Unraveling Success: A Predictive Analytics Exploration of Income, Financial Choices, and Housing Dynamics

Divyansh Anand
School of Computing
National College of Ireland Dublin,
Ireland
x22240217@student.ncirl.ie

***Abstract*— In the pursuit of success in our competitive world, this research undertakes a comprehensive analysis of interconnected success indicators, encompassing income, financial choices, and housing dynamics. Focusing on three datasets, the study employs predictive analytics and machine learning to explore these complex relationships. The first dataset predicts an individual's income surpassing a threshold, emphasizing education's impact. The second dataset delves into factors influencing the decision to open a term deposit. The third dataset predicts house prices, recognizing the significance of shelter in one's financial journey. The research poses questions on education, income disparities, and housing dynamics, employing Decision Trees and Random Forests for the first dataset, Support Vector Machine and Logistic Regression for the second, and Linear Regression for the third. Evaluation metrics, including accuracy, RMSE, RSS, precision, recall, and F-measure, reveal the models' performance. The study offers insights into education's role, gender-based income disparities, and the correlation between apartment size and price, contributing to a holistic understanding of success indicators.**

Keywords*— *Predictive Analytics, Machine Learning, Income Prediction, Decision Trees, Random Forests, Support Vector Machine, Logistic Regression, Linear Regression.

I. INTRODUCTION

In our competitive world, individuals relentlessly strive for success, measured across multiple dimensions such as income, education, bank account balance, and property ownership. This research is dedicated to analyze and explore the complex relationships within these parameters, with a primary focus on predicting whether an individual's income exceeds certain threshold amount in the first dataset. As individuals embark on their journey of financial stability, saving becomes a important aspect, often realized through the secure avenue of opening a term deposit in a bank account. The second dataset explores the determinants influencing the decision to open a term deposit, shedding light on the diverse inputs that guide this financial choice. Subsequently, as individuals progress in their earning potential and personal lives, considerations such as marriage and homeownership come to the forefront. The third dataset delves into predicting house prices, recognizing the fundamental need for shelter and addressing the factors influencing this crucial aspect of one's financial journey. This research aims to explore the complex landscape of success indicators, offering perspectives on the interconnected domains of income, financial choices, and housing dynamics. This exploration is conducted through the perspective of predictive analytics and machine learning methodologies.

A. Research questions

Question 1: How does the level of education impact the probability of an individual opening a term deposit?

Question 2: What factors contribute to the observed disparity in income between males and females?

Question 3: How do interactions between variables such as size of an apartment affect the price of an apartment?

II. RELATED WORK

The literature review from the first paper [1] reveals a diverse landscape of income prediction studies on the UCI Adult Dataset, employing various machine learning models such as Logistic Regression, Naive Bayes, Decision Trees, and Gradient Boosting. While these studies offer comprehensive analyses, the research question on the relationship between education levels and term deposit likelihood remains underexplored. Existing works exhibit a lack of emphasis on education, and there's a need for a focused investigation into this specific aspect. The decision to reuse Decision Tree and Random Forest methods aligns with the interpretability and feature importance analysis required for the research question, building upon the strengths of these models observed in prior studies. The UCI Adult Dataset's popularity stems from its rich demographic and employment features, providing a robust foundation for exploring the nuanced relationship between education levels and financial decisions, specifically the probability of opening a term deposit. [1]

The related work extensively explores the integration of association rules with decision trees in object-relational databases, emphasizing practical implementation within an Oracle environment. While it positively aligns with the classification objective, specifically through models like CBA-ODM1 and CBA-ODM2, it falls short in directly addressing the nuanced context of predicting term deposit likelihood. In terms of datasets and methods, the research omits explicit mention of the UCI Adult Dataset but utilizes 18 datasets from UCI for comparative evaluations. The research, focusing on the impact of education on term deposit probability, distinguishes itself by reusing Decision Tree and Random Forest methods, ensuring alignment with state-of-the-art practices and the Oracle environment. These methods promise enhanced model interpretability, crucial for stakeholders, and a robust feature importance analysis, offering nuanced insights into the significance of education levels. While the related work lays a foundation, the research extends this understanding to the specific realm of the relationship between education levels and term deposit likelihood, emphasizing the applicability of Decision Trees and Random Forests. [2]

The thesis provides a thorough examination of predicting bank term deposits through various machine learning models,

specifically investigating the influence of education levels on individuals' propensity to open term deposits. The inclusion of diverse models, such as binomial logistic regression, decision tree classifier, artificial neural networks, and support vector machines, reflects a meticulous consideration of their strengths and weaknesses in handling classification tasks. A limitation, however, lies in the relatively brief exploration of related work, especially in the context of education's impact on deposit decisions. Additionally, the critical evaluation of key related work is somewhat lacking, as a more detailed analysis of the strengths and weaknesses of existing research in this domain could have provided valuable insights into the project's positioning within the broader field of study. Despite this, the thesis demonstrates awareness of potential biases through the application of resampling techniques to address dataset imbalances. Furthermore, the discussion on dataset conditions evolving over time acknowledges the relevance of findings in different macroeconomic contexts, contributing to the adaptability of the selected support vector machine model. [3]

The related work in the fair regression paper, focusing on fairness in machine learning algorithms, offers a valuable foundation but lacks direct relevance to the investigation of education's impact on term deposit likelihood. Positive aspects include emphasizing fairness in algorithmic predictions and systematic approaches using supervised learning oracles. However, a limitation is the absence of a direct connection to the specific research question. While the fair regression paper doesn't detail datasets and methods, in the context of education and deposits, relevant datasets likely involve financial and demographic information. Reusing logistic regression, a common method in predictive modeling, could offer insights into the nuanced relationship between education levels and the likelihood of opening a term deposit, leveraging established fairness metrics. [4]

This groundbreaking research navigates the intricate landscape of gender disparity in income by innovatively incorporating machine learning into the traditional economic theories used to understand the gender pay gap (GPG). Conducted on a sample of 5,742 Argentinean IT-related workers, the study exposes an overall GPG of 20%, shedding light on both direct discrimination (7.7%) and other factors like experience, education, and age (12.3%). Leveraging XGBoost, the study overcomes challenges posed by skewed wage distribution, offering a systematic and iterative approach to GPG estimation. [5]

