Performance Evaluation and Accuracy Improvement in Individual Record Linking Problems Using Decision Tree Algorithm in Machine Learning

Vadood Keramati*
PhD Student; Payam Noor University; Tehran, Iran;
Email: keva_1981@outlook.com; keva1981@gmail.com

Ramin Sadeghian
Associate Professor; Department of Industrial Engineering;
Payam Noor University; Tehran, Iran Email: sadeghian@pnu.ac.ir

Maryam Hamedi
Assistant Professor; Department of Industrial Engineering;
Payam Noor University; Tehran, Iran Email: hamedi@pnu.ac.ir

Ashkan Shabbak
Associate Professor of Statistics Department; Research Institute of Statistics; Tehran, Iran Email: shabbak@srtc.ac.ir

Iranian Journal of Information Processing and Management

Received: 27, Oct. 2023 **Accepted: 02, Jun. 2024**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Special Issue | Winter 2025 | pp. 81-106
Exploring the Relationship Between Data Quality and Business Process Management
https://doi.org/10.22034/jipm.2024.2014515.1470

Abstract: Record linkage is vital for consolidating data from different sources, particularly in Persian records where diverse data structures and formats present challenges. To tackle these complexities, an expert system with decision tree algorithms is crucial for ensuring precise record linkage and data aggregation. Adaptation operations are created based on predefined rules by incorporating decision trees into an expert system framework, simplifying the aggregation of disparate data sources. This method surpasses traditional approaches like IF-THEN rules in effectiveness and ease of use and improves accessibility for non-technical users due to its intuitive nature. Integrating probabilistic record linkage results into the decision tree model within the expert system automates the linkage process, allowing users to customize string metrics and thresholds for optimal outcomes. The model's accuracy rate of over 95% on test data highlights its effectiveness in predicting and adjusting to data variations, confirming its reliability in various record linkage scenarios. The

* Corresponding Author

innovative utilization of machine learning decision trees alongside probabilistic record linkage in an expert system represents a significant advancement in the field, providing a robust solution for data aggregation in intricate environments and large-scale projects involving Persian records. Combining decision tree algorithms and probabilistic record linkage within an expert system offers a powerful tool for handling complex data integration tasks. This approach not only streamlines the process of consolidating diverse data sources but also enhances the accuracy and efficiency of record linkage operations By leveraging machine learning techniques and automated decision-making processes, organizations can achieve significant improvements in data quality and consistency, paving the way for more reliable and insightful analytical results in implementing statistical registers. In conclusion, integrating decision trees and probabilistic record linkage in an expert system represents a cutting-edge solution for addressing data aggregation challenges in Persian records and beyond.

Keywords: Machine Learning, Record Linkage, Decision Tree, Performance Evaluation

1. Introduction

The increase in the production of administrative data has enabled organizations to leverage this information to gain new insights into issues such as population statistics. Statistical errors, like non-response and measurement errors, have driven organizations to leverage existing administrative data to acquire comprehensive lists of target objects and generate pertinent statistics. Despite the advantages of this data, the multitude of sources in similar fields has made it challenging to identify a single primary reference for establishing the desired framework.

The main concern in this field is the abundance of diverse data sources with varying coverage, which can confuse decision-makers when choosing sources. The usual way to handle multiple sources is to use data record-matching techniques. These traditional methods involve certain and possible adjustments in various situations while dealing with noise. The advancement of machine learning and artificial intelligence has enhanced tools in the data science field, including record matching. A key application of machine learning techniques is enhancing data record matching from various sources, which is the primary focus of this article, particularly concerning Persian datasets.

This article uses an expert system with a powerful and dynamic knowledge

base to enhance performance and accuracy in matching individual data using the decision tree algorithm in machine learning for Persian datasets. With advancing technology and the accumulation of big data, ensuring personal data compliance has become a crucial and sensitive issue. Enhancing performance and accuracy in matching individual data can have a wide range of applications in fields such as healthcare, disease diagnosis, user classification, and order recommendations. Utilizing the decision tree algorithm enables the rapid and effective construction of predictive models for individual data matching. The decision tree algorithm, a potent method in machine learning, categorizes data into various groups based on the decisions made at each node and branch of the tree. Employing this algorithm can significantly enhance the performance and accuracy of machine learning models for individual data matching.

In this article, the decision tree method is utilized as a potent tool to adjust the probability record of Persian data. This issue is important due to the unique challenges of the Persian language, such as the absence of word order and complex grammatical structures. This issue is important due to the unique challenges of the Persian language, such as the lack of word order and complex grammatical structures, which have been tried to be solved using a knowledge base and the use of string metrics. It also functions as an automated evaluation tool, enabling accurate and comprehensive analysis.

Compared to previous studies, this article distinguishes itself in three key areas. Firstly, the application of machine learning algorithms in the processing and analysis of Persian data requires specialized methods.

Secondly, the introduction of automated evaluation is a novel approach in this domain. Lastly, the creation and execution of an expert system to consolidate related processes represent another distinctive aspect of this research compared to previous works. Also, this system can automatically learn the adaptation rules from the training data and efficiently use them to adapt the new data.

In essence, this article represents a significant advancement in the use of decision trees for Persian data from both a technical and innovative perspective, with these distinctions holding considerable importance.

From a technical and innovative perspective, this article represents significant progress in utilizing decision trees for Persian data, with these differences being

highly important. In some Persian studies, machine learning methods have been employed, focusing solely on the numerical aspects of the data. This study, however, diverges by linking Persian text data.

2. Literature Review

Some studies investigate the combination of decision trees with other methods to enhance performance, particularly in big data scenarios. Fattoum et al. (2020) introduced a hybrid method for identifying duplicates in large datasets. Their approach integrates blocking techniques to pre-filter similar records based on specific criteria, thereby reducing the search space for the decision tree classifier. This dual-stage procedure aims to achieve high accuracy while maintaining efficiency in processing extensive datasets. The researchers evaluated their technique on a massive dataset and demonstrated its effectiveness in terms of both accuracy and efficiency.

