Transactions on Engineering and Computing Sciences - Vol. 12, No. 2
Publication Date: April 25, 2024
DOI:10.14738/tecs.122.16892.
Abu Alrub, M., & Shaout, A. (2024). Power DataMate Tool: Leveraging Logistic Regression Classification for Interactive Data Modeling. Transactions on Engineering and Computing Sciences, 12(2). 195-216.
SERVICES FOR SCIENCE AND EDUCATION

Power DataMate Tool: Leveraging Logistic Regression Classification for Interactive Data Modeling

Mahmoud Abu Alrub
ORCID: 0000-0002-8916-0259
The Electrical and Computer Engineering Department,
The University of Michigan – Dearborn, USA

Adnan Shaout
The Electrical and Computer Engineering Department,
The University of Michigan – Dearborn, USA

ABSTRACT

The demand for efficient predictive modeling techniques has become crucial due to the growing occurrence of binary classification problems in diverse fields. This paper presents a new and innovative software tool "Power DataMate" (PDM) which performs logistic regression (LR) classification as a potent technique for data modeling. PDM is a powerful tool in capturing and analyzing correlations across varied datasets. One objective of PDM is to focus on logistic regression for inquiry based on its capability to represent intricate interactions between predictors and the binary response variable. Another important goal of the new tool is to forecast the likelihood of discovering Primary Keys (PK) and Foreign Keys (FK) within datasets. The new tool PDM also allows users not only to automatically have their data for a project modeled, but also interactively review and confirm primary keys and features for further data analysis and modeling. While the research entails a comprehensive evaluation of model performance indicators, including accuracy, precision, and recall, results show that the accuracy of PK prediction is 89% and 82% for the FK. Hence, these results are the first of their kind and could be a starting point for further model enhancements and data analytics research, especially for projects which include large data files where PDM end user has the choice to interactively feed the learning algorithm for better outcomes.

Keywords: Data Mining, Data Modeling, Data Classification, Logistic Regression, Primary Key, Foreign Key.

INTRODUCTION

In a time of unparalleled data abundance, extracting valuable knowledge from large and complex datasets has become a critical task for researchers and data scientists in many fields. It is more important than ever to use data modeling techniques as we find ourselves at the nexus of information and technology. The present paper delves into the field of data modeling as the combination of methodologies from statistics, machine learning, and database systems which are employed to extract probabilities and insights that may be utilized for decision-making and strategic planning purposes.

While data classification is an essential component within the expansive domain of data mining, classification refers to the procedure of assigning predetermined labels or categories to instances, taking into consideration their distinctive features [1]. On the other hand, logistic regression has also recently been a fundamental component in the domain of statistical modeling and machine learning [2]. The applications of this method are many, encompassing the prediction of consumer behavior and medical results, the detection of fraudulent operations, and the assessment of risk factors. For example, [3].

Data modeling entails the methodical depiction of an entity's data resources and the interconnections that exist among them [4]. It acts as a foundational plan for creating databases, providing a theoretical structure that facilitates the connection between unprocessed data and practical knowledge. In the contemporary era of digitalization, the quantity of data produced daily has attained unparalleled magnitudes. On the other hand, data mining (DM) refers to the systematic exploration and analysis of extensive information with the objective of identifying patterns, trends, correlations, and essential insights. The process integrates methodologies from diverse disciplines, including statistics, machine learning, and database management, to convert unprocessed data into practical insights [5].

The process of data classification entails the methodical grouping of data based on predetermined criteria, including confidentiality and integrity [5]. Classification problem in the context of data mining is the practice of giving predetermined labels or categories to instances by considering their distinct properties. The task includes the pursuit of creating precise, effective, and comprehensible models that can automatically assign predetermined labels to instances of data. This achievement would enable the automation of processes and provide improved decision-making support in a wide range of applications [5-6]. Within the expansive realm of statistical modeling and machine learning, logistic regression emerges as a potent and extensively utilized methodology that specifically addresses binary or multi-class classification challenges. Logistic regression, a fundamental technique in statistical analysis and probability theory, is utilized to describe the association between input variables and categorical outputs [2][7].

The paper will present a new and unique software tool called Power DataMate (PDM) that will enable the user to create data management projects, review and analyze datasets, understand its initial entities, features, and instances, and automatically or interactively model the data. PDM is a web-based comprehensive data modeling application.

The paper is organized as follows: section 2 explores the related work in data modeling using classification and logistic regression techniques, mainly, in PK and FK identification, section 3 reviews the existing academic knowledge and serves as a natural foundation for the rest of the study, section 4 elaborates PDM functionalities in terms of predictive and automated and interactive classification models, section 5 describes how PDM analyzes, trains, tests, and builds a data model, section 6 draws attention to the study's limits and scope, and lastly section 7 summarizes the key findings, gives interpretations, and concludes the recommendations of PDM data models.

RELATED WORK

Several studies explained the classification problem and how logistic regression helped in modeling data using machine learning techniques. However, many have not concluded mining the data to understand its relations by discovering the PK and FK. Our interest relies on automatically modeling the data from distinct types of data sources by classifying PKs and FKs and interactively allowing one to have his input on keys and features to improve the maturity of our prediction model. This section investigates the existing work in terms of data modeling through data mining, data modeling using logistic regression, and lastly PK and FK discovery.

Data Modeling through Data Mining

Data mining can be classified from various viewpoints: 1. In terms of data, it can be categorized as supervised, semi-supervised, or unsupervised. 2. In the context of data mining, it can serve the aim of predictive analysis or descriptive analysis. 3. The used algorithm can be categorized into classification, regression, clustering, pattern mining, etc. [8]. While we aim to use supervised data for predictive classification, it is necessary to ensure manageability and traceability. Thus, it is recommended to follow a guided approach when undertaking data mining activities. To facilitate this, various data mining standard processes have been developed [9-12].

Applying a data mining model to input data without a well-organized application, often known as "data dredging" [13], can result in the production of meaningless or unintelligible patterns, ultimately leading to data mining or modeling failure. The utilization of a standardized data mining framework could serve as a quantifiable guide for individuals to adhere to when engaging in problem definition, data preprocessing, classification model selection, and data modeling. In this contribution, PDM has the capability for one to create a well-structured data modeling project and apply a nine-step framework to ensure quality classification results as will be presented in Section 5.

The rapid expansion of data and databases necessitates the creation of novel technologies and tools to transform data efficiently and autonomously into valuable information and knowledge. Data mining has gained significant prominence as a study field, as evidenced by its growing importance in several studies [14-16]. Unlike existing work that either discusses the data mining techniques and algorithms or applies them onto domain-specific areas such as the field of education, PDM encapsulates the process of extracting valuable information from extensive datasets in a simplified user journey so a data analyst can interpret the knowledge of a refined data model.

Data Modeling using Classification, Logistic Regression

Supervised learning algorithms are characterized by having prior knowledge of the class attribute values in the dataset before executing the algorithm. The data is referred to as labeled data or training data [17]. The elements in this collection are tuples represented as (x, y), where x is a vector and y are the class property, often a scalar value. Supervised learning constructs a model that establishes a mapping between the input variable x and the output variable y [12]. In accordance with the above, we utilize the supervised learning algorithm through the logistic

function where we study the relationship between the dependent variable, logit(pi), and the independent variable, x.