The paper investigates saving behavior, emphasizing services quality, religious belief, knowledge, and incorporates the use of the XGBoost machine learning algorithm for empirical insights. While providing valuable empirical insights, limitations such as a small sample size and a narrow focus on private sector employees may affect generalizability. Moreover, the absence of details on the dataset's previous uses, including the application of XGBoost, raises questions about its reliability. For future research on education's impact on deposit behavior, refining the dataset and leveraging the original study's methodologies, particularly regression analysis and the use of XGBoost, are crucial steps. [6]

The paper contributes to the understanding of the gender pay gap (GPG) in Germany by employing a machine learning approach, specifically the post-double-LASSO procedure, to estimate the adjusted GPG. It successfully highlights the significant variations in the estimated gap when compared to conventional models, emphasizing the necessity for more flexible specifications in capturing gender differences in pay. The introduction succinctly positions the research within the context of the persistent GPG in Germany, aligned with global concerns. Utilizing the German Socio-Economic Panel Study (SOEP) lends credibility to the study, harnessing a large and nationally representative dataset. While the innovative methodological approach opens avenues for future research, the paper conscientiously acknowledges potential limitations associated with machine learning. [7]

In this study, a robust decision support system (DSS) using a data mining (DM) approach is proposed to predict the success of telemarketing calls in Portuguese banks post the 2008 financial crisis. Through a semi-automated feature selection process, the study identifies 22 relevant attributes. Four data mining models—logistic regression, decision trees, neural networks, and support vector machines—are rigorously compared using metrics like the area under the receiver operating characteristic curve (AUC) and the area under the LIFT cumulative curve (ALIFT). While the study demonstrates methodological rigor and practical significance by improving campaign efficiency, the potential trade-off between model complexity and interpretability is acknowledged. [8]

This research, building upon existing studies aims to advance housing price prediction methodologies with a specific focus on the apartment dataset. Previous research has offered valuable insights into housing market dynamics, but limitations, such as overlooking nuanced variable interactions, persist. While [specific strengths] characterize positive aspects of prior work, the identified gaps present an opportunity for improvement in understanding the complex relationships within datasets, including the role of size, bedrooms, bathrooms, and city names. Notably, this study employs the (KNN) method to address these gaps and explore how the collective influence of size, bedrooms, bathrooms, and city names affects apartment prices. By critically assessing and potentially enhancing existing methodologies for variable interaction analysis. [9]

This research navigates the intersection of fairness in ML algorithms, particularly in high-impact domains. Current fairness-aware approaches face challenges in balancing accuracy and fairness in class-imbalanced datasets. AdaFair, an AdaBoost-based classifier, is introduced to explicitly address fairness and class imbalance concerns. The research acknowledges the positive strides in fairness, yet emphasizes the need for nuanced consideration, especially in imbalanced datasets. The decision to use AdaBoost reflects its potential insights into housing price dynamics. The study probes the interactions of apartment size, bedrooms, bathrooms, and city names on prices, shedding light on fairness concerns in this context. The work identifies an opportunity for improvement in addressing class imbalance challenges and leveraging AdaBoost's unique strengths in housing price prediction. [10]

In this innovative ensemble learning paper, a locally optimal tree of predictors is developed, constructing an overall predictor through weighted averages along specific paths. The approach adeptly aligns learners and training sets with dataset characteristics, preventing overfitting. The rigorous proofs in Appendices A and B establish bounds for predictive models, providing a robust theoretical foundation. The computational complexity analysis in Appendix C underscores the efficiency of the proposed method. Building upon related work on BAGGING and BOOSTING, the paper critically evaluates their contributions, highlighting the need for adaptability to diverse datasets. The choice of ensemble methods, particularly bagging and boosting, addresses limitations in previous linear regression models applied to apartment pricing datasets. [11]

The research delves into the intricate dynamics of apartment pricing, examining the influence of variables like size, bedrooms, bathrooms, and city names. Drawing inspiration from the success of ML methods, particularly Random Forests and Neural Networks, in a retail banking paper, the study aims to reveal nuanced relationships affecting housing costs. Previous works, rooted in statistical and ML approaches, highlight the significance of identifying critical features for price determination. Despite successes, limitations, such as covariate analysis gaps and dataset biases, have been acknowledged. Leveraging ML methods, especially Neural Networks, anticipates a high predictive accuracy for understanding variable interactions. Reusing ML approaches from analogous datasets holds the potential for improved predictions and the identification of pivotal variables. [12]

This research, inspired by a thesis on bank marketing, delves into the relationships between apartment features and city names affecting prices. Noteworthy machine learning methods, such as Logistic Regression, Support Vector Machine, and K-Nearest Neighbor, are employed. The thesis efficiently addresses imbalanced datasets using resampling techniques. Positive aspects include versatile ML algorithms, though real estate specifics are not deeply explored. The dataset's previous applications emphasize understanding customer behavior, aligning with predicting apartment prices. Leveraging high-performing ML classifiers, particularly KNN, with necessary adaptations, provides a foundational understanding for exploring real estate price determinants. [13]

III. Data Mining Methodology

Data mining includes various aspects which includes extracting knowledge from data, analysing data pattern, uncovering historical data and analysing data with multiple tests or models until statistically significant results are found. It is also referred as Knowledge Discovery in Databases, it is the nontrivial extraction of implicit, previously unknown, and potentially relevant information from data stored in databases.

Figure 1 shows the process of KDD methodology [14]-

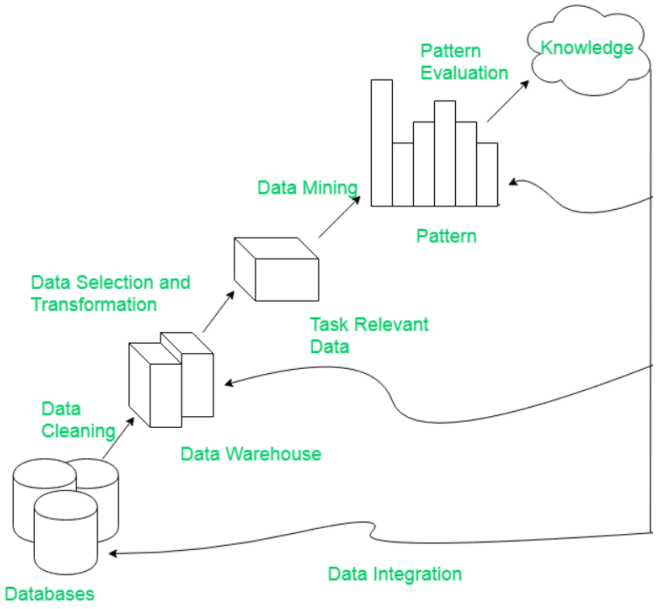

Figure 1: Knowledge Discovery in Database Methodology

Following below are the steps of the KDD process and its application taken in this research.