Su et al. (2021) proposed a machine-learning model for chronic disease datasets utilizing an integrated attribute evaluator and an improved decision tree classifier. The model aimed to achieve high accuracy while reducing processing time. The evaluation focused on heart disease, diabetes, and breast cancer datasets. Results showed that the decision tree with the integrated evaluator achieved promising performance compared to traditional methods.

Ensuring the quality of record linkages generated by decision trees is crucial. Wang et al. (2022) addressed this challenge by proposing a framework for assessing link quality specifically for decision tree-based methods. Their framework analyzes the decision tree structure and data characteristics to identify potential issues with the linkage results. This evaluation process helps guide further refinements and improve the overall accuracy and reliability of the record linkage task.

However, traditional decision trees may not account for the varied costs associated with different error types in record linkage. For instance, mistakenly linking two unrelated records (false positive) could result in a higher cost than missing a true match (false negative). Zhang et al. (2021) addressed this issue by introducing a cost-sensitive decision tree learning algorithm designed for record linkage tasks. This approach incorporates the costs linked to different errors into

the decision tree learning process. By optimizing the decision tree based on these costs, the model aims to improve the overall performance of the record linkage task.

While interpretability is a key advantage of decision trees, recent research has concentrated on enhancing explainability even further. Li et al. (2021) introduced an explainable decision tree framework specifically designed for record linkage. This framework not only facilitates linkage decisions but also offers insights into the reasoning behind those decisions, such as the significance of various attributes (feature importance analysis). This degree of explainability is crucial for comprehending how the decision tree arrives at its conclusions and can help build trust in the results, especially in critical applications.

However, a key element of decision tree-based techniques is establishing suitable thresholds for decision-making on linkages. Incorrect thresholds may result in imprecise outcomes, either by failing to identify true matches (false negatives) or by erroneously connecting unassociated records (false positives). Li et al. (2023, In Press) tackle this issue by suggesting a versatile method for setting thresholds in decision tree-based record linkage. This method takes into account data attributes and user inclinations to adapt the threshold dynamically, potentially enhancing matching precision. Such adaptability can prove especially advantageous in situations where data integrity or user preferences might differ.

Researchers are continually exploring methods to improve the performance of decision trees. Li et al. (2022) proposed an innovative approach that integrates active learning with cost-sensitive learning. Active learning allows the decision tree to focus on critical data points, thereby increasing accuracy while utilizing less training data. Cost-sensitive learning takes into account the varying costs associated with errors, such as failing to identify a true match or incorrectly linking unrelated records. The objective of combining these techniques is to enhance the performance of decision tree-based record linkage.

Jiang et al. (2021) investigated a method that combines blocking with a phonetic encoding technique known as Metaphone. Blocking filters out similar records based on specific criteria, thereby reducing the search area for the record linkage algorithm. Metaphone phonetically encodes names, improving matching accuracy by identifying spelling variations. This method is particularly beneficial

for public health data, as even minor inconsistencies in names can lead to missed linkages.

Zhang et al. (2020) explored the use of decision trees for record linkage in web data. They recognized the difficulties presented by web data's unique traits and suggested a dual approach. Their technique employs decision trees for matching logic and integrates blocking methods to filter out similar records beforehand. This sequential method strives to enhance efficiency and precision in handling extensive and noisy web datasets.

Smith et al. (2020) investigated various record linkage methods, emphasizing their advantages and drawbacks. They evaluated probabilistic, deterministic, and machine learning techniques, analyzing their precision, recall, and F-measure. Their findings indicated that probabilistic methods like the Fellegi-Sunter method excelled in scenarios with significant heterogeneity and data noise, whereas machine learning approaches, such as deep learning algorithms, delivered superior accuracy in well-organized datasets.

In a recent study by Li et al. (2021), researchers explored the use of natural language processing (NLP) techniques for record linkage. They investigated extracting and matching unstructured text data, like clinical notes and discharge summaries, from various sources. The study revealed that integrating NLP methods enhanced the accuracy and efficiency of record linkage, especially with textual data.

Decision trees have been used in different real-world scenarios, such as medical diagnosis, finance, and natural language processing. In the medical field, decision trees have demonstrated potential in diagnosing illnesses using symptoms and patient information, offering doctors valuable decision-making support (Rokach et al., 2008).

Moreover, explanations produced from decision tree models can be crucial in fields where transparency and interpretability are important, such as legal and ethical considerations (Guidotti et al., 2019). Consequently, decision trees remain preferred in situations where interpretability, explainability, and human-understandable reasoning are vital factors.

Técnicas de aprendizaje automático, incluidos los árboles de decisión, también se han utilizado en escenarios avanzados de vinculación de registros.

Por ejemplo, Zhang et al. (2019) propusieron un enfoque de vinculación de registros basado en aprendizaje automático que combinaba árboles de decisión con otros algoritmos de clasificación. Aplicaron este enfoque para vincular registros de salud electrónicos de múltiples proveedores de atención médica. El árbol de decisión desempeñó un papel crucial en determinar la importancia de las características y los criterios de división, lo que llevó a una mayor precisión en la vinculación de registros de pacientes. El estudio demostró que la integración de árboles de decisión dentro de marcos de aprendizaje automático puede mejorar el rendimiento de las tareas de vinculación de registros en conjuntos de datos complejos y diversos.