Class attribute values in classification are categorical. Classification is the process of assigning data elements to certain classes based on the dissimilarity of data elements belonging to distinct classes. The objective is to ascertain the criteria that determine whether an item is part of a specific subset or category of data [18]. According to Han, Kamber, and Pei [19-20], categorization involves a two-step process of constructing a model that describes a selected set of classes. It is expected that each tuple belongs to a predetermined class, which is determined by the class label parameter. The collection of tuples utilized for constructing the model is referred to as the training set. The model is expressed as classification rules, decision trees, or mathematical equations. Subsequently, the model will be employed to classify forthcoming or unfamiliar entities [9]. Indeed, this is part of our development where a column name gets inspected among a data table to assess the success of detecting PKs or FKs. In this contribution, we make use of the threshold value idea, which establishes the likelihood of either 0 or 1. For example, values above the threshold tend toward one, and those below the threshold tend toward zero.

There are three prediction techniques [21]: classification, regression, and density estimation. Each category employs a distinct variety of variables. In classification, categorical or binary variables are utilized. The objective of prediction is to deduce a target attribute or specific element of the data (predicted variable) based on a combination of other aspects of the data (predictor variables). In our research, prediction has been employed to forecast table relations and identify primary and foreign keys [18, 22]. This means that PDM employs binomial logistic regression where there are only two conceivable forms of dependent variables such as 0 or 1, Pass or Fail. Fatih Ertam and Galip Aydin utilize functions such as Rectified Linear Unit (ReLU), Hyperbolic Tangent (tanh), and Exponential Linear Unit (ELU), and classifiers including Convolutional Neural Network (CNN) and SoftMax to obtain the highest level of accuracy in categorization [1]. The performance of the Tensorflow library in image processing systems was evaluated using the MNIST dataset [23]. Dissimilarly to the deep learning neural network, our tool uses the activation output in a typical classifier which is generated by a single unit. However, in Fatih Ertam and Galip Aydin work, a single layer may contain several units, each of which is a classifier with a different activation function.

Additionally, Chen et al. address the high dimensional data [24] arising issue by enhancing the focus on extracting important features from these high dimensional variables. They used statistical and classification techniques such as Random Forest and Support Vector Machines [25]. However, PDM is concerned when the quantity of explanatory variables equals or is less than the quantity of noise variables. In other words, rather than maximizing the margin among class variables (SVM) or having two levels of means, the sample in the tree target cell, and then all trees (Random Forest), our tool predicts the probability that an instance belongs to a given class or not. On the other hand, David Williams et al. have primarily developed a missing-data logistic regression classification technique that effectively addresses the incomplete-data problem by making two reasonable assumptions, hence eliminating the need for imputation heuristics [26]. Furthermore, the revised equations for constructing a Gaussian Mixture Model

(GMM) [27] using incomplete data through the Expectation-Maximization (EM) [28-29] and Variational Bayes Expectation-Maximization (VB-EM) methods [30] have also been provided. Nonetheless, the purpose of this work is basically finding procedures of using alternative values in place of missing data, which is basically, the step we take to preprocess and prepare the datasets in our data modeling framework. While the developed logistic regression technique sounds effective in addressing the incomplete-data issue, it does not propose a data modeling solution especially discovering the relationships between data entities.

With the context of logistic regression methodological reviews, Stephan Dreiseitl and Lucila Machado offer guidelines that are valuable for conducting a thorough evaluation of classification models and the outcomes derived from them [7]. They present a summary of observations about the fulfillment of quality standards for logistic regression [2] and artificial neural network models. However, it is found that the popularity of logistic regression can be attributed to its interpretability of model parameters and ease of use. Artificial neural networks [31-32] are also popular because they can be seen as nonlinear generalizations of logistic regression, making them at least as powerful as that model.

Zan Yang and Dan Li proposed a new logistic regression approach with a filter [33] to overcome the traditional logistic regression, which employs the gradient descent method to determine the best parameters of the loss function, which exhibits a particular level of convexity. Their study utilizes filter theory [34] from the domains of operations research and cybernetics to determine the optimal parameters of a loss function [35-36]. Despite the inclusion of the filter, its application is on datasets where you are interested in the variation in the proportion of forward and reverse reads in two replicate sets, each including at least three data stores. It is believed that such a procedure would be more convention if used in specific data analysis tasks and within domain-limited projects. Moreover, using PDM, we do not aim to bi-directionally read the data when modeling. In addition to the above, Fortunato et al. studied and compared the logistic regressions estimated by the maximum likelihood model (LRMML) and the logistic regression model estimated using the Boosting algorithm [37-38], specifically the Binomial Boosting algorithm (LRMBB). To elucidate this scenario, the chosen illustration involved categorizing the occurrence of coronary heart disease (CHD) [39, 11, 40]. By any means, Fortunato et al. evaluate classification models whereas we contribute to studying relational datasets, discovering PKs and FKs, and finally modeling data.

Jihong Yan and Jay Lee presented a methodology that uses logistic regression to assess machine degradation and classify its primary cause [41]. The combination of Wavelet Packet Transform (WPT) and Fisher's criterion contributed to having an effective method for extracting features from non-stationary data [8]. Thus, although the categorized performance and causes help avoid machine degradation the paper misses the threats to validity concerning the linear regression. Compared to PDM, the work may be useful specifically for industrial manufacturing. Jawad Raza et al. work aimed to categorize the maintenance data obtained from both the strainer's clean state and its half-full state in the oil and gas industry. The gathered data was categorized as "1", indicating a healthy and clean condition, and "0", indicating a malfunctioning and unclean condition of the filter. They employed three distinct methodologies for data categorization, logistic regression, principal component analysis-based artificial neural

network (PCA-based ANN), and support vector machine (SVM) [42-43] to assess their suitability and effectiveness in categorizing the provided data [44]. In contrast to our tool, this work does not utilize neither relies on logistic regression knowledge. Nevertheless, we have not seen a use case for managing data projects for extensive datasets not only in the form of data files, but also through connecting to a database.

Primary and Foreign Key Prediction

While the establishment of a foreign/primary key relationship between relational tables is a crucial restriction within a database and identifying foreign keys is an essential step in comprehending and manipulating the data [45-47], Meihui Zhang et al. represented a technique, referred to as Randomness, that encompassed a range of laws by devising highly effective approximation methods for assessing Randomness, thus, employing a mere two iterations over the data [48]. The research identified foreign keys in both single-column and multi-column formats. While Randomness focuses on detecting FKs within relational tables, it is important to consider the other constraint of relational data, thus, PKs. Moreover, our research brings great user experience into the table where one can himself verify and validate the success guesses of PKs and send the results to the learning model for better accuracy rate. We believe that a refined and tested data model that would PDM produce challenges the discovery of single-column or multi-column FKs. Significant advancements have been achieved in addressing the issue of Consistent Query Answering (CQA) in relation to primary keys since 2005 [49-50]. The given challenge involves an input that consists of a database instance that potentially violates one or several PK restrictions. Therefore, Miika and Jef examined the issue of CQA in relation to both PKs and FKs. That's said, it was logical to embrace the concept of symmetric-difference repairs, as they allow for the correction of FKs by adding new tuples [51]. Although the developed work incorporates a resolution to the CQA where the data that remains constant over all consistent database instances is typically referred to as consistent data in an inconsistent database it has a different area of focus from our tool. Unlike Miika and Jef, we propose a tool that mainly supports in analyzing datasets with the purpose of accurately discovering PKs and FKs using LR.