A. *Datasets Selection*

This is the first step of KDD which consists of developing final dataset from the raw data available in the sources. This step emphasis on attribute collection and sampling of data to minimize the records for optimal use in the further steps of KDD.

1) Adult/ Income Dataset

The Adult dataset, also called as the Census Income dataset, is a widely used standard machine learning dataset. It contains demographic information from the 1994 US Census and is used to predict whether an individual's income exceeds $50,000 per year based on various features. The dataset contains 15 attributes, including some important attributes such as age, work class, education, marital status, occupation, sex, native country, and income level. The features include both categorical and integer types, and the dataset consists of 32,562 instances. The dataset is used to explore and demonstrate machine learning algorithms, particularly those designed for classification. [15] Below figure 2 illustrates the summary of the loaded Adult/Income dataset.

```
      age            workclass          fnlwgt          education
 Min.   :17.00   Length:32561       Min.   :  12285   Length:32561
 1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character
 Median :37.00   Mode  :character   Median : 178356   Mode  :character
 Mean   :38.58                      Mean   : 189778
 3rd Qu.:48.00                      3rd Qu.: 237051
 Max.   :90.00                      Max.   :1484705
 education.num   marital.status     occupation        relationship
 Min.   : 1.00   Length:32561       Length:32561       Length:32561
 1st Qu.: 9.00   Class :character   Class :character   Class :character
 Median :10.00   Mode  :character   Mode  :character   Mode  :character
 Mean   :10.08
 3rd Qu.:12.00
 Max.   :16.00
    race              sex            capital.gain    capital.loss
 Length:32561       Length:32561       Min.   :    0   Min.   :   0.0
 Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0
 Mode  :character   Mode  :character   Median :    0   Median :   0.0
                                       Mean   : 1078   Mean   :  87.3
                                       3rd Qu.:    0   3rd Qu.:   0.0
                                       Max.   :99999   Max.   :4356.0
 hours.per.week  native.country     income
 Min.   : 1.00   Length:32561       Length:32561
 1st Qu.:40.00   Class :character   Class :character
 Median :40.00   Mode  :character   Mode  :character
 Mean   :40.44
 3rd Qu.:45.00
 Max.   :99.00
```

Figure 2. Summary of Adult dataset

2) Bank Marketing Dataset

The dataset pertains to the direct marketing campaigns conducted by a Portuguese banking institution, with the objective of forecasting whether a client will opt for a term deposit. Comprising 45,211 instances and encompassing 16 features, including categorical and integer attributes. The features describe the client's demographic information, such as age, job type, marital status, and education, as well as information about the contact, such as the communication type and the month of the last contact. The target variable indicates whether the client subscribed a term deposit (yes or no). [16] Summary of the dataset after loading it is given in figure 3.

```
      age            job              marital          education
 Min.   :18.00   Length:11162       Length:11162       Length:11162
 1st Qu.:32.00   Class :character   Class :character   Class :character
 Median :39.00   Mode  :character   Mode  :character   Mode  :character
 Mean   :41.23
 3rd Qu.:49.00
 Max.   :95.00
   default            balance        housing            loan
 Length:11162       Min.   :-6847   Length:11162       Length:11162
 Class :character   1st Qu.:  122   Class :character   Class :character
 Mode  :character   Median :  550   Mode  :character   Mode  :character
                    Mean   : 1529
                    3rd Qu.: 1708
                    Max.   :81204
   contact              day            month              duration
 Length:11162       Min.   : 1.00   Length:11162       Min.   :   2
 Class :character   1st Qu.: 8.00   Class :character   1st Qu.: 138
 Mode  :character   Median :15.00   Mode  :character   Median : 255
                    Mean   :15.66                      Mean   : 372
                    3rd Qu.:22.00                      3rd Qu.: 496
                    Max.   :31.00                      Max.   :3881
    campaign          pdays            previous         poutcome
 Min.   : 1.000   Min.   : -1.00   Min.   : 0.0000   Length:11162
 1st Qu.: 1.000   1st Qu.: -1.00   1st Qu.: 0.0000   Class :character
 Median : 2.000   Median : -1.00   Median : 0.0000   Mode  :character
 Mean   : 2.508   Mean   : 51.33   Mean   : 0.8326
 3rd Qu.: 3.000   3rd Qu.: 20.75   3rd Qu.: 1.0000
 Max.   :63.000   Max.   :854.00   Max.   :58.0000
   deposit
 Length:11162
 Class :character
 Mode  :character
```

Figure 3. Summary of Bank Marketing dataset

3) Apartment/House rent Dataset

The Apartment for Rent Classified dataset is a multivariate dataset containing 10,000 instances of classified advertisements for apartments for rent in the USA. The dataset has 22 features, including attributes related to the apartment's price, size, location, and amenities, as well as the text of the advertisement. The features include both categorical and integer types, dataset can be used for various machine learning tasks such as clustering, classification, and regression.[17] This research has performed linear regression on this dataset. Summary for the dataset is provided in below figure 4.

```
       id                 Run all          title              body
 Min.   :5.509e+09   Length:10000       Length:10000       Length:10000
 1st Qu.:5.509e+09   Class :character   Class :character   Class :character
 Median :5.669e+09   Mode  :character   Mode  :character   Mode  :character
 Mean   :5.623e+09
 3rd Qu.:5.669e+09
 Max.   :5.669e+09
  amenities          bathrooms          bedrooms           currency
 Length:10000       Length:10000       Length:10000       Length:10000
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
     fee             has_photo          pets_allowed          price
 Length:10000       Length:10000       Length:10000       Min.   :  200
 Class :character   Class :character   Class :character   1st Qu.:  949
 Mode  :character   Mode  :character   Mode  :character   Median : 1270
                                                          Mean   : 1486
                                                          3rd Qu.: 1695
                                                          Max.   :52500
 price_display       price_type         square_feet          address
 Length:10000       Length:10000       Min.   :  101.0    Length:10000
 Class :character   Class :character   1st Qu.:  649.0    Class :character
 Mode  :character   Mode  :character   Median :  802.0    Mode  :character
                                       Mean   :  945.8
                                       3rd Qu.: 1100.0
                                       Max.   :40000.0
   cityname            state              latitude         longitude
 Length:10000       Length:10000       Length:10000       Length:10000
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
    source              time
 Length:10000       Min.   :1.569e+09
 Class :character   1st Qu.:1.569e+09
 Mode  :character   Median :1.577e+09
```

Figure 4. Summary of Apartment Rent Dataset

B. Data Preprocessing

This second stage of KDD involves cleaning and preprocessing the dataset. It involves activities like checking and handling null values, checking and handling outliers in the columns if necessary and cleaning the other data anomalies. This is a crucial step as unprocessed data can lead to poor models and may give incorrect accuracy.