3. Evaluation Criteria

Any matching procedure should aspire to three important criteria: it should be efficient, accurate, and unbiased. I define these terms in the record linkage context:

◇ Efficient: A high share of the records to be searched for are found and matched. The match rate will naturally vary across applications and source or target databases, but generally, a procedure that requires thousands of records to match only a handful would be quite inefficient and not very useful for econometric analysis. An efficient match process will have a low share of type I errors. In the machine learning context, one measure of efficiency is the true positive rate or TPR. This records the ratio of true positives with the total number of positives:

$$1)\ PR\ = \frac{TP}{TP+FN}$$

◇ Accurate: A high share of the records matched are true matches and not false positives. Ideally, this rate would be close to 100%, but naturally the higher the bar for declaring two records matched, the less efficient it will be. An accurate match process will have a low share of type II errors. In machine learning, accuracy could be measured with the positive predictive value or PPV. This measures the ratio of the true positives to all of the records identified as matches by the algorithm:

$$2)\ PPV\ = \frac{TP}{TP+FP}$$

◇ Unbiased: A matching procedure will generate a dataset for downstream analysis. To what extent is this final dataset representative of the records that the researcher attempted to link in the first place? Improvements in efficiency and accuracy will necessarily decrease the bias in the resulting dataset. However non-random variation in either error rate will generate bias. One manifestation of bias would be an unrepresentative linked sample. Using spouse names to create links, for example, would increase the match rate among married people and over-represent them in the final analysis; similarly matching on county or state of residence would bias against including interstate migrants in the sample (James. 2016)

In Table 1, we present a detailed comparison of various splitting criteria employed in machine learning methods for record linkage. This analysis aims to provide a comprehensive understanding of the different approaches utilized in this field. By evaluating the performance and characteristics of each criterion, we can gain valuable insights into their strengths and weaknesses. This information serves as a reliable reference for researchers and practitioners seeking to make informed decisions when selecting the most appropriate splitting criteria for their record linkage tasks.

To evaluate the learning methods employed in record linking and to determine the appropriateness of the selection method used in this research, we conducted a brief review of previous studies, the results of which are presented in Table 1.

The table compares various splitting criteria commonly used in machine learning methods for record linkage. It includes details such as the name of the method used, splitting criteria, improvements, and limitations. This comparative analysis aims to clarify the differences between these criteria, enabling us to make informed decisions when selecting the most suitable approach for specific record linkage tasks.

For this research, we aimed to identify the most effective machine learning method for record linking by reviewing the results of previous studies. Based on our findings and the objectives of this research, we determined that the decision tree method is the most suitable approach. The data presented in this table indicate that the fields utilized in this study require blocking, and the majority of these fields necessitate binary comparison. Given the results obtained, we have

identified a viable method to assess the quality of the outcomes, leading us to choose the decision tree method. The research that employed the decision tree approach demonstrated greater alignment with the goals of this study and could significantly contribute to its success.

Table 1. COMPARISONS BETWEEN DIFFERENT SPLITTING CRITERIA (Sheth, et al., 2015)

No	Method Used	Splitting Criteria	Improvements	Limitations
1	hybrid machine learning	Gain Splitting with Pre-Pruning	Higher Classification Accuracy	Data Dependency
Ref: Su, S., Xiao, Y., & Wang, H. (2021)				
2	combines blocking and decision tree techniques	based on similarity features between pairs of records	Scalability to Big Data	Error Propagation
Fattoum, N., Issaoui, D-E., & Moussaoui, M. A. (2020)				
3	utilizes the decision tree's structure and node information	based on similarity features between records	Enhanced Linkage Confidence	Parameter Sensitivity
Wang, J., Pei, J., & Zhang, Y. (2022)				
4	A cost-sensitive decision tree	based on both information gain and misclassification costs	Adaptability to Cost Structures	Computational Cost
Zhang, W., Fan, X., & Wu, X. (2021)				
5	active learning approach	Medical Records	Reduced Labeling Effort	Query Selection Strategy
Li, J., Zhu, Y., & Wang, H. (2022, April)				
6	ensemble learning approach	Information Gain	Reduced Error Rates	Parameter Sensitivity
Jiang, N., Desruisseaux, L., & Swanson, D. A. (2021)				

4. Datasets

For this system, which is for Persian characters for record linkage issues using

machine learning, we need input files must have a .csv extension and there is no limit on the number of records. Although the number and specifications of the fields can be defined by the user, in this system for the test phase, their number is fixed and 8 fields are considered. The structure of these input files is as follows:

Table2 . Data set fields and their types used for link and decision tree making

Field Name	Name	Family	FatherName	Sex	DateOfBirth	NationalCode	Province	PostalCode
Type	Text	Text	Text	Text	Date	Text	Text	Text

After loading the information, the system performs cleaning operations on the input data. This cleaning of the rules provided in the system knowledge base section is done which which are described below:

- ◇ Convert all Arabic letters and numbers to Farsi
- ◇ Removing any non-numeric characters from the national code (except the letter o which may be inserted instead of zero)
- ◇ Convert the letter o to zero if it exists in the national number
- ◇ Adding one or two zeros to the beginning of the national number if the length of the national number is less than 10
- ◇ Convert the character “–” to the character “/” in the date of birth
- ◇ Convert the name of the month to a number; like April to “01”
- ◇ Checking the authenticity of the national number (compliance with the national number algorithm) and removing the wrong ones

For generating data for this research, we used a web scrap solution that is explained comprehensively in section 5-1.

Some examples of the records utilized in this study are shown in Table 3.

