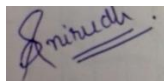


# **STUDY ON RECORD LINKAGE COMPARSION** **PATTERN**

STUDENT NAME: Anirudh shukla

ENROLMENT NUMBER: 00619011921

SIGNATURE:



EMAIL ID: anirudh.00619011921@ipu.ac.in

CONTACT NUMBER: 8826728647

GOOGLE DRIVE LINK:

Report:

[https://drive.google.com/file/d/1Ao3gbcwz8ziMga8jD5yJdZpxPxHD0sr4/view?usp=drive\\_link](https://drive.google.com/file/d/1Ao3gbcwz8ziMga8jD5yJdZpxPxHD0sr4/view?usp=drive_link)

Code:

<https://colab.research.google.com/drive/1c3SIUWWnwdY6eQdhC6l7uT24F7B9LHsV?usp=sharing>

Data Sets:

<https://archive.ics.uci.edu/dataset/210/record+linkage+comparison+patterns>

[https://drive.google.com/drive/folders/15XtB-qENV24skbBYLN03m60A2J-bwW2g?usp=drive\\_link](https://drive.google.com/drive/folders/15XtB-qENV24skbBYLN03m60A2J-bwW2g?usp=drive_link)

GOOGLE WEBSITE LINK:

<https://sites.google.com/view/record-linkage-comparison-patt/home>

YOUTUBE VIDEO LINK:

<https://youtu.be/C6UqP8xY330>

# **REPORT**

**TITLE:** Record Linkage Comparison Pattern

**ABSTRACT:**

Record linkage, also known as entity resolution or data deduplication, is a vital task in data integration and data cleansing processes. It involves identifying and linking records across different datasets that correspond to the same real-world entity. This abstract explores the concept of record linkage and its significance in various domains, such as healthcare, finance, and customer relationship management. It highlights the challenges associated with record linkage, including handling errors, dealing with large-scale datasets, and selecting appropriate comparison patterns. Effective comparison patterns play a crucial role in determining the accuracy and efficiency of record linkage algorithms. This abstract provides an overview of the importance of comparison patterns and their impact on record linkage outcomes.

**KEYWORDS:**

1. Record linkage
2. Data deduplication
3. Entity resolution
4. Comparison patterns
5. Data quality

**INTRODUCTION:**

Record linkage, also known as data deduplication or entity resolution, is the process of identifying and matching records that refer to the same entity across different data sources or databases. It plays a crucial role in various applications such as data integration, customer relationship management, and fraud detection. One of the key challenges in record linkage is designing effective comparison patterns that determine the similarity between pairs of records.

