDD process:

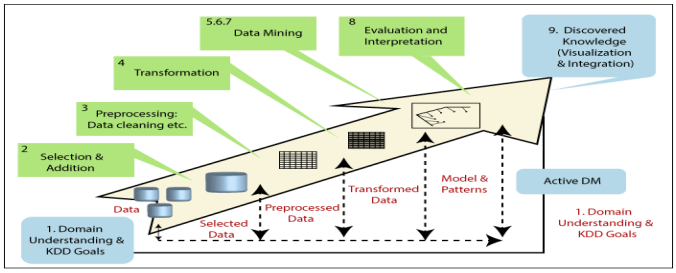

Fig. 1. Methodology Flow of KDD

A. Data Selection

The selection of data is the first step in the KDD Process choosing the right data for the modeling in machine learning is very important because if there is anything missing in the data like null values present, or missing values present then it creates problems for the models and also decreases the accuracy of the model.

Among the available data sources, the desired dataset is selected, and it is then used in the subsequent procedures. To find all three data sets various websites available but I used the Kaggle website, Kaggle is the largest database collection for various datasets which is useful for data science students, the researcher also there are various competitions running on the Kaggle website which is helpful for students.

1) Dataset 1: Salary Predictions data : I got the salary prediction dataset from the kaggle website the data is presented in the CSV format and the file name of the salary data is (salary_data.csv). The Salary dataset have the 32561 rows and 15 column. For the dataset I have to predict that the person make over 50K yearly or not. Choosing the independent variables X which exclude the salary column and included age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, region and the dependent variables Y (Salary). Based on the above independent vairabels I have to evaluate whether a person makes a salary over 50000 or not.

2) Dataset 2: Housing price Prediction Data: I got the Housing Dataset from the kaggle website the data is presented in CSV format and the file name of the data is Miami Housing Dataset (miami-housing.csv). The house data have the 13932 rows and 17 columns. For this dataset, I have to predict the price of the house depending on the various factors. The independent variables X includes latitude area of the house, longitude area of the house, parcel no., land squ. foot, total living area of the house, central distance, sub-center distance from the house, highway distance from the house, age, avno60plus, structure quality of the house, month sold. The dependent variable Y is the sale price which is the house price of the house.

3) Dataset 3: Retail dataset: This is very large Retail store dataset provided by our lecturer. A group of student working on this dataset all have to work on the same dataset. The data is from the UK chain of retail stores served as the dataset for this study. The data contains details on the store transactions. The dataset has already been examined in the past, and research has been done and recorded. The column name of the dataset is OperationName, RequestTimestamp, RequestSiteId, RequestSalePointId, RequestBasketIdTypeId, RequestBasketValue, ResponseProcessingTimestamp, RequestMessageId, RequestBasketId, ResponseMessageId, RequestTransmitAttempt, ResponseRgBasketId, ResponseCode, RequestNumberBasketItems, ResponseFinancialTimestamp, RequestBasketJsonString, LoyaltyId. We will to determine the recommended product using the Apriori algorithum and market basket analysis (MBA).

B. Data Preprocessing and Data Transformation

The Data Cleaning is the second steps in the KDD process. It is very important to clean the noise, null vlaues, missing values from the data set. If any column found the missing values or null values then it will create a problem for the model and the model is not become good and also not give the good accuracy. This process improves the data's dependability. There are several method is used to determine what are the unnecessary things in the data set. The panda's package in python is very helpful for the data preprocessing. When we do the data cleaning and transform our data into the proper way then we move to the next steps in the KDD.

1) Dataset 1: Salary Predictions Data : In the dataset before we going in the preprocessing process we have do some data analysis. First I checked what are the data types of the all features using the .dtypes() function. After that finding the unique values of all the features then find the unique value for all the object data types and found that the native-country column shows the very high unique values this create a problem for our model then I have created a function continents and segments the country name into the continents and make the new column region which store the continents data and reduce the unique values of native countries column and remove the native country column from the dataset. Furthermore, checking is there any missing value in the dataset unfortunately there is no missing values in the entire dataset. In our data set there are many categorical columns I have to transform that categorical column into the continuous column so that the calculation becomes easy because when we working with the regression model all the features must be in numerical for doing this work we use the Label Encoder () function which is again the come from the sklearn preprocessing package. It transformed all the categorical column in numerical values. After this checking the outliers of all the columns using box plot chart as shown on Fig 2, we found that outliers are present in the data set. After finding the outliers checking the correlation between the features by using the heatmap chart which comes form the seaborn library as shown in Fig 3.

