Journal of Scientometric Research, 2025; 14(1):300-308.
https://www.jscires.org

Research Article

Predicting Collaborations among Research Scientists: A Datathon Experience

Maria Del Pilar Angeles[1,*,#], Helena Gomez-Adorno[1], Sinuhe David Hernandez-Guevara[2], Victor Manuel Corza-Vargas[2,#]

[1]Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de Mexico (UNAM), MEXICO.
[2]Divulgación Académica, Centro de Estudios en Computación Avanzada, Universidad Nacional Autónoma de Mexico (UNAM), MEXICO.
[#]These authors contributed equally to this work.

ABSTRACT
This paper presents the challenge of predicting collaborations between research scientists as a datathon experience. The focus of the challenge task is determining whether or not the author of a research paper is keen to collaborate with another author in the future. The main aims of the datathon challenge are: (i) to show the feasibility of automatically identifying potential collaboration in a research network as a link prediction task; (ii) to propose a methodology for the environment configuration that covers data collection, selection and preparation stages required for link prediction in a massive event. (iii) to join the efforts of students from different fields of study in solving the task from a multi-disciplinary perspective. For this purpose, we created a corpus with DBLP data covering publications from 1990 to 2004. The created dataset has been made available for further research. Altogether, the datathon attracted 78 registered students, yielding 13 submissions of teams composed of 6 students each. In this paper, we compare their approaches and analyze their performance.

Keywords: Link prediction problem, Topological features, Authorship predictions, Imbalanced classification model, Datathon.

Correspondence:
Maria Del Pilar Angeles
Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de Mexico (UNAM), MEXICO.
Email: pilarang@unam.mx
ORCID: 0000-0002-1754-2514

Received: 25-07-2024;
Revised: 08-09-2024;
Accepted: 20-11-2024.

INTRODUCTION

A datathon is an event where groups of multidisciplinary students and early career researchers can work together on a new interdisciplinary project for a concentrated period, usually over the course of around three days. The aim is to bring together students with complementary skills and knowledge; they then work together to create an initial plan to solve a specific link prediction problem. This allows the participants to establish new and completely independent collaborations with minimal intrusion into their normal duties and, in the process, create a basis for the development of interdisciplinary data management projects. The number of datathons has increased dramatically since its initiation. It is used by industries, scientists and others to solve a wide variety of data management problems and to develop new strategies within a short period of time.[1,2]

Datathons have emerged recently as a new way of developing solutions in any application area, promoting rapid learning, innovation and collaboration. The organization of a datathon is an excellent opportunity for a data science challenge. So, the Institute for Research in Applied Mathematics and Systems (IIMAS) of the National Autonomous University of Mexico (UNAM) organized a Datathon in January 2020 that lasted three days. The first day was dedicated to delivering three workshops: Introduction to Network Analysis: Statistical Measurements and Network Connectivity; Influence and Centrality in Networks and Prediction of Links. During the second day, the challenge statement was explained, the teams were formed and they were given access to the programming environment and the data. On the third day, the teams focused on the link prediction modeling. After the evaluation, the University awards the best predictive models. Such an event involved students with no experience in data science and experts in the area as mentors, which resulted in an enriching and formative experience for all the participants.

The challenge presented in the IIMAS-UNAM datathon 2020 consists of predicting co-authoring collaborative social networks in the DBLP database and applying graph analysis, statistics and machine learning techniques. The presented paper is organized as follows: The Datathon Challenge section describes the main issues concerning creating a predictive model on the collaboration of university scientists. The next two sections detail the steps the

DOI: 10.5530/jscires.20251453

Datathon organizers took in preparing the data and selecting the metrics to evaluate the teams. Subsequently, the outcomes obtained from the generated models are analyzed, the findings are concluded and future work is identified.

The Datathon Challenge: Predicting Collaborations among Scientists

The National Autonomous University of Mexico (UNAM) has 24 scientific research institutes, 7 research centers, 12 humanities institutes, 25 schools, 5 foreign campuses within the country and 14 offices abroad that carry out teaching projects, research and technical development related to computer science.

UNAM is a distinguished university worldwide for its teaching, scientific and technological contributions. Furthermore, these contributions have enormous potential for growth, specialization and application if collaboration is strengthened across all entities and university sites.[3]

Several researchers are working on similar or complementing research projects within the university. Due to the lack of contact or collaboration among such colleagues, many academics might spend more time obtaining significant results.