Table 3. Some example data

No	Name	Family	F_Name	Sex	DateOfBirth	NationalCode	Province	PostalCode
۱	محمد	احمدی	علی	مرد	۱۳۷۰/۰۵/۲۰	۰۰۱۲۳۴۵۶۷۸	تهران	۱۲۳۴۵۶۷۸۹۰
۲	سارا	رضایی	محمد	زن	۱۳۶۵/۱۲/۱۰	۰۰۹۸۷۶۵۴۳۲	اصفهان	۵۴۳۲۱۶۷۸۹۰
۳	علی	خاندانی	حسین	مرد	۱۳۷۵/۰۸/۳۰	۰۰۵۶۷۸۱۲۳۴	خراسان جنوبی	۹۸۷۶۵۴۳۲۱۰
٤	مریم احمدی	---	حسین	مرد	۱۳۷۲/۰۹/۳۰	۰۹۸۷۶۵۴۳۲۱	اصفهان	۵۴۳۲۱
٥	حسن	رضا	محمد		۱۳۶۰	۱۲۳۴۵۶۷۸۹۰	طهران	۱۲۳۴۵
٦	علی کرینی	----	احمد	مرد	۱۳۷۸/۰۴/۲۵	۵۶۷۸۱۲۳۴	یزد	۹۸۷۶۵۴۳۲
۷	حسن	حسن‌زاده	یونس			۰۰۹۸۷۶۵۴۳۲	اردبیل	-----
۸	حسین	حسین‌زاده	یوسف	مرد	۱۳۹۳/۰۳/۰۳	۱۱۱۱۱۱۱۱۱	تهران	-----

5. Suggested Method

To enhance the accuracy and efficiency of record linkage, a decision tree approach was employed. Decision trees are a widely used machine learning method that utilizes a tree-like model to represent decisions and their potential outcomes. In this case, a decision tree was constructed using a dataset comprising 30 to 70 training and testing data points. The decision tree algorithm analyzed features from the probabilistic record linkage and identified patterns in the data to determine whether two records should be linked. The tree was trained on the training data to understand the relationships between input features and target labels (linked or non-linked status). Subsequently, the trained tree was evaluated on the testing data to assess its performance and generalization capabilities.

To calculate distance in this case, we utilized the Jaro-Winkler metric. The Jaro-Winkler metric quantifies the similarity between two sequences or strings. This metric is widely employed in data mining and record matching due to its effectiveness in matching names (Wang et al., 2017). A higher Jaro-Winkler score indicates a greater likelihood of similarity between the two strings. The score ranges from 0, indicating dissimilarity, to 1, indicating similarity.

The main components of the sweep algorithm are:

◇ Calculate string length.

◇ Number of common characters in two strings.

◇ Counting displacements.

where the common definition means that the matching character must be present within half the length of the shorter string (Winkler et al, 2006). The string comparator value is obtained from the following relation:

$$3)\ \phi(s_1, s_2) = \frac{1}{3}\left(\frac{N_C}{len_{s_1}} + \frac{N_C}{len_{s_2}} + \frac{0.5N_t}{N_C}\right)$$

That:
s_1, s_2: are strings with length len_{s_1} and len_{s_2}

N_C: The number of common characters between two strings so that the distance for common characters is half of the minimum length of s_1, s_2.

N_t: It is a displacement count that is calculated somewhat differently from the obvious method.

When conducting record linkage, we compare combinations of records from both datasets, resulting in a dataset of pairs accompanied by comparison vectors. The objective is to categorize these pairs into two groups: one containing pairs where both records pertain to the same object (the matching set) and another containing pairs where the records do not pertain to the same object (the unmatched set). Record linkage is fundamentally a classification challenge; by determining the classification of certain pairs, we can train a supervised classification model.

Utilizing a decision tree significantly improved the record linkage process by automating and streamlining it. The decision tree made choices based on learned patterns, which reduced reliance on manual rules or heuristics. This approach enhanced accuracy and minimized the need for manual intervention, ultimately saving time and effort.

In the proposed system of this study, a decision tree is utilized for record matching. The potential record-matching outcomes are employed as input data for machine learning training and testing. The process is outlined below:

1. First, two datasets X and Y are selected, each containing n and m records respectively, with specific fields such as national number, name, surname, father's name, gender, province, date of birth, and postal code.
2. These two datasets are then provided as input to the system.
3. Next, the user must determine the following items for the system: Threshold of the probability of linkage to the record (a number between 0 and 1)
4. Minimum number of linkage fields
5. String metric method (Winkler, Lonstein, Jaro, and …)
6. Determining the blocking variables for classification (the default three variables are national number, gender, and province)
7. Determining the percentage of test data to evaluate and calculate the F-Measure
8. A Cartesian multiplication is performed between two matrices X and Y
9. Probabilistic record matching is done based on the provided inputs according to the metric method and the obtained distance
10. The output of this step is produced in the form of an n*m matrix, which is a binary matrix. From matching the possible records with the above conditions, 1 indicates a match and 0 indicates a non-match of two records i and j from datasets X and Y.
11. The matrix of step 10 is selected as labeled data
12. Next, the two matrices X and Y are multiplied again in a Cartesian manner and the cells of the matrix are calculated based on the specified metric.
13. By selecting 70% training data and 30% test data, a decision tree is made using the results of the two matrices of the previous steps.
14. At the end, the evaluation values of the tree, the final tree, and the accuracy of the tree are presented
15. This article uses the Jaro-winemaker metric method for both steps.