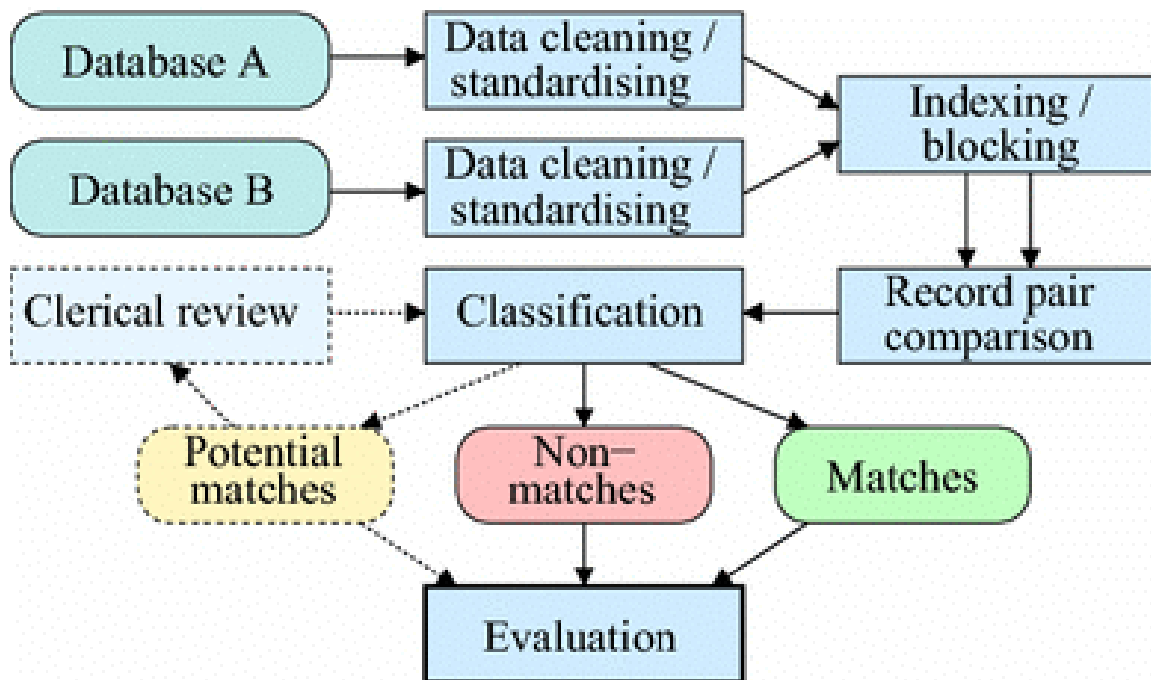
Comparison patterns define the criteria and algorithms for comparing different attributes or features of records to determine their similarity or dissimilarity. These patterns can include string similarity measures, such as edit distance or Jaccard similarity, as well as domain-specific rules or heuristics. The choice of comparison patterns greatly influences the accuracy and efficiency of the record linkage process.

Designing effective comparison patterns involves considering factors such as the quality and completeness of data, the characteristics of the attributes being compared, and the specific requirements of the application. It requires a balance between precision (correctly identifying matches) and recall (ensuring all matches are identified).

In conclusion, the design and selection of comparison patterns in record linkage are critical for achieving accurate and efficient entity resolution. By considering the characteristics of the

data and the specific application requirements, researchers and practitioners can develop effective patterns to improve data quality and enable reliable decision-making processes.

### **PROPOSED METHODOLOGY:**



### **Datasets:**

The records represent individual data including first and family name, sex, date of birth and postal code, which were collected through iterative insertions in several years. The comparison patterns in this data set are based on a sample of 100.000 records dating from 2005 to 2008. Data pairs were classified as "match" or "non-match" during an extensive manual review where several documentarists were involved. The resulting classification formed the basis for assessing the quality of the registry's own record linkage procedure.

The data set is split into 10 blocks of (approximately) equal size and ratio of matches to non-matches.

The separate file frequencies.csv contains for every predictive attribute the average number of values in the underlying records.

Number of Instances: 5.749.132

Number of Attributes: 12 (9 predictive attributes, 2 non-predictive, 1 goal field)

Attribute Information:

id\_1: Internal identifier of first record.

id\_2: Internal identifier of second record.

cmp\_fname\_c1: agreement of first name, first component

cmp\_fname\_c2: agreement of first name, second component

cmp\_lname\_c1: agreement of family name, first component

cmp\_lname\_c2: agreement of family name, second component

cmp\_sex: agreement sex  
cmp\_bd: agreement of date of birth, day component  
cmp\_bm: agreement of date of birth, month component  
cmp\_by: agreement of date of birth, year component  
cmp\_plz: agreement of postal code  
is\_match: matching status (TRUE for matches, FALSE for non-matches)

### **Pre-processing:**

#### **Missing Attribute Values:**

id_1	0
id_2	0
cmp_fname_c1	1007
cmp_fname_c2	5645434
cmp_lname_c1	0
cmp_lname_c2	5645434
cmp_sex	0
cmp_bd	795
cmp_bm	795
cmp_by	795
cmp_plz	12843
is_match	0

Pre-processing refers to the set of activities or techniques applied to raw data before it can be used for analysis, modelling, or any other data processing tasks. It involves transforming, cleaning, and organizing the data to improve its quality, consistency, and suitability for further analysis.