The challenge consists of analyzing scientific papers to predict future collaborations between academics that produce more and better scientific results. Regarding the data science challenge, the approach focuses on identifying, comparing and improving algorithms for link prediction in a social-scientific network. Previous work has already been done in.[4]

To predict new collaborations between computer science paper authors, co-authorships are represented as edges in a graph (or social network). A graph consists of a set of nodes (in this case, author) that relate to each other to coauthor a scientific article.

The problem of predicting links in a graph corresponds to finding the probability of a future association (co-authorship) between two nodes, knowing that there is no association between them in the current state of the network.

Considering an undirected graph $G=V, E$ where each edge $e=(u, v)$ E represents an interaction between nodes u and v at a particular time t. Such interaction, in the domain of our problem, is defined as the co-authorship of a research article.

There is link prediction algorithms designed to estimate different influence rates within the links in a graph. They assume that a node with multiple connections may be more likely to receive additional link influences.[5]

For example, predicting co-authorship among scientists involves understanding how the authors relate to each other and, for instance, measuring the tendency of scientists who share connections in a research group to connect with each other to achieve their goals and publish new articles.[6] The sample dataset utilized in the IIMAS-UNAM Datathon 2020 corresponds to the Digital Bibliography and Library Project (DBLP) database,[7] specifically the DBLP monthly release from January 2019.[8] We decided to use this dataset since it is a free resource that provides open bibliographic information on major journals and reports of computer science conferences.

The following subsection analyzes the challenge to be presented to the teams and what considerations should be taken into account during the environment's configuration to provide the competitors with sufficiently prepared data so that they only focus on modeling. As well as identify which metrics would be in accordance with the data and the model.

Analysis of the Challenge

Given the short duration of the datathon, there were many issues that we tackled before it began:

The information stored in the DBLP database was not in optimal size, content and format conditions to be analyzed directly. Typically, the most time-consuming stage in a data science project is pre-processing and transformation. Consequently, an initial pre-processing was required to let the participant teams concentrate on the analysis, generation and comparison of models, to achieve better performance in the prediction.

The link prediction problem can be approached by binary classification. However, in this case, the imbalanced nature of the classification model[9] arises because the sample datasets are not balanced; that is, there is a majority class represented by the number of negative outcomes covering 75% of possible cases and a minority class, represented by the number of positive outcomes the remaining 25% of possible cases.

Authors with less than three publications are irrelevant to productive research groups, so they were not considered for prediction.

The nature of the link prediction problem in a social network requires supervised learning. So, to evaluate the performance of the prediction of the participant's models, data from past and present collaborations must be gathered to compose training and validation sets. Using data from future collaborations to test the participant's models is also necessary. It is, therefore, necessary to divide the DBLP database by periods of publication years.

There are various indicators or metrics for evaluating the performance of the models.[10] They depend on the type of learning to be carried out, the algorithms used, etc. Thus, it was necessary to identify the more suitable performance metric that would be applied during the model assessment.

The time required for the correct and fair evaluation of the proposals would be directly proportional to the number of teams formed. Therefore, a mechanism must be established to automate the model evaluation.

The following section presents the main steps carried out for setting the link prediction environment to guarantee a successful event in terms of time and achievements. Setting the environment: the current state graph, the training and test Datasets

Three data sets were designed containing co-authorships carried out during different periods to address the link prediction problem through an imbalanced binary classification model.[9,11] The first data set included authorships that occurred from 1990 to 2000 and corresponded to the current authorship graph. The second data set is focused on co-authorships carried out during 2001 and 2002 and it will be required to train the model. The third data set contains those co-authorships established from 2003 to 2004 and it will be used to evaluate the link prediction models. As the goal of the models is to predict links in a data set by successfully distinguishing positive classes, this problem was considered a binary classification problem that can be solved using effective features in a supervised learning framework.

Figure 1 shows the temporality and the overall flow of the process of extracting co-authorship sets from the DBLP database. These sets correspond to several-year intervals that do not overlap. The three data sets serve as a framework for addressing the link prediction problem through a binary classification model.

The prediction consists of finding the set of links formed at time $t+\Delta$, called $E(t+\Delta)$, given the current state of a network (co-authorship graph) at time t, $G(t)=\{V(t), E(t)\}$.