1) Adult/ Income Dataset

Below steps were taken to check and handle anomalies in Adult/Income dataset:

a) Missing value check

As shown in below figure 5, there are no missing values in the columns.

```
missing_values <- colSums(is.na(df))

print(missing_values)
```

```
          age      workclass         fnlwgt      education  education.num
            0              0              0              0              0
 marital.status     occupation   relationship           race            sex
            0              0              0              0              0
   capital.gain   capital.loss hours.per.week native.country         income
            0              0              0              0              0
```

Figure 5. Checking for missing value

b) Outliers check

This research has checked the outliers for the non-categorical columns age, capital gain/loss and hours per week. Outlier values was removed from hours column and printed as there were very less people working for more than 80 hours, please find this in the below figure 6 and figure 7.

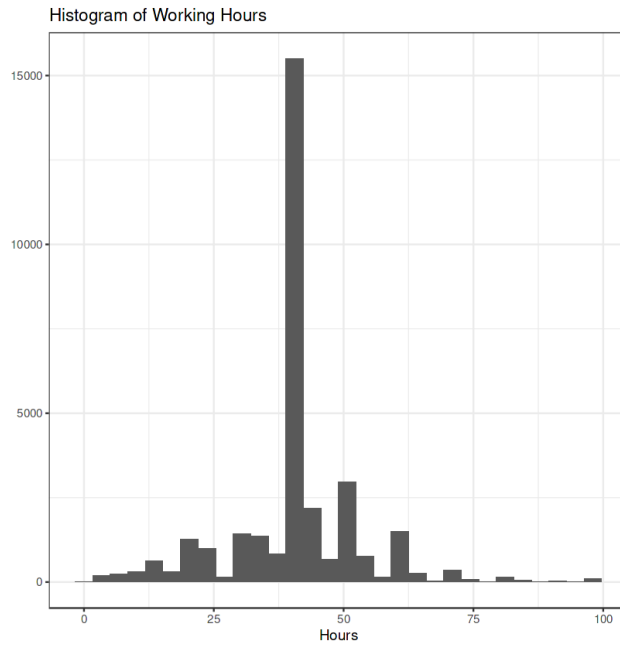

Figure 6. Before removing outlier from hours column

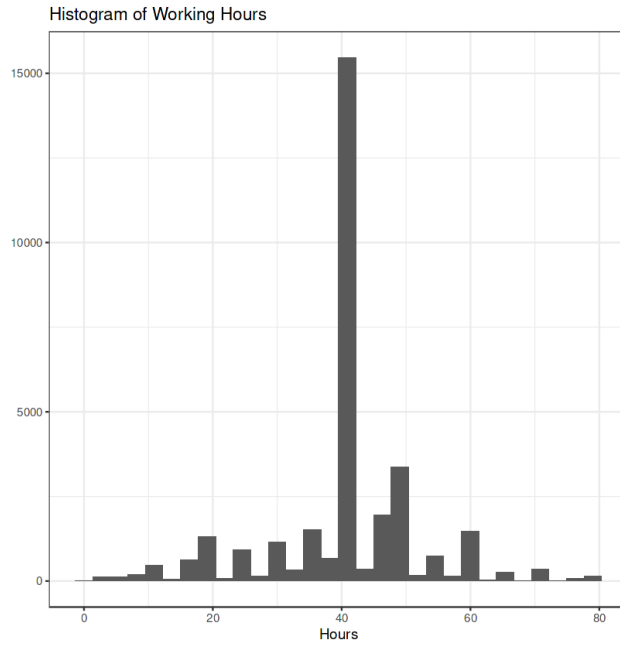

Figure 7. After removing outlier from hours column (max 80 hours)

2) Bank Marketing Dataset

Below steps were taken to handle anomalies in Bank Marketing dataset:

a) Missing value check

```
# Handle missing
df[df == "null"] <- NA

missing_values <- sapply(df, function(x) sum(is.na(x)))

print(missing_values)
```

```
      job   marital education   default   balance   housing      loan  duration
        0         0         0         0         0         0         0         0
 campaign     pdays  previous poutcome   deposit
        0         0         0         0         0
```

Figure 8. Checking for missing values

As depicted from figure 8, there are no missing values in the columns.

b) Outliers check

This research has checked for outliers in the columns balance, duration, campaign, pdays and previous. Out of which highest outlier values in columns balance and duration columns as given in figure 9.

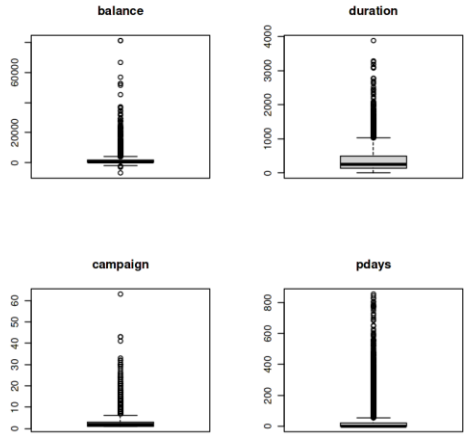

Figure 9. Before removing outliers from balance and duration column

After removing outliers from columns balance and duration column can be observed in figure 10.

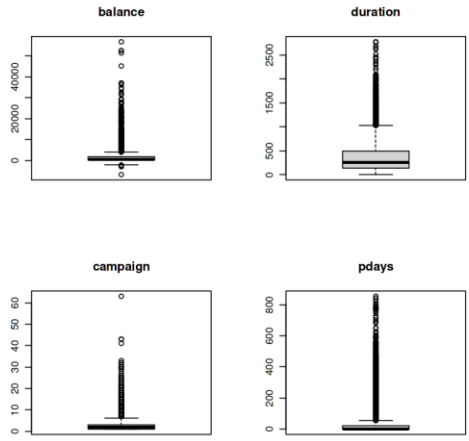

Figure 10. After removing outliers from balance and duration columns

3) Apartment/ House rent Dataset

Below steps were taken to handle anomalies in Apartment/ House rent dataset:

a) Missing value check

Missing values count in the dataset has been displayed in the below figure 11.

```
# Handle missing values
df[df == "null"] <- NA

missing_values <- sapply(df, function(x) sum(is.na(x))

print(missing_values)
```

```
bathrooms    bedrooms square_feet    cityname       price
       34           7           0          77           0
```

Figure 11. Missing values in the Apartment rent dataset

Above displayed columns are used for further analysis and missing values were removed from the dataframe as shown in figure 10.

```
df <- na.omit(df)
missing_values <- sapply(df, function(x) sum(is.na(x))

print(missing_values)
```

```
bathrooms    bedrooms square_feet    cityname       price
        0           0           0          0           0
```

Figure 12. Missing values in the Apartment rent dataset (After removing)

b) Outlier check

Outliers for Apartment rent dataset has been checked for columns number of bathrooms, bedrooms and size (in square feet) of the apartment/house. Out of these only column apartment size (square feet) has shown significant outlier values as given in figure 11.