### **Indexing / Blocking:**

By using different blocking/indexing techniques one reduces the number of pairs to check. If two pairs of record attributes are completely dissimilar; those record comparisons are locked. The blocked pairs do not need to be evaluated and instead the upcoming similarity measures can be used only on the matching pairs.

### **Classification:**

Classification algorithms are used to categorize each pair of records as “matches” or “non-matches”. Our research mostly focuses on this step in the record linkage process. Non matches now all consist of blocked records and additional non-matches are found by using different categories of decision models. Some of the Classification Techniques are:

- 1) DecisionTreeClassifier: from sklearn.tree import DecisionTreeClassifier
- 2) RandomForestClassifier: from sklearn.ensemble import RandomForestClassifier
- 3) SVC (Support Vector Machine): from sklearn.svm import SVC
- 4) LogisticRegression: from sklearn.linear\_model import LogisticRegression
- 5) GaussianNB: from sklearn.naive\_bayes import GaussianNB

### **Evaluation:**

Evaluation is done by calculating the scores such as:

- 1) Accuracy
- 2) Precision
- 3) Recall
- 4) F1-Score

### **RESULT:**

#### **DecisionTreeClassifier:**

Classifier: <class 'sklearn.tree.\_classes.DecisionTreeClassifier'>

Accuracy: 1.0000

Precision: 1.0000

Recall: 1.0000

F1-Score: 1.0000

#### **GaussianNB:**

Classifier: <class 'sklearn.naive\_bayes.GaussianNB'>

Accuracy: 0.9998

Precision: 0.9998

Recall: 0.9998

F1-Score: 0.9998

#### **RandomForestClassifier:**

Classifier: <class 'sklearn.ensemble.\_forest.RandomForestClassifier'>

Accuracy: 1.0000

Precision: 1.0000

Recall: 1.0000

F1-Score: 1.0000

#### **LogisticRegression:**

Classifier: <class 'sklearn.linear\_model.\_logistic.LogisticRegression'>

Accuracy: 1.0000

Precision: 1.0000

Recall: 1.0000

F1-Score: 1.0000

### **SVC (Support Vector Machine):**

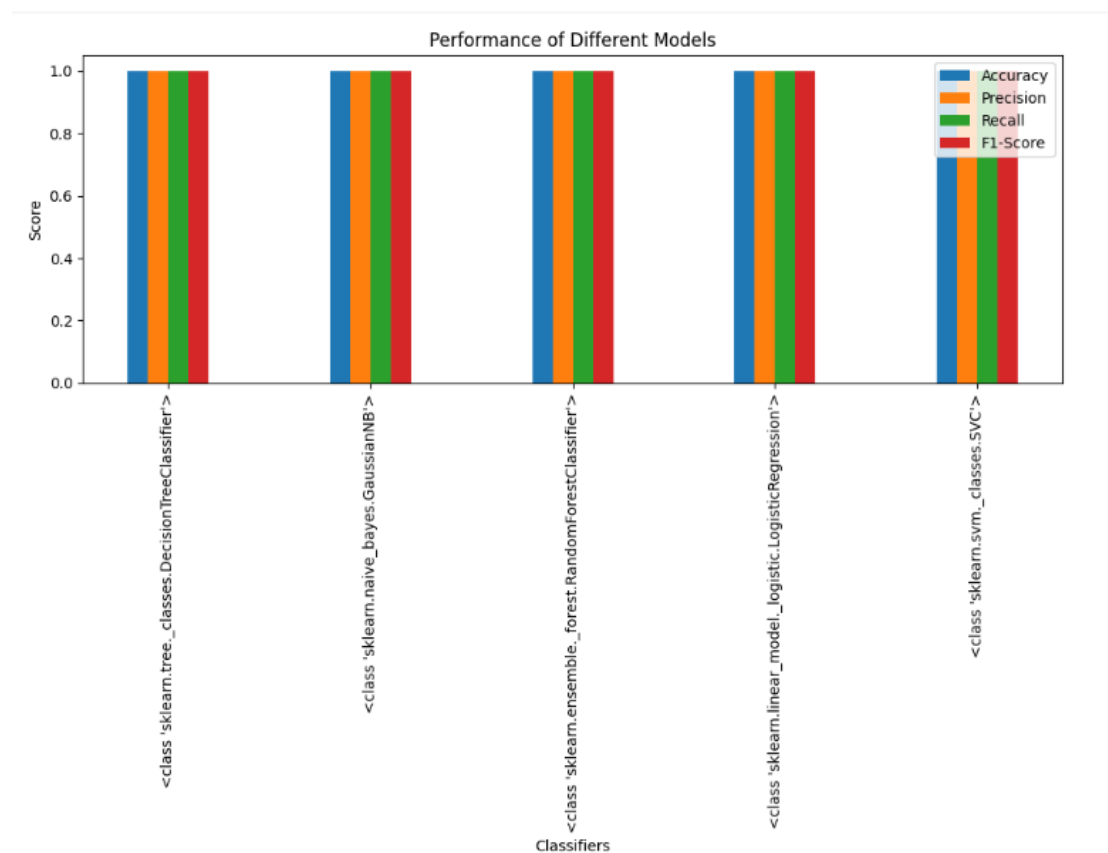
Classifier: <class 'sklearn.svm.\_classes.SVC'>

Accuracy: 1.0000

Precision: 1.0000