Generating the current state of the graph (from 1990 to 2000)

The Current state of the Graph is a co-authorship network that can be used as the reference point to extract the features that allow the classification model to identify the positive classes in the classification data sets. The positive classes in this problem are the authors who will collaborate in the future. This section describes the preprocessing performed on the original DBLP dataset to create the current state of the graph with publications from the years 1990 to 2000.

The DBLP database is a free Extensible Markup Language (XML) structure resource that provides open bibliographic information on major journals and reports of computer science conferences worldwide. It was originally created at Trier University in 1993. Nowadays, the DBLP database is operated and developed by Schloss Dagstuhl. Further information can be found in.[8]

The XML dataset provided by DBLP consists of a series of structured Tags. The root element is the Tag <dblp> which contains a sequence of bibliographic records. The DBLP dataset contains publications of articles in a newspaper, magazine, or conference, proceedings, books, in-collection (a part or chapter in a monograph) and master and doctorate theses. The DTD document of the DBLP dataset is shown in Figure 2.

The DBLP elements used to build the current state of the graph $G(t)$ are:

- **Authors:** Represent the nodes of the network and are denoted by $V(t)$.
- **Co-authoring:** Represent the edges of the network and are denoted by $E(t)$.
- **Absence of co-authoring:** Datapoint, a pair of nodes that are not related in the current state of the network.

One feature that significantly impacts link prediction is the sum of articles that the pair of authors has published. The importance of this feature arises from the fact that authors with higher article counts are more prolific. If one or both authors are prolific, the probability that this pair will collaborate in a co-authorship is greater than the case of unprolific authors.[4] Considering the above, to rule out unprolific authors, the current state of the graph should include only authors with at least three publications. This minimum number of publications was considered empirically.

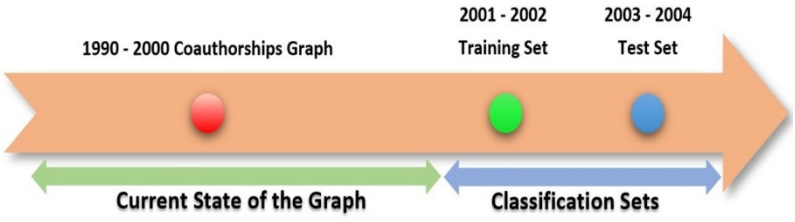

Figure 1: Temporality and general flow of the problem.

```
<!ELEMENT dblp (article|inproceedings|
  proceedings|book|incollection|
  phdthesis|mastersthesis|www)*>
```

Figure 2: DBLP DTD Structure.

The generation of the current state of the graph was structured as a pipeline composed of 4 steps. Figure 3 describes the flow of the pipeline:

Step 1-Extracting authorships from DBLP

In this step, the authors' data were extracted from the DBLP data set. The original first and last names were restricted to ISO-8859-1 characters, so they were converted to UTF-8. To optimize time processing during the programming and test phases of the datathon, the input data set was reduced to articles and conferences and included only information from 1990 to 2000. The data cleansing and pre-processing is summarized as follows:

- Translation of the Latin-1 XML document ISO-8859-1 to UTF-8.
- Selection of the specific period of time (1990-2000).
- Obtaining articles and conference records.
- The output of this step is the *authorships.csv* file with the fields' id article, author and year.

Step 2-Generation of Nodes Catalog

In this step, a node catalog was generated with the following considerations:

- Each author is a node.
- The ID of each author is their full name.
- Duplicate records were deleted.
- The authors with 2 publications or less were discarded.
- Generation of authors catalog with fields: author and id article.
- The *authorships.csv* file with 482998 authors yielded a catalog with 43912 nodes named
- *nodes.csv*.

Step 3-Authorships filtering

In this step, publications were filtered to preserve only those whose authors appear in the node catalog, according to the following:

- Read nodes catalog (*nodes.csv*).
- Read authorships file (*authorships.csv*).
- Output authorships that have authors with more than two articles to the filtered authorships file (*filteredAutorships.csv*) with fields: id article, author and year.
- Once the publications were filtered to preserve only those whose authors appear in the node catalog, the number of coauthorships decreased to 306197, corresponding to 43912 authors.

Step 4-Generation of Edges Set (Current state of the graph)

This step generates a list of edges that represent the co-authorship network in its current state. Edge weights are not taken into account, so duplicate edges were removed. The per-mutations of the edges were also discarded, for example, edge (author1, author2), edgePermutation (author2, author1). The following process is performed to generate the co-authorship network:

- Read filtered authorships (*filteredAutorships.csv*).
- Create a pair of the authors that appear in the same article.
- Generate the corresponding edges with the authors' pairs.
- Delete duplicate authors' pairs.
- Finally, the current state of the graph contains 43912 nodes and 95703 edges, which is undirected and unweighted.