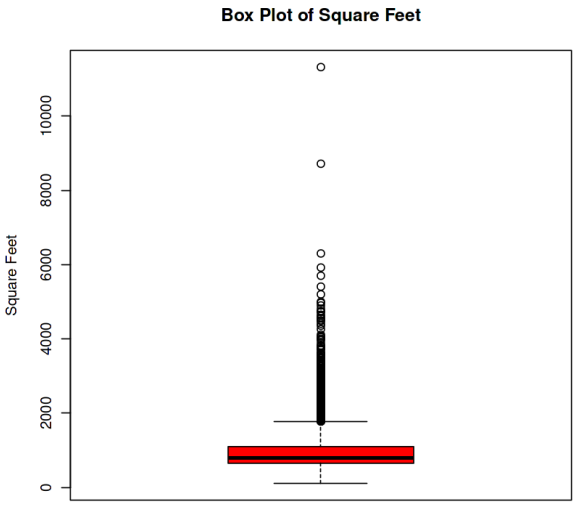

Figure 13. Box plot of apartment size (square feet) column (Before)

Below is the figure 14, after removing the outliers above

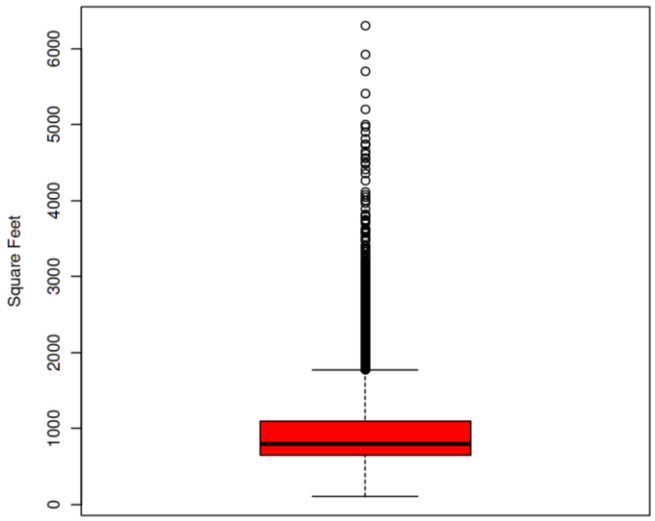

Figure 14. Box plot of apartment size (square feet) column (After)

C. *Data Transformations*

After preprocessing of data in all the three datasets, data transformation phase which is the next phase of KDD methodology comes into picture. In this phase, different operations like data aggregation, feature engineering and dimension reduction is worked upon.

This research has performed below functions/steps on the datasets:

1) Adult/ Income Dataset

For this dataset this research has handled all the categorical columns and converted them into integer variables. This helps in preparing data for machine learning algorithms like Decision tree and random forest which will be further applied for this dataset. Additionally, workclass column values has been replaced by index values for proper handling of this column given in figure 15.

	age	workclass	fnlwgt	education	education.num	Marital	occupation	relationship	race	sex	capital.gain	capital.loss	Hours	Country	inc
	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<
1	90	1	77053	12	9	7	1	2	5	1	0	4356	40	40	
2	82	2	132870	12	9	7	5	2	5	1	0	4356	18	40	
3	66	1	186061	16	10	7	1	5	3	1	0	4356	40	40	
4	54	2	140359	6	4	1	8	5	5	1	0	3900	40	40	
5	41	2	264663	16	10	6	11	4	5	1	0	3900	40	40	
6	34	2	216864	12	9	1	9	5	5	1	0	3770	45	40	

Figure 15. After transformation of work-class column

2) Bank Marketing Dataset

For this dataset one-hot encoding has been applied to categorical columns of the dataset. One-hot encoding is a technique used to convert categorical variables into binary vectors, making them suitable for machine learning algorithms that require numerical input given in figure 16.

jobentrepreneur	jobhousemaid	jobmanagement	...	educationsecondary	educationtertiary	educationunknown	defaultyes	housingyes	loanyes	pou
<dbl>	<dbl>	<dbl>	...	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
0	0	0	...	1	0	0	0	1	0	
0	0	0	...	1	0	0	0	0	0	
0	0	0	...	1	0	0	0	1	0	
0	0	0	...	1	0	0	0	1	0	
0	0	0	...	0	1	0	0	0	0	
0	0	1	...	0	1	0	0	1	1	

Figure 16. Snapshot after one-hot encoding

3) Apartment/ House rent Dataset

For this dataset code, code has dropped other columns keeping important columns such as bathrooms, bedrooms, square feet (size of apartment) in the initial phase of the research itself.

Research ensures that the "bathrooms" and "bedrooms" columns are treated as numeric variables in the data frame. It's common to perform these conversions when dealing with data that might have been read in as factors or characters but needs to be treated as numeric for numerical analyses or modeling purposes.

During the analysis, research removes rows from the data frame where the "cityname" column has missing values.

There is one feature engineering step involved which helps in creating a new feature ("price_per_sqft") by calculating the price per square foot using existing columns ("price" and "square_feet") in the dataset it is shown in below figure 17.

Also, the research removes rows from the data frame where the "cityname" column has missing values.

Lastly, one-hot encoding has been applied to cityname column keeping top 10 cities and labelling remaining city names as “other”.

	bathrooms	bedrooms	square_feet	cityname	price	price_per_sqft
	<dbl>	<dbl>	<int>	<chr>	<int>	<dbl>
3	1	0	107	Arlington	1390	1299065.4
4	1	0	116	Seattle	925	797413.8
6	1	0	130	Manhattan	2475	1903846.2
9	1	0	138	San Francisco	1495	1083333.3
15	1	0	190	San Francisco	1695	892105.3
16	1	1	200	New Bern	1560	780000.0

Figure 17. After feature engineering step(adding price per sq feet)

D. Data Mining

Data mining is the fourth phase of KDD methodology, which consist of choosing necessary data mining algorithms for the chosen datasets. Data mining algorithm extracts patterns/insights from the transformed data. For optimum algorithm operation relevant parameter settings would need to be determined.