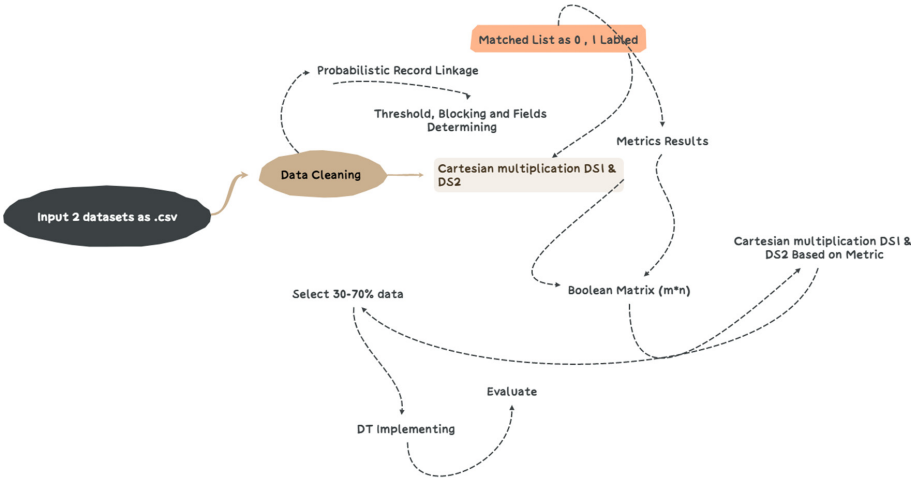

Figure 1. The research process generally flows

5-1. Extracting Possible Matches

In this study, we have created two datasets using web scraping methods. Web scraping involves extracting data from websites and serves as an effective means of gathering information from various online sources. By employing this technique, we have obtained datasets that will serve as the foundation for your record linkage research. This method enables us to collect real-world data from online sources, providing a comprehensive and current view of the information essential for your study. Our use of web scraping demonstrates our commitment to acquiring high-quality, relevant data for your research, facilitating in-depth analysis in the field of record linkage. Through this approach, we can acquire and organize data that may not be accessible through traditional means, thereby enhancing the scope and impact of our work. Due to the confidentiality of personal data and limited access to relevant institutions, and in compliance with the specifications of the information field in this study, all dataset records have been gathered and modeled using web scraping techniques that accurately reflect real data. By utilizing web scraping, we have produced two datasets tailored to your research, containing specific variables crucial for our investigation. Web scraping offers researchers numerous advantages. Firstly, it allows us to swiftly and efficiently gather large volumes of data, which is particularly beneficial when dealing with extensive information

that would be laborious or time-consuming to collect manually. Furthermore, web scraping helps overcome the limitations posed by traditional data sources, as online data is often more comprehensive and up-to-date. Reason: Improved clarity, vocabulary, and technical accuracy while maintaining the original meaning.

First, probabilistic record linkage was employed to evaluate the likelihood of matching pairs of records based on similarities such as name, age, gender, and address (as detailed in the Datasets section). Two datasets in .csv format were prepared for this purpose: one containing 80,000 records and the other containing 70,000 records. These datasets were extracted randomly and with noise from a centralized database of 100,000 records (generated by the provided model) and then input into the system for processing. Ideally, accurately linking the two datasets should yield a total of 100,000 records; any deviation from this number indicates errors. The primary objective of this study was to develop a model capable of verifying its output against a specific benchmark.

The model utilized probability distributions of match and non-match pairs to assess the likelihood of a given pair being a true match. The resulting scores were then used to identify potential matches that exceeded a predetermined threshold. In this process, the records from two datasets were compared in a Cartesian manner (X1 × X2). Probabilistic record linkage was implemented using national and provincial number blocking, as well as gender. The Jaro-Winkler metric was employed to calculate distances, enabling the differentiation between matches, non-matches, and possible matches. A blocking method was applied to streamline the comparison process by utilizing the PIN, gender, and province fields.

The next step involved the use of decision trees. These decision trees were created to identify potential matches among records. The algorithm analyzed features derived from probabilistic record linkage and recognized patterns in the data to ascertain whether two records should be linked. Initially, the algorithm examined features from the probabilistic record linkage method, including demographic details such as name, age, gender, occupation, address, and other relevant attributes.

The decision tree algorithm operates by recursively dividing the dataset based on various feature values. At each stage, it selects the optimal feature to split the data by evaluating metrics such as entropy or Gini impurity, which measure

the homogeneity or purity of the resulting subsets. As the decision tree grows, it identifies patterns in the data that help distinguish between matched and non-matched records. For example, it may learn that individuals with similar ages and addresses are more likely to be the same person. Once the decision tree is trained, it can determine whether two records should be linked based on the identified patterns. This process involves navigating the tree from the root to a leaf node, following the path defined by the feature values. At the leaf node, a decision or prediction is made regarding the record linkage. One benefit of using decision trees is their ability to provide interpretable outcomes. The tree structure allows for straightforward visualization and comprehension of the decision-making process. Moreover, decision trees can accommodate both categorical and numerical features, making them suitable for various datasets. In Figure 2, you can observe the comparison result example for two datasets using a 60% threshold for Jaro-Winkler.

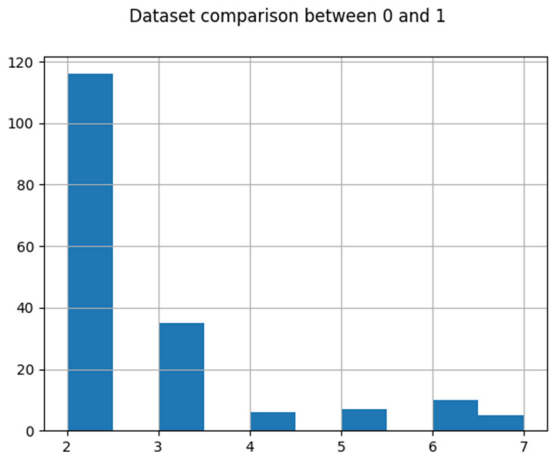

Figure 2. Dataset Comparison between two datasets for each number of comparison fields

5-2. Creating Decision Tree

To construct a decision tree, the dataset was divided into training and testing sets, comprising 70% and 30% of the total records, respectively. The training set was used to train the decision tree using the C4.5 algorithm, which generated decision rules based on the input features. Subsequently, the trained decision tree was

evaluated on the testing set to assess its ability to identify matches. In a record linkage scenario, a decision tree can categorize records into matches and non-matches. Initially, the tree may split at a critical value of the Family Name string distance at the first node and a critical value of the last name distance at the second node. Subsequent nodes could further divide based on matching first or last name Soundex results, agreement of middle initials, and other criteria. Figure 3 illustrates a potential decision tree with two levels. At the first node, records are segregated based on the Jaro-Winkler string distance in the last name. The descriptions of each node are as follows:

NationalCode

- ◇ NationalCode <= 0.969 indicates that pairs with a Jaro-Winkler distance of 0.969 or lower will be directed to the True path (left), while the rest will go to the False path (right).
- ◇ gini = 0.002 signifies the split quality, ranging between 0.0 and 0.5. A value of 0.0 implies all samples yield the same result, while 0.5 signifies an exact middle split.
- ◇ samples = 166019 denotes 166019 pair matches remaining at this decision point, encompassing all pairs in this initial stage.
- ◇ value = [165865, 154] shows that out of 165865 pair records, 165865 will be labeled “NO,” and 154 will be labeled “Yes.”