Generating the Training and Testing Datasets

The training set contains data from 2001 to 2002 and the testing set contains data from 2003 to 2004. These files were generated following a pipeline of seven steps, where the first four steps were

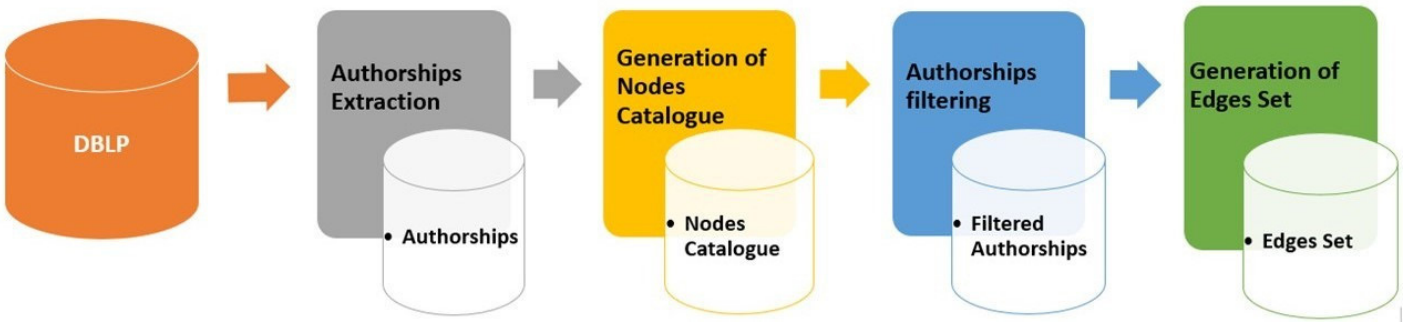

Figure 3: Pipeline to generate the current state of the graph.

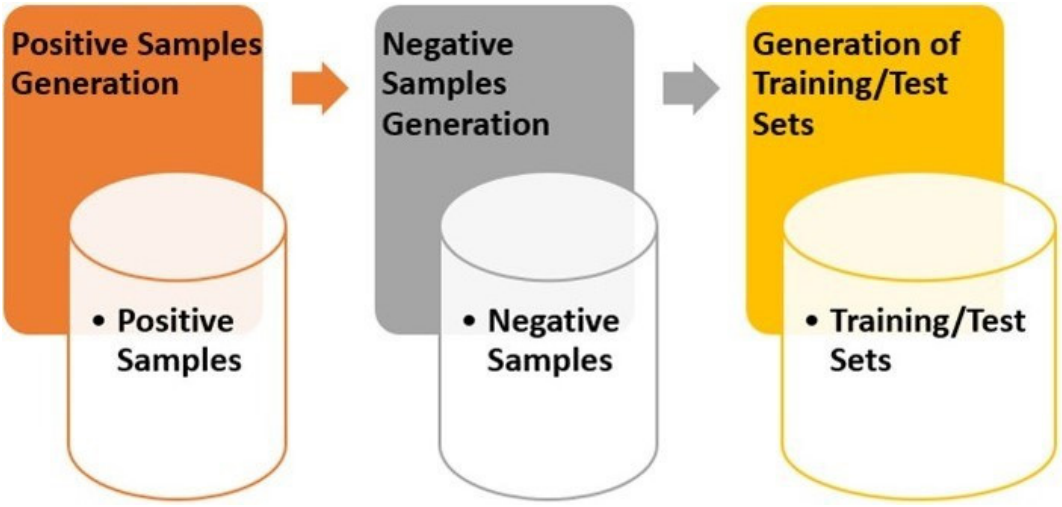

Figure 4: Pipeline of the additional steps to generate the training and test datasets.

already explained in Figure 3 and each set of author pairs must meet the following conditions:

- Both authors must appear in the current state of the graph (1990-2000).
- The authors did not publish any articles together in the current state of the graph
- (1990-2000).
- If both conditions are met and the pair of authors published in one of these datasets, there is a positive sample. Otherwise, there is a negative sample.
- Figure 4 describes the pipeline of the additional three steps required to generate the training and testing datasets.