1) Adult/Income Dataset

Initially for data preparation, R code in this research utilizes the 'caret' package to facilitate the data mining process, specifically focusing on the preparation of a dataset for model training and evaluation. To ensure reproducibility, a seed is set using set.seed(1000).

The "income" column is explicitly converted to a factor variable to handle categorical data appropriately.

Subsequently, the dataset is partitioned into training and testing sets with a 70-30 split using the createDataPartition function from 'caret'. The training set ('train') comprises 70% of the data, while the testing set ('test') contains the remaining 30%. This partitioning is a crucial step in building and assessing machine learning models, allowing for independent training and evaluation phases to ensure the robustness and generalizability of the models.

Out of all the classification algorithms namely Logistic Regression, Decision trees, Randomforest, Support vector machine, K- Nearest Neighbour and Naïve bayes, this research has worked on Decision trees and random forest.

Decision Trees are known for their interpretability, offering clear decision paths. By reusing this method, this research benefits from a transparent understanding of the education-related factors influencing term deposit decisions. Additionally, Random Forests enhance this interpretability by providing a collective assessment of feature importance, contributing to a more nuanced understanding. [2]

By reusing these methods, your research anticipates improved generalization to new, unseen data. Decision Trees and Random Forests are known for their ability to generalize well, making them suitable for predicting term deposit behavior beyond the confines of the training dataset. [2]

2) Bank Marketing Dataset

For this dataset, since it is a classification dataset, we have used Support vector machine model and Logistic regression model. As per the majority of the research papers which this research has gone through has found out that majority of the papers have used ensemble methods like AdaBoost, Bagging, Neural network and clustering for this dataset. Choosing SVM and logistic regression would help this research to produce new insights and significant results for this research. [9] [10] [11] [12]

3) Apartment/ House rent Dataset

This dataset is used to predict prices of an apartment/house on the basis of major factors like number of bathrooms, bedroom, city names and area in square feet. Previously, algorithm which was applied to this dataset is lasso regression with feature selection as per the research papers. [5] [7]. For this research we have used linear regression due to less complexity of our dataset. Lasso Regression with feature selection is often preferred when dealing with datasets with a large number of features and potential multicollinearity. Linear Regression might be chosen when interpretability is a top priority or when the dataset is not overly complex.

IV. Results and Evaluation

This phase consists of evaluating the results and outcomes of data mining algorithms implemented in the Data mining phase of KDD to check if usable and acceptable trends have been found from the datasets. In this phase various visualizations and predictive assessments have been plotted. Below are the results of all the models applied to their corresponding 3 datasets.

1) Adult/ Income Dataset

For this dataset we have applied decision tree and random forest algorithms, please find the results below-

*a) **Decision Tree:***

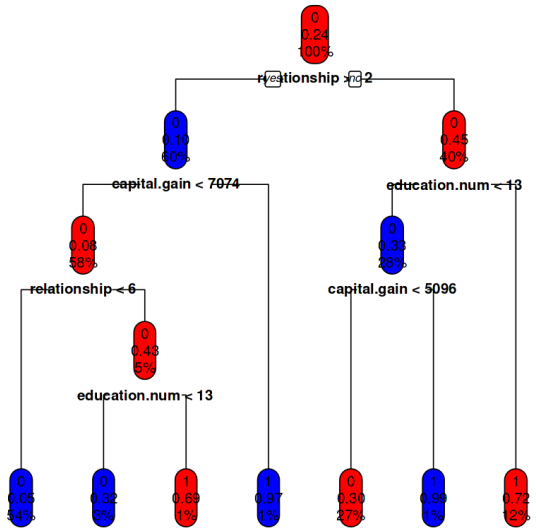

Figure 18. Decision Tree output

Observation: As illustrated from figure 18, the tree is very deep. This means that the model is making a lot of decisions before it arrives at a prediction. This could be an indication that the data is complex or that the model is overfitting.

The splits are based on a variety of features. This is a good sign, as it means that the model is not relying on just one or two features to make predictions.

The left branch of the tree is generally associated with higher values of the target variable, while the right branch is associated with lower values. This suggests that the model is able to capture some of the underlying structure of the data.

The percentages at each node represent the proportion of data points that fall into each branch. This information can be used to assess the importance of different features in the model.

Overall, the figure 18 suggests that the random forest model is able to make complex predictions based on a variety of features. Below is the accuracy, RMSE, RSS, Precision, Recall and f-measure values of the model.

```
Accuracy: 0.8454405
RMSE: 0.3931406
RSS: 1500
Precision: 0.8613983
Recall (Sensitivity): 0.9492675
F-measure: 0.9032008
```

Figure 19. Values from Decision tree model

Based on figure 19, here are the observations:

Accuracy (0.8454405) - The model correctly predicts outcomes approximately 84.54% of the time, indicating a reasonably high overall accuracy.

RMSE (0.3931406) - The Root Mean Squared Error of 0.3931406 suggests that, on average, the model's predictions deviate by approximately 0.39 units from the actual values.

RSS (1500) - The Residual Sum of Squares of 1500 indicates the sum of squared differences between predicted and actual values, providing a measure of the model's overall goodness of fit.

Precision of 0.8613983 signifies the proportion of correctly predicted positive instances among all instances predicted as positive, reflecting a high level of accuracy in positive predictions.

The Recall of 0.9492675 indicates the proportion of actual positive instances correctly predicted by the model, highlighting the model's effectiveness in capturing positive cases.

The F-measure of 0.9032008 provides a balanced metric combining Precision and Recall, suggesting a robust overall performance in correctly identifying positive instances.

b) ***Random Forest:***

```
Call:
 randomForest(formula = income ~ ., data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 13.6%
Confusion matrix:
      0    1 class.error
0 16186 1017   0.0591176
1  2062 3383   0.3786961
```

Figure 20. Random Forest output in the form of confusion matrix.

From above figure 20, we can calculate the accuracy of the model.