Mozhgan Memari et al. presented techniques for analyzing unary and multi-column FKs in real-world scenarios, explicitly considering the ways null markers are interpreted according to the SQL standard. The study concluded that the comprehensive tests provide evidence that the author's algorithms achieve similar performance in many case studies [52]. This research considers the null markers presence to identify FKs where we believe further steps must be taken, as our classification framework does, to build on data modeling. Vladimir focused on the issue of automatically constructing non-repetitive sets of referential constraints on the database schema [53]. While the inclusion dependencies serve as the foundational concept for referential restrictions and to ensure theoretical accuracy, the concept of an acyclic database design was introduced and studied. The rules for decomposing relations were provided to establish referential restrictions automatically based on the PK of the relations. However, this research treats the referential restrictions in a database which helps in database design. Alternatively, PDM works with both data files and databases with a wider range of analytics. Alexandra et al. investigated the issue of autonomously identifying semantic connections among FKs. They highlighted the significance of the issue in any scenario where there is a need

to combine datasets that are organized in tables but lack clear FK limitations. Therefore, the contribution proposes a machine-learning methodology to address this issue [54]. In accordance with the above, a binary classification technique is utilized to evaluate dependencies, however, it is necessary to clarify that identifying potential candidates for FK restrictions cannot be sufficient to analyze and model the data. Our research leverages the LR not only to discover referential data features but also unique instances where it greatly connects the dots and results in conceptual and logical data modeling.

CONCEPTUAL FRAMEWORK

In this section, we aim to provide a roadmap for our research including a theoretical framework of data mining techniques, data classification problem and logistic regression, and other important data management practices and tools such as data normalization and data warehouse.

Data Mining Techniques

Data mining refers to the systematic examination and exploration of extensive datasets with the objective of identifying meaningful patterns and correlations. This analytical approach aids in addressing business challenges by leveraging data analysis techniques. The utilization of data mining approaches and tools empowers organizations to forecast forthcoming trends and make well-informed company decisions [55]. Figure 1 shows the data mining techniques which fall into two categories, predictive and descriptive [5]. Our focus is on the predictive ones where classification comes into the picture.

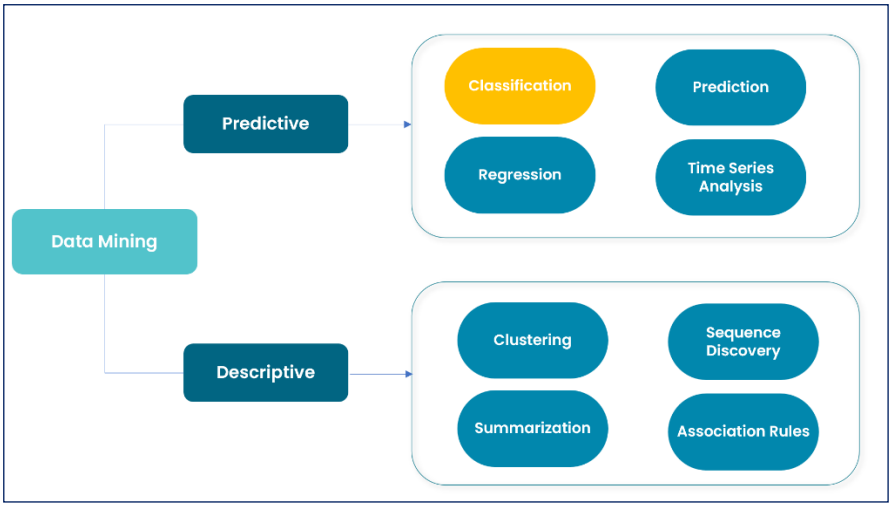


Figure 1: Data mining techniques.

Classification

The task of classification involves determining the appropriate category or subpopulation to which a new observation belongs, using a training set of data that includes prior observations and their corresponding category memberships [6] [20]. Its framework includes problem definition, where the classification problem is explained, data collection and exploration, data preprocessing, where data integrity is examined, feature selection, algorithm selection, model training, model evaluation, and interpretability and visualization.

Logistic Regression

The statistical model, commonly referred to as the logit model, is frequently employed in the fields of categorization and predictive analytics. The LR is a statistical method used to assess the likelihood of an event taking place, such as voting or not voting, by utilizing a dataset consisting of independent variables. Given that the result is expressed as a probability, the dependent variable is constrained inside the range of 0 to 1. The LR involves the application of a logit transformation to the odds, which represents the ratio of the probability of success to the probability of failure. This is frequently referred to as the logarithm of odds or the natural logarithm of odds [7].

In terms of machine learning, logistic regression is a supervised machine learning approach commonly employed for classification tasks, wherein the goal is to find an instance belonging to a specific entity (a PK or FK record). The term "regression" is used to describe this process, as it involves taking the result of the linear regression function and utilizing a sigmoid function to estimate the probability associated with the specific class. Moreover, logistic regression represents the relationship between the dependent or response variable, and the independent variable. This approach involves conducting several iterations to test various values of beta to maximize the log odds for the best match. In the context of binary classification, it is commonly understood that a probability value below 0.5 is indicative of a predicted class label of 0, whilst a probability value over 0.5 is indicative of a predicted class label of 1.

Data Warehouse

A data warehouse is a specific form of data management system that has been strategically developed to facilitate and uphold business intelligence (BI) endeavors, with a particular emphasis on analytics. Data warehouses are primarily designed for the purpose of executing queries and analysis, typically housing substantial volumes of historical data [56]. The data included within a data warehouse typically originates from several sources, including application log files and transactional applications. A well-designed data warehouse [57] can be regarded as the primary and authoritative repository of information inside an organization.

POWER DATAMATE DESIGN

PKs and FKs play a crucial role in relational schema data across diverse applications, including query optimization and data integration. Nevertheless, it is often the case that these limitations are either undisclosed or lack proper documentation. The manual detection of these entities is a laborious and impractical task, mainly when dealing with datasets of significant scale. In this paper, we address the issue of automatic and interactive discovery of PKs and FKs.

We present a novel tool, referred to as PDM, which is capable of loading and reading datasets, automatically and interactively detecting both types of keys. PKs and FKs are seen as subsets of unique column combinations (UCCs) and inclusion dependencies (INDs), respectively. Efficient discovery techniques have been developed for both subsets. By adapting logistic regression binary classification and employing scoring functions, our methodology successfully identifies the accurate primary keys and foreign keys from extensive collections of valid unique column combinations (UCCs) and inclusion dependencies (INDs). Multiple pruning rules are utilized to enhance the efficiency of the prediction. Precision and recall are assessed using four real-world

datasets and two databases. The findings indicate that the employed methodology demonstrates a capability to retrieve an average of 89% of PKs and 82% of FKs.

Figure 2 shows PDM nine-step framework starting with identifying the data classification issue and ending with visualizing the dataset with PK and FK identified. In step 5, logistic regression is used where its context is explained in Figure 3. The classification model gets trained with the training data whereas the metadata, later, gets classified.