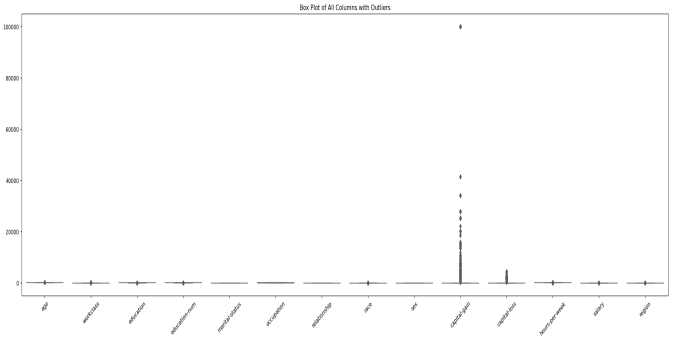

Fig. 2. Outliers in the Data

Now, separated the independent variables (x) which exclude the salary column and dependent variabel(Y) only include the

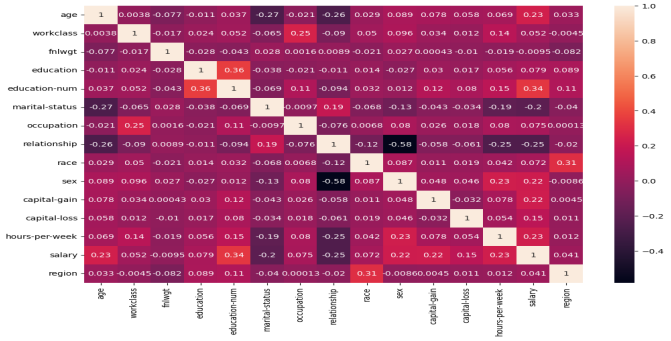

Fig. 3. Heatmap

salary column after this doing the train-test splitting by using the sklearn model selection package where the test size is 20 percent and the train size is 80 percent and the random state is 0. Moreover,, I found that the ouliers is present in the dataset so we have to remove the outliers from the training data set for this we use the z-scores method. In the z-score fixed the threshold value for our case the threshold value is 3 if the z-scores value is greater than the threshold we drop those outliers from the columns. At last, we perform the feature scaling on the training and testing data using the Standard Scaler function from the sklearn preprocessing package.

2) Dataset 2: Housing price Prediction Data: In the dataset before we going in the preprocessing process we have do some data analysis. First I checked is there is any null values in the dataset by using the .isnull() function and then checking the dtypes of the all features using the .dtypes() function. After that created a function (numerical_column_summary) which gives the outliers and the missing values of the numerical features and it give a complete summary of the numerical values which includes the min, max, IQR, mean, std, var, nmiss, cardinality and many more things which is help for the preprocessing. In the dataset I have founded that outliers are present as shown in Fig 4 but there is no missing values in the housing dataset. Furthermore, checking the correlation between the variabels and founded the correlation between some variables as shown on Fig 5. At this point divide the data into dependent (Y) and independent variables (X) and performing the train-test splitting using the sklearn library name as model selection. The test-size of the data is 30 percent and train test size is 70 percent, random state is 14. After this we perform the StandardScaler() which is used to transformed the data and make all the data in a similar scale for this again using the sklearn library.