True Positives (TP): 3383
True Negatives (TN): 16186
Total: 16186+1017+2062+3383=23748

Accuracy= (3383+16186)/23748 = **0.8234** (Approx) from confusion matrix

```
Accuracy: 0.8673879
RMSE: 0.3641594
RSS: 1287
Precision: 0.88852
Recall (Sensitivity): 0.9438416
F-measure: 0.9153457
```

Figure 21. Random forest accuracy parameters

The model achieves an accuracy of 86.74%, indicating a high percentage of correct predictions overall.

Observations from figure 21:
With a Root Mean Squared Error of 0.3641594, the model's predictions deviate by approximately 0.36 units on average from the actual values, reflecting a relatively low level of prediction error.

The Residual Sum of Squares is 1287, representing the sum of squared differences between predicted and actual values. A lower RSS indicates a better fit of the model to the data.

The Precision of 0.88852 denotes the proportion of correctly predicted positive instances among all instances predicted as positive, showcasing a high level of accuracy in positive predictions.

The Recall of 0.9438416 signifies the proportion of actual positive instances correctly predicted by the model, indicating the model's effectiveness in capturing positive cases.

The F-measure of 0.9153457, a balanced metric combining Precision and Recall, suggests a robust overall performance in correctly identifying positive instances.

Class Error Analysis: From figure 16, The class error, which is a measure of misclassification rate, is approximately 10.48%. This indicates that, on average, around 10.48% of instances are misclassified by the model.

Overall, both the models Decision tree and random forest have worked efficiently with the given Adult/Income dataset.

2) Bank Marketing Dataset

This research has applied Support vector machine model and logistic regression model.

*a) **Support vector machine:***

It has worked nicely as per the accuracy and other metric values given in the figure 22.

```
Accuracy: 0.7958146
F1-Score: 0.8095902
Precision: 0.7964893
Recall (Sensitivity): 0.8231293
```

Figure 22. SVM accuracy parameters

Observations from figure 22:

The Support Vector Machine achieves an overall accuracy of 79.58%, indicating the proportion of correctly classified instances.

The F1-Score of 0.8095902 signifies a balanced metric combining Precision and Recall, offering a robust evaluation of the model's performance.
With a Precision of 0.7964893, the model accurately predicts positive instances, minimizing false positives.

The Recall of 0.8231293 indicates the model's ability to capture a high proportion of actual positive instances, highlighting its sensitivity.

b) Logistic Regression model:

Below are the resultant values given in figure 23-

```
Accuracy: 0.7922272
F1-Score: 0.7679466
Precision: 0.8132956
Recall: 0.7273877
```

Figure 23. Logistic regression model accuracy parameters

As observed from figure 23,

The logistic regression model achieves an overall accuracy of 79.22%, indicating the proportion of correctly classified instances.

The F1-Score of 0.7679466 represents a balanced metric between Precision and Recall, offering a robust evaluation of the model's performance.

With a Precision of 0.8132956, the model accurately predicts positive instances, minimizing false positives.

The Recall of 0.7273877 indicates the model's ability to capture a significant proportion of actual positive instances, highlighting its sensitivity.

As the accuracy for both the models are almost same, both of them worked efficiently with the Bank marketing Dataset.

3) Apartment/House Dataset

This research has applied linear regression on this dataset, below are the results, given in figure 18:

```
[1] "MSE: 119839.231114896"
[1] "R-squared: 0.827289141792265"
RMSE: 346.178
RSS: 244591871
MAPE: 12.40957 %
```

Figure 24. Linear regression model's accuracy parameters

As observed in the figure 24,

The high R-squared value of 0.8273 indicates that a significant proportion (approximately 82.73%) of the variability in the dependent variable is captured by the model, reflecting a good overall fit.

The model's MSE of 119839.23 suggests a relatively large average squared difference between predicted and actual values, indicating potential variability in prediction accuracy.

The RMSE of 346.178 signifies that, on average, the model's predictions deviate by approximately 346.178 units from the actual values, indicating a moderate level of prediction error.

The RSS of 244591871 provides a sum of squared differences between predicted and actual values, offering an overall measure of how well the model fits the data.

The MAPE of 12.40957% indicates the average percentage difference between predicted and actual values, with a lower MAPE generally considered better.

A value of 12.40957% suggests a moderate level of accuracy in predicting percentage differences.

Overall, while the R-squared value demonstrates a good fit, the MSE and RMSE suggest that there might be room for improvement in reducing prediction errors.

Coming to the research questions:
Question 1: How does the level of education impact the probability of an individual opening a term deposit?

Answer: As observed in figure 25, higher education may increase the likelihood of opening a term deposit. People with tertiary education often have more financial knowledge, income stability, and long-term goals, making saving through deposits more attractive.

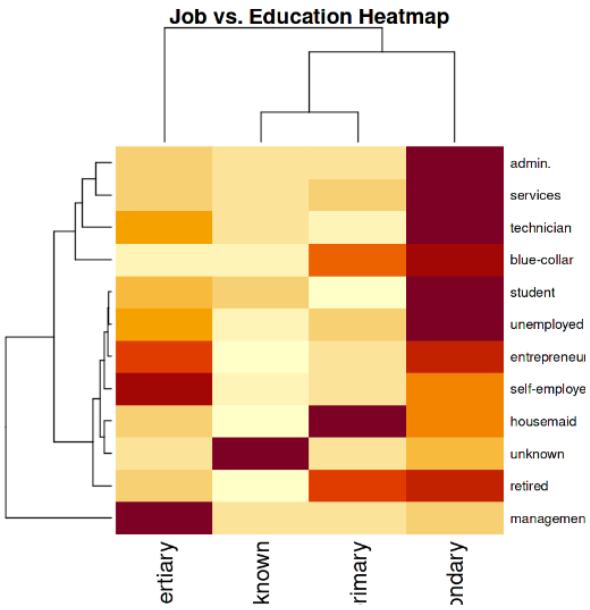

Figure 25. Level of education V/s term deposit

Question 2: What factors contribute to the observed disparity in income between males and females?

Answer: As observed in figure 26, women's larger share of unpaid domestic work further limits their earning potential. This gap varies across education levels and work classes, but tackling it requires promoting equal opportunities, combatting bias, and supporting work-life balance.