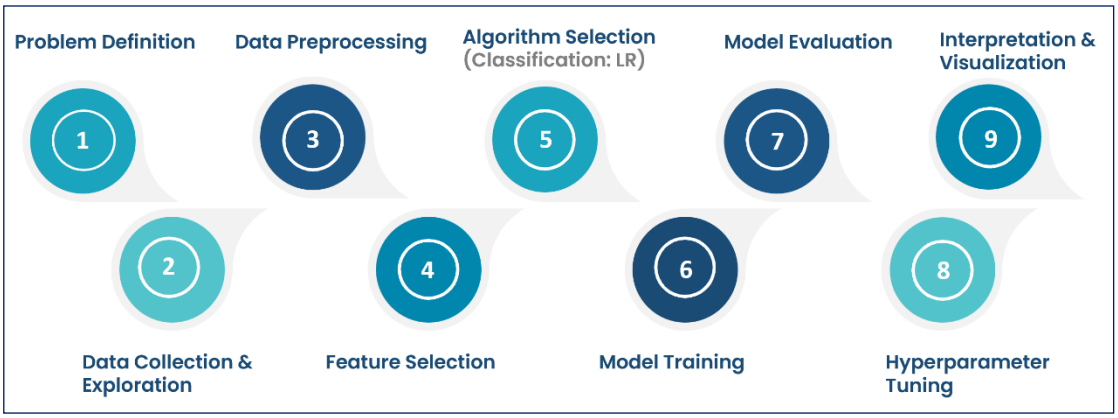
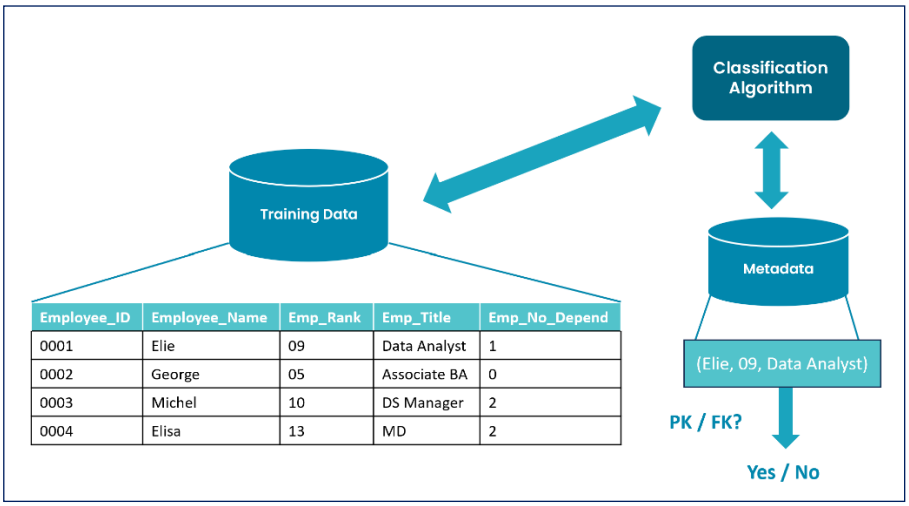


Figure 2: Power DataMate (PDM) classification framework overview.

Per the context demonstrated in Figure 3, the training data is collected and preprocessed. The preprocessing step includes data transformation, where the training data gets restructured, and data clean. The training model, then, gets deployed for prediction. For example, in Figure 3, the Employee table has five data features and four data instances. The data will be transformed into Table Name, Column Name, and Column Data Values (Sample Data), thus, the first records in the newly transformed dataset must include Table Name: Employee, Column Name: Employee ID, Sample Data: 0001, 0002, 0003, 0004.



Employee_ID	Employee_Name	Emp_Rank	Emp_Title	Emp_No_Depend
0001	Elie	09	Data Analyst	1
0002	George	05	Associate BA	0
0003	Michel	10	DS Manager	2
0004	Elisa	13	MD	2

Figure 3: PDM classification problem context example.

Data Modeling Project

The process of designing a data modeling project includes constructing a framework that arranges and depicts data in a manner that fulfills the precise demands of a system or application. In our contribution, PDM enables developers, data scientists, and or business users to create and manage their own data management projects.

One can benefit from PDM by:

- Precisely expressing the aims and purposes of the data modeling effort.
- Enumerating and providing concise definitions for the primary entities and features present within a data file or a database.
- Creating associations between entities, considering the cardinality and participation requirements.
- Specifying the features linked to each entity.
- Creating a conceptual data model to depict the overarching relationships and entities.
- Automatically or interactively constructing a logical data model that encompasses specific information such as primary and foreign keys.
- Verifying the data model to ensure it aligns with the defined requirements.
- Evaluating the performance, scalability, and data integrity of the implemented data model.
- Enhancing PK and FK discovery and prediction by incorporating insights into the machine learning embedded model with the actual output and feedback.

Logistic Regression Classification

PDM develops, using a machine learning classification model through logistic regression, an algorithm that undergoes training to make predictions on the categorical class labels of novel instances. These predictions are based on patterns that have been acquired from labeled training data. As classification is a form of supervised learning in which an algorithm is trained to discern the connection between input features and their corresponding output classes, PDM considers a labeled training dataset where each instance is associated with a class label.

Process (As shown in Figure 2):

1. Define the Problem: The categorization problem we are trying to address is mining the raw data and data tables and getting their relations identified. This is done using logistic regression and the outcome is an efficient and effective data model.
2. Collect and Prepare Data: The initial step in the research process involves the collection and preparation of data. We import and tag a dataset wherein each individual instance is linked with a corresponding class label.
3. Select a Classification Algorithm: The utilized classification algorithm we are using is Logistic Regression.
4. Train and Test: The data is divided into several sets for training and testing purposes. The training set is employed for the purpose of training the model, whereas the testing set is utilized to assess its performance on novel and unexplored data.
5. Train the Model: The training data is inserted into the classification method to train the model. The algorithm acquires knowledge of the patterns present in the data and adapts its internal parameters to generate precise predictions on primary and foreign keys.

6. Evaluate the Model: The model's performance is evaluated by assessing its functioning using the testing set. Prominent evaluation criteria for the classification models encompass accuracy (Model training 0:1, 0 Low, 1 Highest, 25K testing data, for PK 0.89 and for FK Sample Data 2K, 0.82), precision and recall (how many times the model can predict IsPK true or false. how many times the model can predict IsFK true or false).
7. Make Predictions: In this activity, we engage in the act of making predictions. After the completion of training and evaluation, the model is utilized to provide predictions on novel and unobserved data. The model assigns a predicted class label to each occurrence by leveraging the acquired PK and FK.

Figure 4 shows the process flow of using PDM where one can create data modeling project, upload a data file, or connect to a database, discover PKs and FKs, and model data.