3) Dataset 3: Retail dataset: Our lecture provided the retail transaction dataset, which has undergone extensive preprocessing and data transformation, preparing it for efficient exploratory data analysis (EDA). The dataset has been carefully prepared and altered to satisfy the required requirements for quality and usability. Association Rules used, it means if a person purchases A item then also buy B item. The important

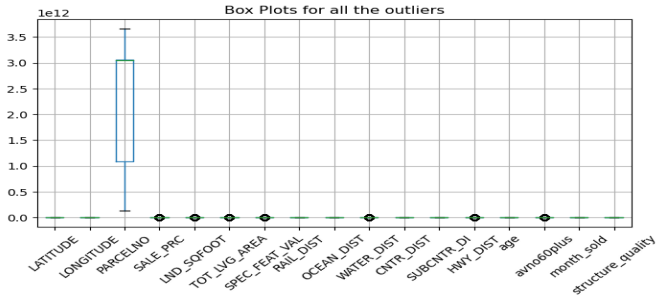

Fig. 4. Outliers in the House Data

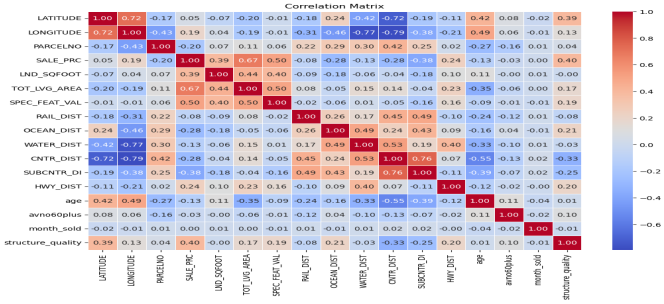

Fig. 5. Housing Price Correlation HeatMap

thing is when using the apriori algorithm the dataset must be in the list format we have the transaction data so we first convert it into a list format and all the items in the list must be string otherwise the apriori is not learned and trained. An analysis that is more ordered and cohesive has been made possible by the transactions' thoughtful association based on unique identifiers. All variables have been formatted according to appropriate standards in order to improve the dataset's uniformity and clarity. The robustness of the dataset is enhanced by these preprocessing and data transformation processes taken together, which also provide a strong foundation for insightful discoveries made during later phases of exploratory data analysis.

C. Data Mining

Data Mining refers to methods that are used to identify potentially helpful patterns. It turns the data into the pattern which is used to classify data. In KDD this is the fourth step process, using the mining process if any pattern is followed in the data, it will be easily found. That's why selecting the correct model parameters for the data is compulsory to optimize the process and the algorithm. All three of the data sets used in this study have been classified using machine learning techniques, I used the regression and classification machine learning model applied on two data sets and for the Retail dataset used the Apriori algortium to find the similar items using the association rules.

1) Dataset 1: Salary Predictions Data : Finding meaningful patterns and information in large datasets requires the use of data mining, a step in the Knowledge Discovery in Databases (KDD) process. Two supervised machine learning algorithms, DecisionTreeClassifier and GaussianNB, were used in the context of the Salary_data dataset to handle the classification task of determining whether or not an individual's salary exceeds 50K. Using a set of training data, the DecisionTreeClassifier is a potent tool that builds a tree-like model and makes decisions by recursively dividing the dataset into subsets. Applying the DecisionTreeClassifier specifically to the Salary_data, it divides the data according to pertinent features in order to categorize instances into pay groups. By contrast, the GaussianNB algorithm is a probabilistic classifier that uses the Gaussian distribution to generate predictions. A comparative analysis was carried out to assess the efficacy of these algorithms in classifying salaries. Dimensions like accuracy, F-1 score, precision, and recall were looked at to identify the best performing model for the task at hand. Deciding which algorithm is best for the classification problem at hand is made easier by having a thorough understanding of the subtle differences between GaussianNB and DecisionTreeClassifier in the context of the Salary_data dataset.

2) Dataset 2: Housing price Prediction Data: The task of predicting home prices was approached using the two supervised machine learning algorithms, KNeighborsRegressor and Linear Regression. Both algorithms belong to the category of supervised learning, in which predictions about unknown instances are made by the model after it has been trained on labeled data. Sale price was the dependent variable (Y) and the other variables (X) were the independent features of the miami-housingdata dataset. In order to fit a line that best represents the underlying pattern in the data, a linear relationship between the dependent and independent variables is the premise of the linear regression model., KNeighborsRegressor is a non-parametric algorithm that makes predictions about the target variable by averaging its k-nearest neighbors. Evaluating the complexity of both algorithms within the framework of the Miami-housing data dataset yields important information about how well-suited they are for predicting home values. In order to identify the best regression model, a comparative analysis based on metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared can be performed. This will help identify the algorithm that best captures the underlying patterns in the housing dataset.