Gini

There are several ways to divide the samples, in this tutorial, we will utilize the GINI method. The Gini method employs the following formula:

$$4) \quad Gini = 1 - (x/n)^2 - (y/n)^2$$

Where x represents the number of positive answers (“Yes”), n represents the number of samples, and y represents the number of negative answers (“No”), the calculation is as follows:

$$5) \quad 1 - (154 / 166019)^2 - (165865 / 166019)^2 = 0.002$$

Family

Family<= 0.73 indicates that when the Jaro-Winkler distance between two records

is 0.73 or lower, they will be directed to the True path (left), while others will go to the False path (right).

- ◇ gini = 0.475 signifies the split quality, ranging from 0.0 to 0.5. A value of 0.0 implies all samples yield the same outcome, whereas 0.5 indicates an exact middle split.
- ◇ samples = 252 implies 252 pair matches remain at this decision point, encompassing all pairs as this marks the initial stage.
- ◇ value = [98, 154] shows that out of 165865 pair records, 98 are classified as "NO" and 154 as "Yes".

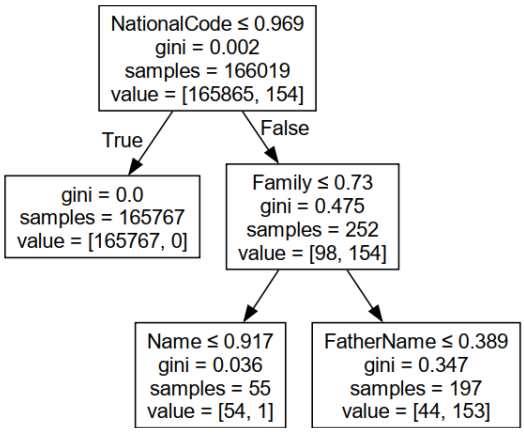

Figure 3. Two levels decision tree results

In the same manner, nodes are created within the tree and should be extended until reaching the tree's conclusion as illustrated in Figure 3. No pair of records is left unassigned by the end. Figure 4 shows the complete tree.

5-3. Evaluation

The decision tree evaluation revealed high accuracy in matching records. It attained an F-measure of approximately 0.97, indicating minimal false positives and false negatives. This showcases the algorithm's efficiency in automating record linkage, reducing manual intervention and saving time. The confusion matrix is as follows:

TP: 48880, FN: 1120, FP: 352, TN: 49648

$$6)\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{98528}{100.000} = 0.98528$$

$$7)\ Recall = \frac{TP}{TP + FN} = \frac{48880}{48880 + 1120} = 0.9776$$

$$8)\ Precision = \frac{TP}{TP + FP} = \frac{48880}{48880 + 352} = 0.9928$$

In summary, this study demonstrated the efficacy of a machine-learning approach for record linkage in census data. The utilization of probabilistic record linkage and decision tree-based linkage facilitated the development of accurate and efficient models for record matching. These models serve as valuable resources for researchers and organizations seeking reliable and automated record linkage techniques. The results of this study highlight the potential of machine learning algorithms to improve the accuracy and efficiency of record linkage processes in large datasets. By leveraging advanced techniques such as probabilistic record linkage and decision trees, researchers can streamline the matching of records from different sources, ultimately leading to more reliable data integration and analysis. Moving forward, further research in this area could explore the application of other machine learning algorithms and data preprocessing techniques to enhance the performance of record linkage models even further.

Some results related to this system that are linked with different thresholds for two datasets 1 and 2 are as follows:

Table 4. Some example results

DS	Name	Family	F_Name	Sex	DateOfBirth	NationalCode	Province	PostalCode	Threshold	Result
۱	سارا	محمدی	علی	زن	۱۳۵۹/۰۴/۱۲	۱۲۳۴۵۶۷۸۹۰	اصفهان	۱۳۱۴۶۷۶۵۹۸	٪۹۰	Linked
۲	سارا	احمدی	علیرضا	زن	۱۳۵۹/۰۴/۲۲	۱۲۳۴۵۶۷۸۹۰	اصفهان	۴۳۱۲۶۷۶۱۱۸		
۱	علی	احمدی	محمد	مرد	۱۳۷۰/۰۵/۲۰	۰۰۱۲۳۴۵۶۷۸	تهران	۱۲۳۴۵۶۷۸۹۰	٪۸۰	Not-Linked
۲	محمد	علوی	علی	مرد	۱۳۷۰	۱۲۳۴۵۶۷۸۹۰	تهران	۱۲۹۸۳۴۵۶۷۸		

DS	Name	Family	F_Name	Sex	DateOfBirth	NationalCode	Province	PostalCode	Threshold	Result
۱	پوریا	نیک‌پی	احمد	مرد	۱۳۷۸/۰۴/۲۵	۰۹۸۷۶۵۴۳۲۱	خراسان شمالی	۹۸۷۶۵۴۳۲	٪۷۰	Not-Linked
۲	کیان	نیک‌پور	محمود	مرد	۱۳۷۸/۰۴/۲۵	۰۹۸۷۶۵۴۳۲۱	خراسان‌رضوی	۵۴۳۲۱۶۷۸۹۰		
۱	حسن	حسن‌زاده	یونس			۰۹۸۷۶۵۴۳۲۱	تهران		٪۷۰	Linked
۲	حسین	حسین‌زاده	یوسف	مرد	۱۳۹۳/۰۳/۰۳	۱۱۱۱۱۱۱۱۱۱	تهران			

Table 5 has been prepared to compare this research approach with others for linking two datasets. A reference to this table can be made to facilitate a better comparison between the two methods.