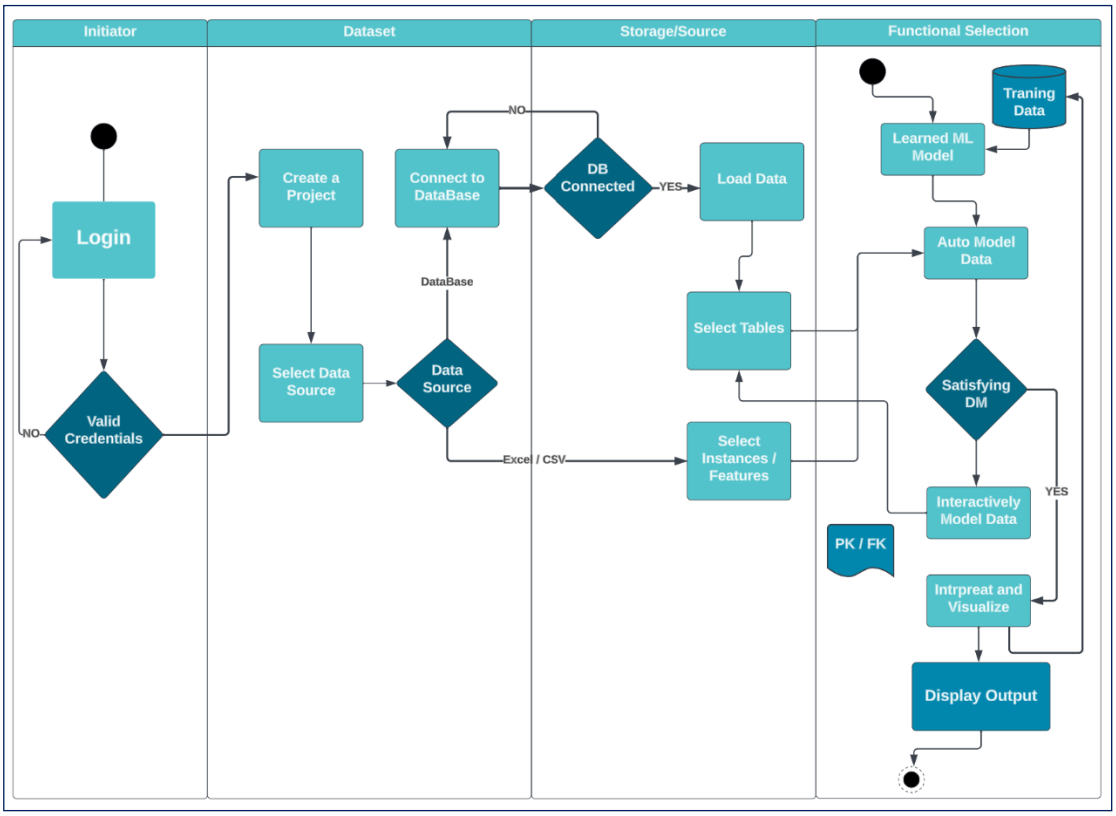


Figure 4: PDM process flow.

Logistic Regression Prediction

To forecast a PK in a dataset, PDM tries to determine a column or a group of columns that can distinctly identify each individual record in the dataset. It utilizes classification model, logistic regression, which gets trained to determine the primary and foreign keys. Moreover, as a preliminary step, it also searches for columns that could potentially function as inherent identifiers, such as an ID column. The algorithms predict the probability of an attribute as a PK or not and an FK or not. The first one takes the Table Name, Column Name, Sample Data, and Target Output as input, while the second one processes the Reference Table Name, Reference

Column Name, Primary Key Table, and Primary Key Column, Sample Data, and Target Output to predict an FK. Additionally, PDM efficiently introduces automatic data modeling and interactive data modeling, which are two distinct methods used to develop and improve models for a specific dataset. The decision between automatic and interactive data modeling relies on particular objectives of the modeling assignment, the intricacy of the data, and the required degree of user skill and involvement. A hybrid strategy that combines both methods can sometimes be employed, capitalizing on the respective advantages of each during certain stages of the modeling process. We discuss each approach in the below sections.

Figure 5 explains the logistic regression context where features as inputs get processed with a logistic classifier to produce the probability of detecting a PK or FK.

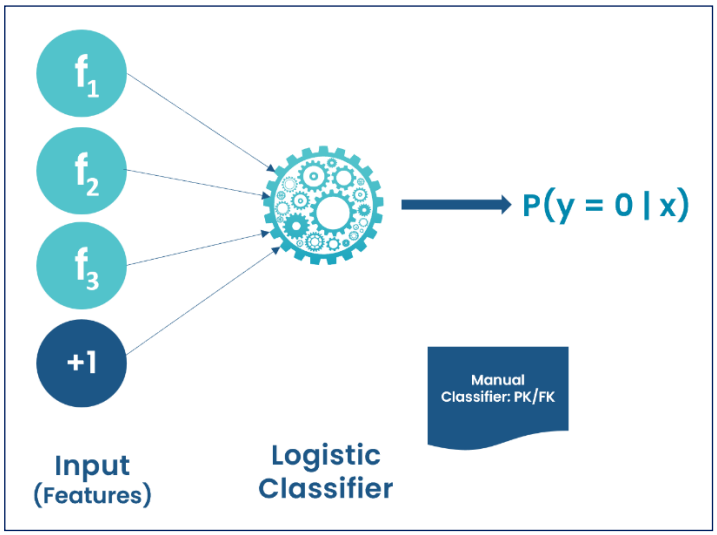


Figure 5: PDM logistic regression context.

Automated Logistic Regression Modeling

The automatic logistic regression modeling using PDM is considered the first phase of data modeling, where the tool utilizes the ML and prediction models to anticipate and discover PKs and FKs.

The following are the characteristics of PDM automated logistic regression modeling:

- Automated Process: Automatic logistic regression modeling refers to the process of utilizing the logistic function to produce models without the need for direct human involvement. The process involves the analysis of the dataset, the selection of features, and the construction of the model.
- Machine Learning Algorithm: Entails the utilization of the ML algorithm, i.e., classification – logistic regression. The algorithm typically handles the complexity of the model and feature selection automatically, and later interactively.
- Suitability: This tool is particularly suitable for situations when there is a huge amount of data, the modeling activity is predictable or repetitive, and there is less requirement for human experience in the modeling process.

Interactive Logistic Regression Modeling

The rapid expansion in classification logistic regression has led to the development of intelligent programs that can make crucial judgments autonomously. Nevertheless, those lacking technical expertise frequently encounter challenges when placing trust in machine learning models, as a significant number of these models are black-boxed. The interactive features enable end-users to intervene and enhance the categorization model created by the machine, hence enhancing the overall performance of the model. To maximize the effectiveness of the classification technique, PDM allows end users to navigate through the automatically created data model, revealing the intricate underlying relationships in the dataset. Subsequently, selected and user-valid data labels are constructed and fed to the machine learning training model.

Thus, the following points are considered while interactively modeling data:

- User Involvement: Refers to the direct participation and direction provided by users during the modeling process. It involves users engaging in the active selection of features.
- Human Expertise: Users, who are typically domain experts or data scientists, actively participate in the decision-making process, resulting in models that are more easily understood and relevant to the specific field.
- Iterative Refinement: Iterative refinement is a common approach that involves users experimenting with various model configurations and evaluating the model's performance in real-time.
- Suitability: Scenarios that are well-suited are those that include complex activities, innovative issues, or circumstances where domain expertise is crucial.

Power DataMate (PDM) Algorithmic Execution

This section explains the main algorithms involved in the model training and prediction for PK and FK:

Primary Key and Foreign Key Model Training:

The following pseudocode provides a structured approach for implementing the PK and FK model training:

1. Start PK or FK Model Training
2. Load Data: from CSV file or Database
3. Convert columns to numerical values using label encoding
   For PK:
   - TABLE_NAME, COLUMN_NAME, SAMPLE_DATA, TARGET OUTPUT

   For FK:
   - TABLE_NAME, COLUMN_NAME, REFERENCE_TABLE_NAME, REFERENCE_COLUMN_NAME, SAMPLE_DATA, TARGET OUTPUT
4. Convert OUTPUT to Boolean
5. Get the feature columns and target column
   For PK:
   - Features of TABLE_NAME, COLUMN_NAME, SAMPLE_DATA
   - Target = OUTPUT

For FK:
- Features of TABLE_NAME, COLUMN_NAME, REFERENCE_TABLE_NAME, REFERENCE_COLUMN_NAME, SAMPLE_DATA
- Target = OUTPUT

6. Feature scaling
7. Split the data into training and testing sets
8. Train the logistic regression model and get it fit
9. Evaluate the model
10. Save the trained model
11. Display Accuracy
12. End PK or FK model training