3) Dataset 3: Retail Dataset: Using data mining techniques is critical to finding important patterns and associations in the Knowledge Discovery in Databases (KDD) process for retail data. There are two algorithum in association rule learning, Apriori and Eclat. The Apriori algorithm is one frequently used algorithm for this purpose it has 3 different parts Support, Lift, and confidence. An association rule mining technique called Apriori is used to find frequently occurring itemsets in a dataset. Apriori can be used to find product sets that are frequently bought together in the context of retail data. Through the process of sorting through transaction

data, the algorithm finds product combinations that satisfy a predetermined support threshold. These frequently occurring itemsets are then filtered to produce association rules, which show the connections and interdependencies between different products. Applying Apriori to the KDD process for retail data allows retailers to optimize their marketing strategies and make informed decisions by providing valuable insights into customer purchasing behaviors and preferences.

D. Pattern Evaluation

The Pattern Evaluation is used to Examine the patterns of the model based on the above measures of Linear Regression and the classification models this is a crucial stage in the Knowledge Discovery in Databases (KDD) process is the evaluation step, where the effectiveness of the created models is closely examined. Many parameters are necessary for a thorough assessment when it comes to regression models, like KNeighborsRegressor and Linear Regression. To measure the accuracy and goodness of fit of regression models, metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared are frequently used. The average squared difference between expected and actual values is measured by MSE and RMSE, which emphasize the accuracy of the model. R-squared, on the other hand, evaluates the percentage of the independent variables' variance that accounts for the variation in the dependent variable. Recall, accuracy, precision, and F1-score are important parameters in the context of classification models, such as DecisionTreeClassifier in GaussianNB. Preciseness and recall concentrate on the balance between false positives and false negatives, whereas accuracy assesses how accurate findings are overall. With a balanced evaluation, the F1-score is mean of recall and precision. Assuring a thorough grasp of the model's performance through the evaluation of these parameters helps in the KDD process by directing the choice of the best algorithm for the particular dataset and objective at hand.

1) Dataset 1: Salary Predictions Data : For the first dataset an accuracy score of 78.5% was obtained by the model-1 DecisionTreeClassifier, while the corresponding F-1, Precision, and Recall scores were all 0.785 as shown in fig 6. Conversely, the F-1, Precision, and Recall scores of the second model which is GaussianNB model were all 0.803, indicating a higher accuracy score of 80.38% as shown in Fig 7. These metrics give each model's performance a numerical understanding.

2) Dataset 2: Housing price Prediction Data: Two regression models were used to predict house prices in the evaluation of the House dataset as part of the Knowledge Discovery in Databases (KDD) process: Linear Regression and KNeighborsRegressor. With an Explained Variance Score of 0.73, Mean Absolute Error of 101877.32, R-squared Score of 0.73 as shown in Fig 8, Mean Squared Error of 21815636646.26, and Root Mean Squared Error (RMSE) of 147701.17, the Linear Regression model was found to be effective. However, with an Explained Variance Score of 0.88, Mean Absolute Error of 50243.21, R-squared Score of 0.88, Mean Squared Error of 9408039505.76, and an RMSE of 96995.04, the