Table 5. Short comparison between this research method and other researchs

Research	Knowledge Base	Compare any Thresholds and metrics	Different Fields	Automatic Evaluation	Flexible Blocking	Suitable for Persian Records
Su, S., Xiao, Y., & Wang, H. (2021)	✗	✓	✗	✗	✗	✗
Fattoum, N., Issaoui, D-E., & Moussaoui, M. A. (2020)	✓	✓	✗	✗	✓	✗
Wang, J., Pei, J., & Zhang, Y. (2022)	✓	✓	✓	✗	✗	✗
Zhang, W., Fan, X., & Wu, X. (2021)	✗	✓	✗	✗	✗	✗
Li, J., Zhu, Y., & Wang, H. (2022, April)	✗	✓	✓	✗	✗	✗
Li, J., Zhu, Y., & Wang, H. (2021)	✗	✗	✓	✗	✗	✗
Li, J., Zhu, Y., & Wang, H. (2023)	✗	✓	✓	✗	✗	✗
Li, J., Zhu, Y., & Wang, H. (2022)	✗	✓	✗	✗	✗	✗
Jiang, N., Desruisseaux, L., & Swanson, D. A. (2021)	✓	✓	✓	✗	✓	✗
Zhang, J., Dong, X., & Sun, A. (2021)	✗	✗	✓	✗	✗	✗

Research	Knowledge Base	Compare any Thresholds and metrics	Different Fields	Automatic Evaluation	Flexible Blocking	Suitable for Persian Records
Zhang, W., Fan, X., & Wu, X. (2020)	✗	✓	✓	✗	✓	✗
This Research	✓	✓	✓	✓	✓	✓

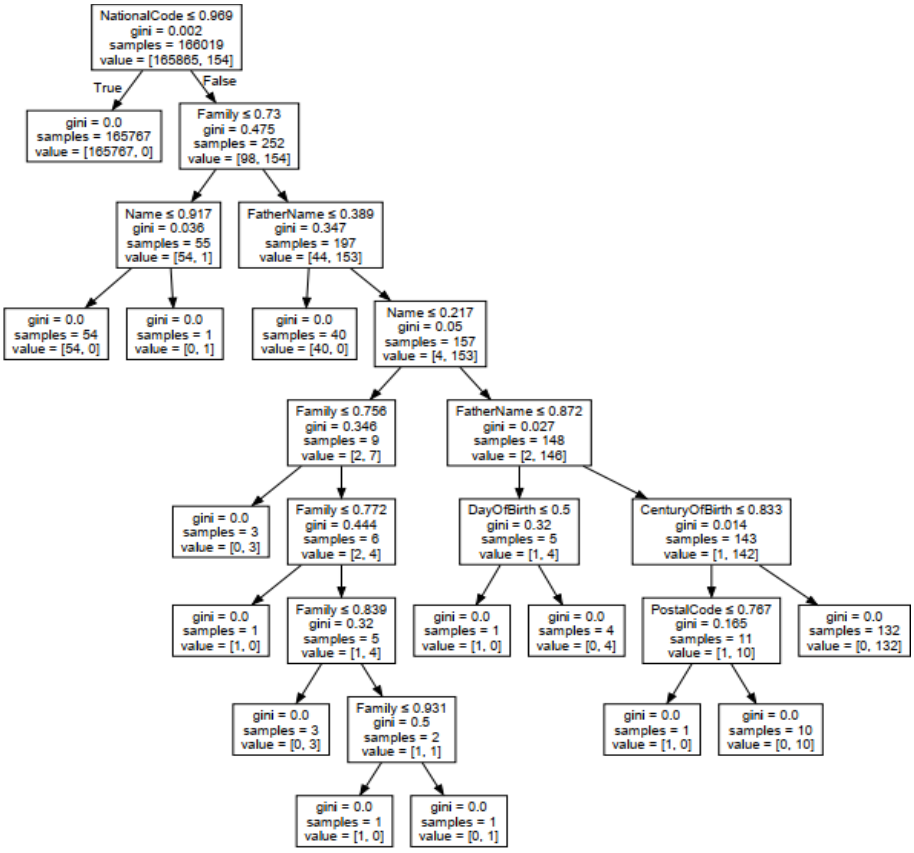

Figure 4. Decision tree results

6. Conclusion

Based on previous research findings on record linkage and the scarcity of studies on Persian record linkage involving string fields, we opted to employ the decision tree method within a machine learning framework using an expert system to enhance precision. Leveraging the outcomes of probabilistic record linkage, we constructed a decision tree utilizing 70% of the training data and 30% of the test data, incorporating a robust knowledge base for Persian data refinement. In this study, we utilized two data sets comprising 8 fields.

According to the results presented in this research, it can be seen that the decision tree works with very high accuracy in performing the act of recognition and matching and separating the records from the concordance. The values obtained for Accuracy, Precision, and Recall, all of which are above 95%, indicate the good performance of this wash for matching data records related to persons in terms of 7 information fields presented in two files. The values obtained for Recall, Precision, and Accuracy, all of which are above 95%, indicate the good performance of this wash for matching data records related to persons in terms of 7 information fields presented in two files.