Primary Key and Foreign Key Automated Prediction:

The following pseudocode provides a structured approach for implementing the PK and FK automated prediction:

1. Start PK or FK Automatic Prediction
2. Load the trained model from the file
3. Get transformed and cleaned form of data
   For PK:
   - TABLE_NAME, COLUMN_NAME, SAMPLE_DATA

   For FK:
   - TABLE_NAME, COLUMN_NAME, REFERENCE_TABLE_NAME, REFERENCE_COLUMN_NAME, SAMPLE_DATA
4. Convert columns to numerical values using label encoding
   For PK:
   - TABLE_NAME, COLUMN_NAME, SAMPLE_DATA

   For FK:
   - TABLE_NAME, COLUMN_NAME, REFERENCE_TABLE_NAME, REFERENCE_COLUMN_NAME, SAMPLE_DATA
5. Make a prediction using the trained model
6. Convert the prediction to Boolean
7. Save the prediction result to the database
8. Generate predicted PKs or FKs file
9. End PK or FK Automatic Prediction

ANALYSIS AND RESULTS

The main goal is to present our research's empirical analysis and findings. This covers how models were trained, quantitative analysis of datasets preprocessed, and the results of the modeled data that achieved our objectives and strengthened the capabilities of our tool.

Experiments

In this section, we illustrate our approach to datasets and their specifications as we utilize four data files in the different business domains and two databases in two different business domains. Additionally, data preprocessing activities have been conducted.

Before Data Preprocessing:

Per our previous sections, the number of datasets involved in our analysis and modeling is six. Table 1 includes the statistics of the four datasets in the form of data files. We have varied the business domain in accordance with achieving better results; ERP, CRM, Healthcare, and HR. Whereas Table 2 includes the statistics of the two datasets in the form of databases. Respectively, they have been extracted from Utility and Violation Management domains. It is expected to have less numbers of instances and features once the data preprocessing activity is completed.

Data Files: Four samples of data files exported from Kaggle [58].

Table 1: Data files description – before data preprocessing.

Sr. No.	Dataset	Instance	Feature
1	ERP	20K	12
2	CRM	<4K	17
3	Healthcare	>21K	15
4	HR	>3K	16

Databases (PostgreSQL): Two samples of database files exported from Kaggle [58].

Table 2: Database description - before data preprocessing.

Sr. No.	Dataset	Instance	Feature
1	Utility Management	>20K	30
2	Warning / Violation	<2K	24

After Data Preparation and Preprocessing

The incorporated datasets, whether data files or databases, have been prepared for better data modeling results. Table 3 and Table 4 show reduced numbers of instances and features per data file and database concluding the validity and effectiveness of the subsequent analysis or modeling tasks. The datasets domains remained unchanged.

Table 3: Data files description – after data preprocessing.

Sr. No.	Dataset	Instance	Feature
1	ERP	15K	11
2	CRM	3K	17
3	Healthcare	20K	15
4	HR	3K	16

Table 4: Database description – after data preprocessing.

Sr. No.	Dataset	Instance	Feature
1	Utility Management	15K	25
2	Warning / Violation	2K	22

Model Training

Training data is a crucial element of machine learning, as it is utilized to instruct models. Training a model entails iteratively optimizing its parameters by analyzing the input data and its matching output labels. To perform classification tasks effectively, we have a comprehensive training dataset that encompasses a substantial number of input-output examples. This is crucial from a modeling perspective. Our model determines the appropriate input data samples mapped to specific class labels using the training dataset. Thus, to ensure the effectiveness of the training dataset, we included several samples for each class label and ensured that it accurately represented a better data model.

For class labels import into a modeling process, we converted string values "IsPK: Yes" or "IsPK: No" and "IsFK: Yes" or "IsFK: No" into numeric values. For instance, the class label "IsPK: Yes" was assigned the integer value of 1, whereas the class label "IsPK: No" was assigned the integer value of 0. Hence, label encoding is employed which involves assigning a unique integer value to each class label.

Prediction

It is typically advised that a data scientist undertakes controlled tests to work with the logistic regression algorithm and algorithm configuration to produce the most remarkable performance for the classification task. Classification predictive modeling is assessed based on its output. A common statistic for assessing a model's performance based on dependent and independent variables is classification accuracy. Although not perfect, classification accuracy is a reasonable place to start for many classification jobs. While some tasks may call for a class membership probability prediction for each example rather than class labels, PDM avoids this as it adds more uncertainty to the prediction, which a user or application can subsequently interpret.

Binary Classification

Binary classification refers to the process of categorizing data into two distinct classes or categories. Classification tasks that involve only two class labels are commonly known as binary classification. A commonly employed approach to depict a binary classification problem is to use a model that predicts a Bernoulli probability distribution for each occurrence. The Bernoulli distribution is a discrete probability distribution that specifically applies to situations when an event can only have one of two possible outcomes: 0 or 1.

The following is the systematic procedure to predict the PK:

- Examine Column Values: Identify columns that exhibit unique values for every entry.
- Check for Unique Identifiers: Detect columns that seem to serve as unique identifiers, such as user IDs, product IDs, or order IDs.
- Evaluate Data Quality: A primary key must be regularly populated without any missing or null values.
- Validate Uniqueness: Examine the candidate primary key columns either manually or programmatically to confirm that there are no duplicate values, thus ensuring uniqueness.

On the other hand, determining foreign keys entails discerning patterns and connections across disparate datasets or tables. The following is the systematic procedure to predict the FK:

- Analyze Data Relationships: Identify columns that indicate connections between items.
- Perform a search for distinct IDs: Identify columns that function as unique identifiers in a specific table and are also utilized in other tables.
- Analyze Data Values: Examine the values in columns that could potentially be used as foreign keys.
- Search for Naming Conventions: Occasionally, naming conventions offer clues.
- Validate Data Types: Ensure that the data types of prospective foreign keys align with the data types of the corresponding primary keys in other tables.
- Verify Referential Integrity Constraints: If the data originates from a relational database, examine explicit referential integrity constraints that specify foreign key connections.
- Validate Relationships: Validate relationships by comparing values in prospective foreign key columns with corresponding primary key columns in connected tables.

Modeled Data

Our primary goal is to minimize the occurrence of false negatives (IsPK:1) without necessarily decreasing the number of false positives (IsPK:0). We opt for a decision threshold that yields a low Precision or a high Recall score. In primary key / foreign key prediction, it is crucial to avoid misclassifying any primary keys as foreign or other types of keys, regardless of the possibility of incorrect foreign keys. This is because the disability of discovering a primary key can be identified by additional model training on a specific portion of the data or by utilizing different feature subsets. Four real-world datasets and two databases were analyzed to test our developed model.

Our evaluation for the automatic classification model, for the datasets in Tables 3 and 4, encompasses accuracy as following:

- Model training accuracy range is 0:1.
- 25K sample data for PK, the accuracy is 89%.
- 2K sample data for FK, the accuracy is 82%.

There are no contributions for benchmarking our findings which indicate that our model demonstrates a capability to identify an average of 89% of PKs and 82% of FKs. Our evaluation for the later interactive classification model where the user manually selects or confirms PKs or FKs is increased, thus, PK prediction accuracy as of 92% and FK prediction accuracy as of 85%. Figures 6 and 7 show samples of predicted primary keys within the imported datasets. For instance, in Figure 6, CustomerID, InvoiceID, OrderID, and ProductID are detected as primary keys.