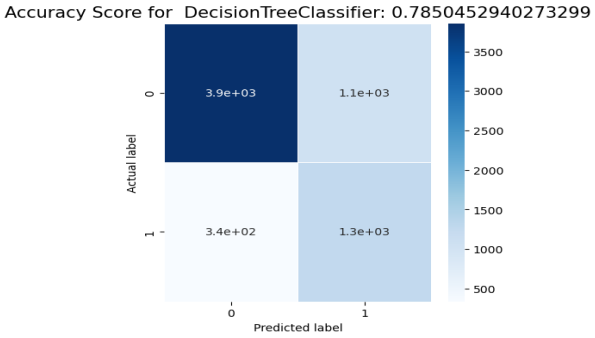

Fig. 6. Confusion Matrix for DecisionTreeClassifier

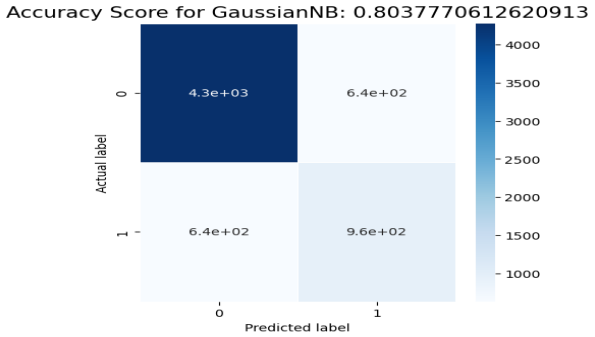

Fig. 7. Confusion Matrix for GaussianNB

KNeighborsRegressor model outperformed the others. These metrics for accuracy offer a thorough assessment of the models' capacity to forecast home prices.

OLS Regression Results

Dep. Variable:	SALE_PRC	R-squared:	0.736
Model:	OLS	Adj. R-squared:	0.735
Method:	Least Squares	F-statistic:	2583.
Date:	Wed, 22 Nov 2023	Prob (F-statistic):	0.00
Time:	16:30:35	Log-Likelihood:	-1.8605e+05
No. Observations:	13932	AIC:	3.721e+05
Df Residuals:	13916	BIC:	3.723e+05
Df Model:	15		
Covariance Type:	nonrobust		

Fig. 8. OLS Models 2

3) Dataset 3: Retail Dataset: The purpose of this experiment is to ascertain how well the selected algorithms perform when the quantity of input transactions varies. The previous authors' work on the execution times of each algorithm will serve as the basis for this comparison. This experiment will assist in determining which algorithm, when applied to the selected dataset, executes more quickly. Later on, the outcomes of this experiment can be compared to the outcomes of the suggested methodology in the following experiment.

E. Knowledge Representation

It is the final results that are helpful for the users to make the decision based on the accuracy of the metrics. The knowledge

representation for the Salary Prediction dataset might entail producing reports and visualizations that convey the conclusions drawn from the GaussianNB and DecisionTreeClassifier models. The models' performance in salary classification could be demonstrated by metrics such as precision-recall curves, or confusion matrices. GaussianNB is a better fit for the salary prediction dataset than DecisionTreeClassifier, according to the comparison of these two models' accuracy scores. Given that it more closely matches the patterns found in the data, GaussianNB appears to offer a better overall prediction performance, as indicated by the higher accuracy score.

Secondly the knowledge representation phase for the House dataset would involve presenting the findings and understandings from the KNeighborsRegressor and Linear Regression models. scatter plots, which contrast projected and actual home prices, are examples of visualizations. Recommendations based on the regression models would be made after analyzing the features' importance in predicting home prices. Encouraging stakeholders to make well-informed decisions by providing them with relevant and useful information is the overall goal of the knowledge representation step. The accuracy metrics mentioned above offer a thorough assessment of the models' capacity to forecast house prices. By outperforming the Linear Regression model on every metric, the KNeighborsRegressor model demonstrates its superior predictive power and accuracy. This model is the best choice for predicting house prices in this particular context because of its higher Explained Variance Score and lower errors, especially the RMSE, which show how well the KNeighborsRegressor model captures the underlying patterns in the house dataset. For the retail transaction data using the Apriori algorithm and Association rules recommended the product based on the previous transaction done by the user.

IV. Future work

In the context of the Salary Prediction dataset, future research could investigate the application of more sophisticated machine learning techniques, like neural networks or ensemble methods, to see if they can attain even higher predictive accuracy. Furthermore, a deeper comprehension of the variables affecting wage forecasts might result from adding more varied features or investigating feature engineering techniques. To guarantee equitable results across various demographic groups, the evaluation of model fairness and bias detection mechanisms could be combined. In order to potentially enhance predictive modeling for the House dataset, future research projects might investigate different regression algorithms or try various approaches to parameter tuning. For the retail transaction dataset, future work is to use the Eclat algorithm which is another algorithm in the assiciation rule.