Recall: 1.0000

F1-Score: 1.0000



### **CONCLUSION & FUTURE WORK:**

The Conclusion of the study is that for classification of 20.931 matches & 5728201 non-matches. Application of different classifiers resulted to produce scores such as Accuracy, Precision, Recall and F1-Score. And, also the comparison of scores of different classifiers by plotting graph.

Future Work in this field are as follows:

- 1) Advanced similarity measures.
- 2) Handling high-dimensional data.
- 3) Incorporating domain knowledge.

- 4) Privacy-preserving record linkage.
- 5) Improved Scalability and efficiency.

## **REFERENCES:**

- 1) Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc* 1969, **64**: 1183– 1210.
- 2) Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. *Science* 1959, 130: 954– 959.
- 3) Randall SM, Ferrante AM, Boyd JH, Semmens JB. The effect of data cleaning on record linkage quality. *BMC Med Inform Decis Mak* 2013, 13: 64.  
<http://www.biomedicalcentral.com/1472-6947/13/64>. (Accessed June 4, 2014).
- 4) Herzog TN, Scheuren F, Winkler WE. *Data Quality and Record Linkage Techniques*. New York: Springer; 2007.
- 5) Cohen WW, Sarawagi S. Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods. In: *Proceedings of the ACM Knowledge Discovery and Data Mining Conference*, 2004, 89–98.
- 6) Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In: *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1990, 354–359.
- 7) Cooper WS, Maron ME. Foundations of probabilistic and utility-theoretic indexing. *J Assoc Comp Mach* 1978, 25: 67– 80.
- 8) Deming WE, Gleser GJ. On the problem of matching lists by samples. *J Am Stat Assoc* 1959, 54: 403– 415.
- 9) Michelson M, Knoblock CA. Learning blocking schemes for record linkage. In: *Proceedings of AAAI*, 2006.
- 10) Kenig B, Gal A. Efficient entity resolution with MFI blocks. In: *Proceedings of VLDB*, Lyon, France, August 2009.
- 11) Titterton DM, Smith AFM, Makov UE. *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley & Sons; 1988.
- 12) Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press; 1975.