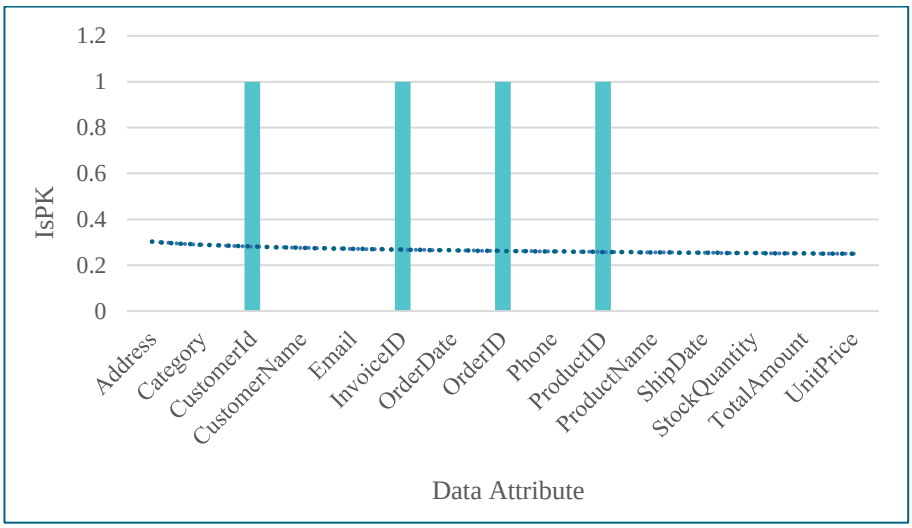


Figure 6: Sample I - modeled data – primary key with logarithmic trend line.

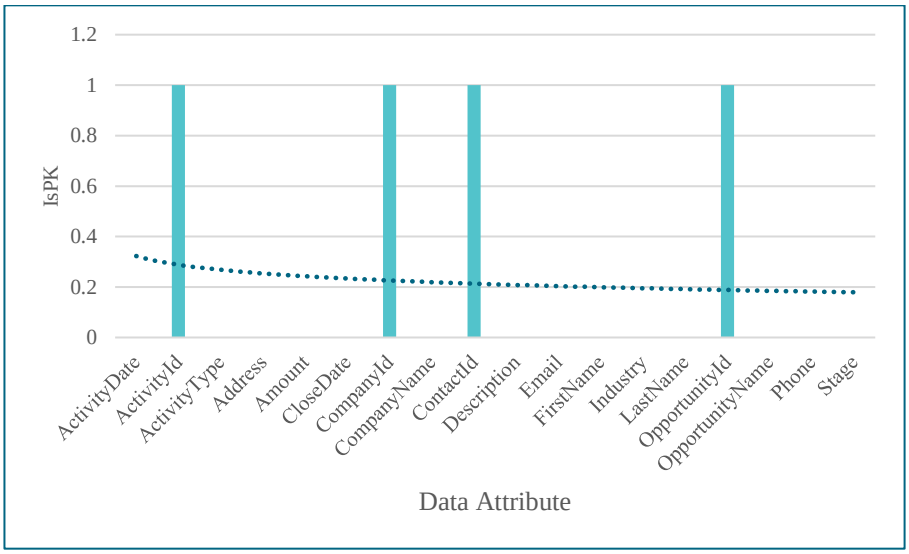


Figure 7: Sample II - modeled data – primary key with logarithmic trend line.

Limitation

In this section, we thoroughly analyze the restrictions and limitations that we faced throughout our investigation. By recognizing the following constraints, our objective is to establish a clear framework that enhances the reliability of our research and provides guidance for future studies in this area.

- Logistic regression postulates a linear association between the Table Name and Column Name as an input, and the logarithm of the probabilities of the PK or FK prediction as an output. Consequently, it is limited to capturing just linear decision boundaries. Therefore, our tool may exhibit lower performance when the underlying connection is nonlinear.
- Binary logistic regression classification assumes the independence of observations. If there is interdependence among observations, PDM may not be suitable.

- We assume that significant multicollinearity among the data almost does not exist as high correlation among Column Names and Instances as an input might result in unstable coefficient estimates, thus, hinder the interpretation of the model.
- Our tool performs well when applied to datasets with a substantial number of observations. Therefore, inadequate sample numbers might result in overfitting or untrustworthy estimations of the coefficients.

DISCUSSION AND CONCLUSION

Logistic regression is a highly adaptable and extensively employed statistical technique in the field of machine learning, specifically for applications involving binary classification. Its simplicity, interpretability, and efficiency make it a preferred option for a wide range of applications, including banking, healthcare, marketing, and more. Logistic regression is advantageous due to its capacity to estimate the chance of an event happening, offering a significant understanding of the connection between independent variables and the probability of a specific result. PDM introduces two types of binary classifications occurring at two levels. From the data relation perspective, the tool enables the analysis and discovery of primary and foreign keys. However, from the process automation one, PDM starts with the automated logistic regression modeling and provides an option if one wishes to continue with an additional step of review and conformance of primary or foreign keys classification.

Furthermore, binary classification is a key concept that PDM utilizes in the contribution of conceptual and logical data modeling provided. It involves categorizing the independent data into one of two distinct classes or outcomes: IsPK: Yes or No and IsFK: Yes or No. The two classes are commonly denoted as positive and negative, represented by the values 1 and 0, or referred to as true and false. The objective is to construct a prognostic model that can precisely allocate new data instances to one of these two categories using input features, noting that separate algorithms have been developed. The first is related to the primary key, where it takes the Table Name and Column Name as input, while the other is related to the foreign key, where it takes the Table Name, Column Name, Reference Table Name, and Reference Column Name as input.

The primary key functions as a method to distinctly identify individual records in the table, ensuring that no two records can have the same primary key. That said, PDM considers the primary key to be an essential attribute for preserving data integrity and establishing connections between tables in a relational database. Typically, a primary key is represented by an integer or a collection of columns. Therefore, our tool frequently searches for unique identifiers, such as auto-incrementing integers. Determining the primary key in raw data requires comprehending the data's arrangement and discerning patterns that signify a distinct identification for every entry. Finally, a foreign key is a data element or a collection of data elements within a database table that points to the primary key of another table; our tool builds a connection or association between the tables depending on the values they contain to be able to identify one. Its importance underlies the function of preserving referential integrity, guaranteeing the consistency and validity of relationships between tables. To conclude, PDM has a promising result of conceptual and logical data modeling through a comprehensive logistic regression binary classification.

References

[1]. Fatih Ertam and Galip Aydin, Data Classification with Deep Learning using Tensorflow, 2nd International Conference on Computer Science and Engineering, IEEE, 2017.

[2]. Hosmer and Lemeshow, Applied Logistic Regression, 2nd Ed., New York, Wiley, 2000.

[3]. Yan, Lee, and Koc, Predictive Algorithm for Machine Degradation Detection Using Logistic Regression, Fifth International Conference on Managing Innovations in Manufacturing (MIM), WI, 2002.

[4]. Graeme Simsion and Graham Witt, Data Modeling Essentials, Elsevier, 2005.

[5]. Charu C. Aggarwal, Data Mining-The Textbook, Springer, January 2015.

[6]. Maglogiannis, Karpouzis, Wallace, and Soldatos, Emerging Artificial Intelligence Applications in Computer Engineering, IOS Press, October 2007.

[7]. Stephan Dreiseitl and Lucila Machado, Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review, Journal of Biomedical Informatics 35, USA, 2002.