References

[1] R. Ranggasari, "Indonesia's Inflation Rate at 1.17% September; Highest in 94 Months," Tempo, Oct. 03, 2022. https://en.tempo.co/read/1641075/indonesias-inflationrate-at-1-17-september-highest-in-94-months.

[2] R. Keely and R. A. Lyons, "Housing Prices, Yields and Credit Conditions in Dublin since 1945," The Journal of Real Estate Finance and Economics, Aug. 31, 2020. https://doi.org/10.1007/s11146-020-09788-z.

[3] A. Sahni, "How To Boost Retail Sales Using an Automated Recommender System." https://blog.ncirl.ie/how-to-boost-retail-sales-with-an-automated-recommender-system.

[4] Y. Chen, "Salary Prediction Based on the Dual-Adaboosting System," EUDL, Jul. 21, 2023. https://eudl.eu/doi/10.4108/eai.26-5-2023.2334428.

[5] R. Kablaoui and A. Salman, "Machine Learning Models for Salary Prediction Dataset using Python," 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, 2022, pp. 143-147, doi: 10.1109/ICECTA57148.2022.9990316.

[6] J. V. Siswanto, L. A. Castilani, N. H. Winata, N. C. Nugraha and N. T. M. Sagala, "Salary Classification Prediction based on Job Field and Location using Ensemble Methods," 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), Jakarta, Indonesia, 2023, pp. 325-330, doi: 10.1109/ICCoSITE57641.2023.10127828.

[7] G. Wang, "Employee Salaries Analysis and Prediction with Machine Learning," 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), Guangzhou, China, 2022, pp. 373-378, doi: 10.1109/MLISE57402.2022.00081.

[8] S. Das, R. Barik, and A. Mukherjee, "Salary Prediction Using Regression Techniques," Social Science Research Network, Jan. 01, 2020. https://doi.org/10.2139/ssrn.3526707.

[9] X. Wu and B. Yang, "Ensemble Learning Based Models for House Price Prediction, Case Study: Miami, U.S," 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 2022, pp. 449-458, doi: 10.1109/AEMCSE55572.2022.00095.

[10] Q. D. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing Price Prediction via Improved Machine Learning Techniques," Procedia Computer Science, Jan. 01, 2020. https://doi.org/10.1016/j.procs.2020.06.111.

[11] Y. Chen, R. Xue and Y. Zhang, "House price prediction based on machine learning and deep learning methods," 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS), Changchun, China, 2021, pp. 699-702, doi: 10.1109/EIECS53707.2021.9587907.

[12] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique," Procedia Computer Science, Jan. 01, 2022. https://doi.org/10.1016/j.procs.2022.01.100.

[13] S. Lu, Z. Li, Z. Qin, X. Yang and R. S. M. Goh, "A hybrid regression technique for house prices prediction," 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, 2017, pp. 319-323, doi: 10.1109/IEEM.2017.8289904.

[14] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, NSW, Australia, 2018, pp. 35-42, doi: 10.1109/iCMLDE.2018.00017.

[15] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," Expert Systems with Applications, Apr. 01, 2015. https://doi.org/10.1016/j.eswa.2014.11.040.

[16] K. Tatiana and M. Mikhail, "Market basket analysis of heterogeneous data sources for recommendation system improvement," Procedia Computer Science, Jan. 01, 2018. https://doi.org/10.1016/j.procs.2018.08.263.

[17] J. Dongre, G. L. Prajapati and S. V. Tokekar, "The role of Apriori algorithm for finding the association rules in Data mining," 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 2014, pp. 657-660, doi: 10.1109/ICICICT.2014.6781357.

[18] W. B. Zulfikar, A. Wahana, W. Uriawan and N. Lukman, "Implementation of association rules with apriori algorithm for increasing the quality of promotion," 2016 4th International Conference on Cyber and IT Service Management, Bandung, Indonesia, 2016, pp. 1-5, doi: 10.1109/CITSM.2016.7577586.

[19] "KDD Process in Data Mining - Javatpoint," www.javatpoint.com. https://www.javatpoint.com/kdd-process-in-data-mining.