Therefore, it can be inferred that utilizing the decision tree in this approach is beneficial for enhancing the precision and effectiveness of the process. For additional exploration and study, different metric techniques can be utilized in probabilistic record linkage and decision trees, and the most suitable approach can be selected according to pertinent criteria.

In conclusion, the decision tree algorithm in machine learning has proven to be effective in improving accuracy and performance evaluation in individual record-linking problems. By harnessing the power of this algorithm, researchers and practitioners can achieve more precise and reliable results in linking individual records. This study highlights the importance of leveraging advanced machine learning techniques to address complex data linkage challenges and underscores the potential for further advancements in this field.

Based on the results of this article, here are some future research recommendations:

1. Comparative study with other machine learning algorithms: Evaluating the decision tree algorithm's performance against popular algorithms like Random

Forest, Support Vector Machines, or Neural Networks in individual record linking problems.

2. Impact of feature engineering techniques: Studying how different techniques affect the accuracy of individual record linking using a decision tree algorithm, including feature scaling, selection, and transformation.
3. Ensemble methods exploration: Investigating the use of Bagging or Boosting with the decision tree algorithm to enhance accuracy in individual record linking tasks.
4. Hyperparameter optimization: Analyzing hyperparameters in the decision tree algorithm thoroughly to enhance performance in individual record linking problems, utilizing techniques like grid search, random search, or Bayesian optimization.

References

Fattoum, N., Issaoui, D.-E., & Moussaoui, M. A. (2020, January 28). A hybrid approach for duplicate detection in big data using blocking and decision tree [arXiv]. arXiv. https://doi.org/10.48550/arXiv.2001.08012

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 93. https://doi.org/10.1145/3287560

Jiang, N., Desruisseaux, L., & Swanson, D. A. (2021). A blocking approach to Metaphone-enhanced record linkage for public health data. International Journal of Environmental Research and Public Health, 18(12), 6338. https://doi.org/10.3390/ijerph18126338

Li, J., Zhu, Y., & Wang, H. (2021, September). Explainable decision tree for record linkage with feature importance analysis. In Proceedings of the 2021 International Conference on Big Data (pp. 123-132).

Li, J., Zhu, Y., & Wang, H. (2023). Flexible threshold setting for decision tree-based record linkage. *Knowledge and Information Systems* (In Press). [DOI: to be added when available]

Li, J., Zhu, Y., & Wang, H. (2022, April). Improving decision tree performance for record linkage using active learning and cost-sensitive learning. In Proceedings of the 2022 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1702-1711). https://doi.org/10.1145/3479029.3479103

Li, J., Zhu, Y., & Wang, H. (2022, April). Enhancing decision tree performance for record linkage with active learning. In Proceedings of the 2022 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1702-1711). https://doi.org/10.1145/3479029.3479103

Rokach, L., Maimon, O. (2008). Data mining with decision trees: Theory and applications (2nd ed.). World Scientific Publishing Co. Pte Ltd.

Smith, J., Johnson, R., & Thompson, M. (2020). A comparative analysis of record linkage techniques for entity resolution. Journal of Data Science, 18(3), 369-392. https://doi.org/10.6339/JDS.2020.18.3.369

Su, S., Xiao, Y., & Wang, H. (2021). Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier. Diagnostics, 11(2), 222. https://doi.org/10.3390/diagnostics11020222

Wang, J., Pei, J., & Zhang, Y. (2022). Link quality assessment for decision tree-based record linkage. Knowledge and Information Systems, 64(3), 1138-1158. https://doi.org/10.1016/j.ksem.2021.108222

Zhang, W., Fan, X., & Wu, X. (2021, November 18). Cost-sensitive decision tree learning for record linkage. arXiv. https://arxiv.org/abs/2111.09042

Zhang, W., Fan, X., & Wu, X. (2020). Record linkage with decision trees and blocking techniques for web data. In Proceedings of the 2020 International Conference on Big Data and Smart Applications (pp. 1-6). Association for Computing Machinery. https://doi.org/10.1145/3429244.3429252

Vadood Keramati

Vadood Keramati was born in Ardabil in July 1981. He received a bachelor's degree in Statistics from Tabriz University in 2003, a master's degree in Industrial Engineering from Islamic Azad University in 2015, and a PhD in Industrial Engineering from Payame Noor University in 2024. His research interests include information systems, data science, artificial intelligence, and machine learning.

Ramin Sadeghian

Ramin Sadeghian was born in Tehran on November 2, 1978. He received a bachelor's degree in Applied Mathematics from Amir Kabir University (Tehran Polytechnic) in 2001, a master's degree from Iran University of Science and Technology in 2003, and a PhD from Iran University of Science and Technology in 2007.

In 2008, he became an Assistant Professor at Bu Ali Sina University in Hamedan. Since 2013, he has been at Payame Noor University in Tehran, where he became an Associate Professor of Industrial Engineering in 2017 and a Professor in 2024. His research focuses on game theory, supply chain management, and statistical and mathematical modeling. He has authored or co-authored approximately 10 books and over 100 scientific articles in reputable journals and conferences. He has supervised and advised more than 70 master's and PhD students. Currently, he serves as the IT Manager of Payame Noor University.

Maryam Hamedi

born in 1982, holds a PhD from the University of Putra, Malaysia, as well as a BA and MSc from AmirKabir University of Technology. Since 2012, she has been an Assistant Professor in the Department of Industrial Engineering at Payam Noor University, Iran. Her research interests include optimization, quality management, data mining, and machine learning.

Ashkan Shabbak

born in 1979, holds a PhD in Applied Statistics from the University of Putra, Malaysia, and BA and MSc degrees from Isfahan University and Tehran Azad University, respectively. He has been an Assistant Professor in the Statistics Department at the Statistical Research and Training Center of Iran since 2014. His research interests include official statistics, quality management, and data governance.

Iranian Journal of Information Processing and Management