[8]. Ming-Syan Chen, Jiawei Han, and Philip S. Yu, Data Mining: An Overview from a Database Perspictive, IEEE Transaction on KD and Data Engineering, Vol 8, No 6, December 1996.

[9]. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[10]. UM Fayyad, G Piatetshy-Shapiro, P Smyth, and R Uthurusamy, Advances in Knowledge Discovery and Data Mining AAAI/MIT Press, 1996.

[11]. G. Piatetsky-Shapiro and W.J. Frawley, Knowledge Discovery in Databases, AAAI/MIT Press, 1991.

[12]. A. Silberschatz, M. Stonebraker, and J.D. Ullman, Database Research: Achievements and Opportunities into the 21st Century, Report, NSF Workshop Futuve of Database Systems Research, 1995.

[13]. G. D. Smith and S. Ebrahim, Data dredging, bias, or confounding, Bmj, vol. 325, no. 7378, pp. 1437-8, 2002.

[14]. Fayyad, Piatetsky-Shapiro, G. and Smyth, P. (1996), From data mining to knowledge discovery: an overview, Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA.

[15]. Technology Forecast: 1997 (1997), Price Waterhouse World Technology Center, Menlo Park, CA.

[16]. Weiss, S.H. and Indurkhya, N. (1998), Predictive Data Mining: A Practical Guide, Morgan Kaufmann Publishers, San Francisco, CA.

[17]. P. Gundecha and H. Liu, Mining Social Media: A Brief Introduction. Tutorials in operations research, 2012.

[18]. J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2016.

[19]. Rung-Ching Chen, Christine Dewi, Su-Wen Huang, and Rezzy Eko Caraka, Selecting Critical Features for Data Classification Based on Machine Learning Methods, Journal of Big Data, SpringerOpen, 2020.

[20]. M. J, Zaki and W. Meira Jr, Data Mining and Analysis-Fundamental Concepts and Algorithms, Cambridge University Press. New York, USA.

[21]. Mahmudah U, Suhartono, Fatimah S. A robust approach to analyzing the factors influencing quality education in Indonesia. New Educ Rev. 2017;49(3):77–90.

[22]. Cristobal Romero and Sebastian Ventura, Data mining in education. WIREs Data Mining Knowledge Discovery 2013, 3: 12–27.

[23]. Y. Lecun, C. Cortes, and C. J. C. Burges, The MNIST Database, Courant Institute, NYU, 2014.

[24]. Wang XD, Chen RC, Yan F, et al. Fast adaptive K-means subspace clustering for high-dimensional data, IEEE, 2019.

[25]. J. Han, M. Kamber and J. Pei, J Data Mining Concepts and Techniques (3rd ed.) Chapter 8, Pp. 99 117. University of Illinois. Illinois, 2010.

[26]. David Williams, Xuejun Liao, Ya Xue, and Lawrence Carin, Incomplete-Data Classification using Logistic Regression, Proceedings of the 22nd International Conference on Machine Learning, Germany, 2005.

[27]. Nasios and Bors, Variational Expectation Maximization Training for Gaussian Networks, Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, 2003.

[28]. Neal, Bayesian learning for neural networks, New York, Springer; 1996.

[29]. Ghahramani and Jordan, Supervised Learning from Incomplete Data via the EM Approach, Advances in Neural Information Processing Systems, CA, 1994.

[30]. Beal and Ghahramani, The variational Bayesian EM algorithm for incomplete data: application to scoring graphical model structures, Bayesian Statistics 7, 453-464, 2003.

[31]. Ripley, Pattern Recognition and Neural Networks, Cambridge, Cambridge University Press; 1996.

[32]. Duda, Hart and Stork, Pattern Classification, Wiley, New York, 2000.

[33]. Zan Yang and Dan Li, Application of Logistic Regression with Filter in Data Classification, IEEE, China, 2019.

[34]. Collins, Schapire, and Singer, Logistic Regression, AdaBoost and Bregman Distances, Machine Learning Journal, 2004.

[35]. Fletcher and Leyffer, Nonlinear Programming without a Penalty Function, Mathematical Programming, 2002.

[36]. Wächter and Biegler, Line Search Filter Methods for Nonlinear Programming: Motivation and Global Convergence, SIAM J. Comput., 2005.

[37]. Fortunato, Gilberto, Marcelo, and Mário, Data Classification with Binary Response through the Boosting Algorithm and Logistic Regression, Elsevier, Brazil, 2016.

[38]. Cao et al., The Boosting: A New Idea of Building Models, Chemometrics and Intelligent Laboratory Systems, 2010.

[39]. Cai, Using Logitboost Classifier to Predict Protein Structural Classes, Journal of Theoretical Biology, 2006.

[40]. Widodo and Yang, Support Vector Machine in Machine Condition Monitoring and Fault Diagnosis, Mechanical Systems and Signal Processing, 2007.

[41]. Jihong Yan and Jay Lee, Degradation Assessment and Fault Modes Classification Using Logistic Regression, NSF I/UCRC Center for Intelligent Maintenance Systems, WI, 2005.

[42]. Caffo B., Developing Data Products in R, R Software, 2015.

[43]. Yang JY, Wang JS, Chen YP., Using acceleration measurements for activity recognition: an effective learning algorithm for constructing neural classifiers, Pattern Recogn Lett. 2008.

[44]. Jawad Raza, Jayantha Liyanage, Hassan Al Atat, and Jay Lee, A Comparative Study of Maintenance Data Classification Based on Neural Networks, Logistic Regression, and Support Vector Machines, Journal of Quality in Maintenance Engineering, 2010.

[45]. Kang and Naughton, On Schema Matching with Opaque Column Names and Data Values, SIGMOD, 2003.

[46]. Koeller and Rundensteiner, Discovery of High-Dimensional Inclusion Dependencies, ICDE, 2003.

[47]. Lopes, Petit, and Toumani, Discovering Interesting Inclusion Dependencies: Application to Logical Database Tuning, Information Systems, 2002.

[48]. Meihui, Marios, Beng, Cecilia, and Divesh, On Multi-Column Foreign Key Discovery, ACM, 2010.

[49]. Leopoldo Bertossi, Database Repairs and Consistent Query Answering: Origins and Further Developments, PODS, ACM, 2019.

[50]. Jef Wijsen, Foundations of Query Answering on Inconsistent Databases, SIGMOD, ACM, 2019.

[51]. Miika Hannula and Jef Wijsen, A Dichotomy in Consistent Query Answering for Primary Keys and Unary Foreign Keys, Proceedings of the 41st ACM Symposium on Principles of Database Systems, 2022.

[52]. Mozhgan, Sebastian, and Gillian, SQL Data Profiling of Foreign Keys, Springer, 2015.

[53]. Vladimir Zykin, Automatization of Foreign Keys Construction, Dynamics of Systems, Mechanisms and Machines, IEEE, 2016.

[54]. Alexandra, Oliver, Jana, Felix, and Ulf, A Machine Learning Approach to Foreign Key Discovery, 12th International Workshop on the Web and Databases, 2009.

[55]. Richard J. Roiger, Data Mining: A Tutorial-Based Primer, Second Edition, 2017.

[56]. Stephen Gardner, Building the Data Warehouse, ACM, 1998.

[57]. Rizzi, Abello, Lechtenborger, and Trujillo, Research in Data Warehouse Modeling and Design: Dead or Alive? DOLAP, ACM, USA, 2006.

[58]. Website: https://www.kaggle.com/datasets, Kaggle.