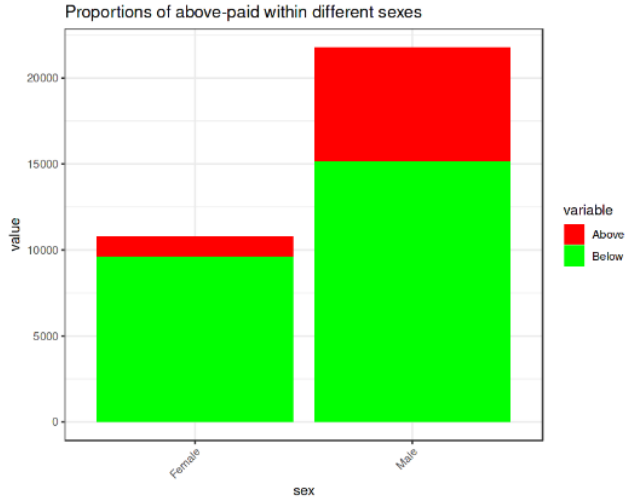

Figure 26. Sex V/s Salary

Question 3: How do interactions between variables such as size of an apartment affect the price of an apartment?

Answer: Scatter plot graph in figure 27 illustrates the association between square footage and price. A notable observation is the concentration of data points in the lower ranges of both variables, suggesting a tendency for smaller properties to have lower prices. Few outliers, featuring higher prices and larger square footage, disrupt this trend. This infers a negative correlation, signifying that as apartment size decreases, prices tend to decrease.

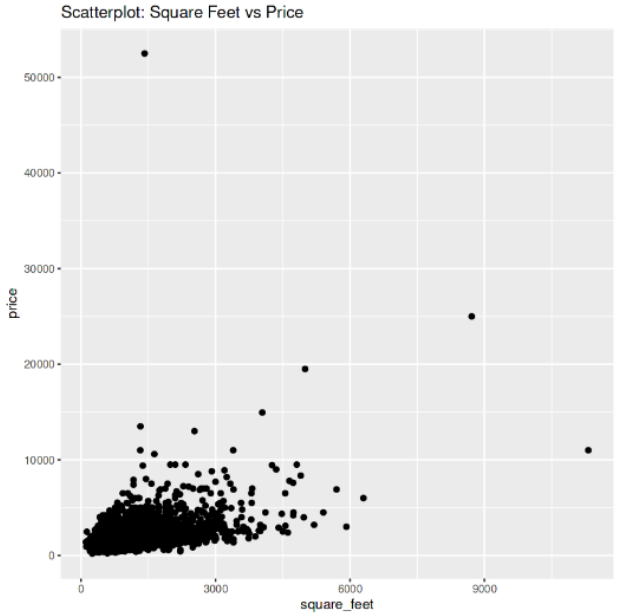

Figure 27. Scatter plot of price V/s square feet (size of apartment)

V. Conclusion and Future Work

This research undertook a comprehensive analysis of success indicators, encompassing income, financial choices, and housing dynamics, using predictive analytics and machine learning on three distinct datasets. The study applied models such as Decision Trees, Random Forests, Support Vector Machine, Logistic Regression and Linear Regression to explore complex relationships. Notable findings include the impact of education on income and the correlation between education levels and the likelihood of opening a term deposit. The analysis of house prices revealed a negative correlation between apartment size and prices. The research suggests potential future explorations in understanding the intricate relationship between education and financial choices, addressing gender disparities, and refining models for improved predictive accuracy. The integration of external data and consideration of additional variables in housing dynamics could further enhance the overall understanding of success indicators. Overall, this research contributes valuable insights for informed decision-making in the realms of income, financial choices, and housing dynamics, setting the stage for future endeavors in this domain.

For future work, refining the predictive models and exploring advanced machine learning techniques could enhance the accuracy and robustness of the analysis. Incorporating additional datasets, considering long-term trends, and investigating cultural and regional variations would offer a more comprehensive perspective. Addressing potential biases in gender-based income disparities through fairness-aware machine learning methods and examining the influence of external economic factors could further enrich the research findings. Collaborative efforts with experts from diverse fields and the development of practical tools for financial decision-making could extend the impact of this research, fostering a more informed and empowered society.

References

1. Papers with Code - Identifying and examining machine learning biases on Adult dataset. (2023, October 13). https://paperswithcode.com/paper/identifying-and-examining-machine-learning

2. Human verification. (n.d.). https://www.semanticscholar.org/paper/Integrating-Association-Rules-with-Decision-Trees-Ayyagari/bddccdae36d08d42a891a9c6fc2c6fef6c9b6cba

3. Becker, D. (2023). Identifying the best machine learning model for predicting bank term deposits: An empirical study using public, post financial crisis data. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3072757

4. [PDF] Fair Regression: Quantitative Definitions and Reduction-based Algorithms | Semantic Scholar. (2019, May 30). https://www.semanticscholar.org/reader/1f868b5839b3126209612f6e2f8c40aa431b46fd

5. Edelsztein, V. C. (2023). Breaking down the Gender Pay Gap through a machine learning model. https://www.redalyc.org/journal/105/10574559005/html/

6. Xi, Z. (2023). Machine learning and prejudice: building theory with algorithm-supported abduction. NTU Singapore. https://dr.ntu.edu.sg/handle/10356/165160

7. Bonaccolto-Töpfer, M., & Briel, S. (2022). The gender pay gap revisited: Does machine learning offer new insights? Labour Economics, 78, 102223. https://doi.org/10.1016/j.labeco.2022.102223

8. Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

9. [PDF] Fair Algorithms for Clustering | Semantic Scholar. (2019, January 8). https://www.semanticscholar.org/reader/34a46c62cb3a7809db4ed7d0c1a651f538b9fe87

10. Human verification. (n.d.-b). https://www.semanticscholar.org/paper/AdaFair%3A-Cumulative-Fairness-Adaptive-Boosting-Iosifidis-Ntoutsi/18fe4800f3c85f315d79063d6b0fe38c7610ad45

11. Yoon, J., Zame, W. R., & Van Der Schaar, M. (2018). TOPS: Ensemble Learning with Trees of Predictors. IEEE Transactions on Signal Processing, 66(8), 2141–2152. https://doi.org/10.1109/tsp.2018.2807402

12. Wu, Y. (2022). Machine Learning Approaches for Retail Bank Marketing practice. BCP Business & Management, 23, 931–937. https://doi.org/10.54691/bcpbm.v23i.1475

13. Exploratory analysis of bank marketing campaign using machine learning; logistic regression, support vector machine and k-nearest neighbour - NORMA@NCI Library. (n.d.). https://norma.ncirl.ie/4574/

14. GeeksforGeeks. (2023, May 23). KDD process in data mining.

15. UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/2/adult

16. UCI Machine Learning Repository. (n.d.-b). https://archive.ics.uci.edu/dataset/222/bank+marketing

17. UCI Machine Learning Repository. (n.d.-c). https://archive.ics.uci.edu/dataset/555/apartment+for+rent+classified