The entire process for generating the training and testing sets is described through seven steps:

Step 1-Authorship Extraction

This step follows the same process as in the co-authorship network generation. The only difference is the considered periods, 2001-2002 for the training set and 2003-2004 for the testing set. After this process, the total authorship extracted for the training set is 167611 and for the testing set are 222186.

Step 2-Generation of Nodes Catalog

As a result of selecting authors from the current state catalog node and authorships, the training data set comprised 66936 nodes. The testing data set corresponding to the years 2003-2004 contained 22663 nodes.

Step 3-Authorship Filtering

As a result of filtering authors from the training set node catalog and the authorships during the years 2001-2002, there are 66936 authorships. In the case of testing data set (2003-2004) there are 69994 authorships.

Step 4-Generation of the Edges Set

A full set of author pairs (edges) was generated by filtering the authors within a specific period of time, 2001-2002 for training with 23246 edges and 2003-2004 for testing with 20678.

Step 5-Positives sample generation

In this step, the positive samples of the training and testing datasets were obtained. The minority class of the training set represents 25% of the possible cases. To do this, the edges in the network's current state were removed from the set generated in step 4. The resulting edges were labeled with the letter *P* to identify them as positive samples. The new 11753 edges become the positive samples for the training set. In the case of the testing dataset (2003-2004), 12764 edges were obtained as positive samples.

Step 6-Negative sample generation

The set of negative samples was randomly obtained. The number of negative samples maintained the proportion established for these sets. The resulting edges were labeled with the letter *N*. The training data results in 35259 negative samples. The testing data set contained 38292 negative samples.

Step 7-Yielding the final training and test datasets

The last step joined the positive and negative samples into a single dataset to generate the final training and testing sets separately. Both sets followed the proportion of 25% for positive samples and 75% for negative samples. The final training dataset consists of 23606 nodes, 23246 edges, 11753 positive samples and 35259 negative samples. The testing dataset contains 22663 nodes, 20678 edges, 12764 positive samples and 38292 negative samples.

Table 1: Generation process of current state, training and test datasets.

Dataset	Authorships	Nodes catalog	Authorships filtering	Edges set	Positive samples	Negative samples
Current state	482998	43912	306107	95703	–	–
Training	167611	23606	66936	23246	11753	35259
Testing	222186	22663	69994	20678	12764	38292

A sample of the data sources generated during the present work is available in the repository of the event.[1]

Table 1 shows the number of authors and authorships obtained at each step of the process to generate the current state, training and testing datasets.

Evaluation Framework

After the training data was released to the participants, the time period to submit the results was 16 hr. Each team received an identification number to evaluate the methods developed by the participating teams during the datathon. All teams submitted their link prediction files to a server. A sample of the results is also available in the repository of the event. Each file was evaluated using the metrics described in subsection.

Evaluation Metrics

As was mentioned in section on the datathon challenge description and it can be observed in Table 1, both training and testing sets have a majority class represented by the number of negative links instances, which covers 75% of possible classes. Thus, we must be careful in how to evaluate the models presented during the competition.

The statistical performance measures of a binary classification model are called rates: True Positives (TP) correspond to the number of instances the classifier predicted correctly in the positive class; False Negatives (FN) correspond to the number of instances incorrectly classified in the negative class, also known type II error; False Positives (FP) correspond to the number of instances incorrectly classified in the positive class, also known type I error; True Negatives (TN) correspond to the number of instances the classifier predicted correctly in the negative class. This section will present the most popular metrics for evaluating link prediction methods.

Precision

The precision metric[10] calculates the precision of the minority class. It is the ratio of the number of correctly predicted positive samples (TP) to the total number of samples predicted in the positive class (which corresponds to the sum of True Positives (TP) and False Positives (FP)).

$$\text{Precision}=\text{True Positives}/(\text{True Positives} + \text{False Positives})$$

Precision is a good measure to consider when the cost of false positives is high. For example, in spam detection, a false positive means that an email that is not spam (actual negative) has been identified as spam (positive). The mail user may lose important emails if the accuracy of the spam prediction model is not high.

Recall

The recall metric[10] is the ratio of the number of samples correctly predicted in the positive class to the number of all positive samples. Unlike precision, recall indicates the positive samples that the model failed to detect. In other words, recall provides a notion of positive class coverage.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Recall should be used when a high cost is associated with false negatives. For example, in fraud detection, if a fraudulent transaction (actual positive) is predicted as non-fraudulent (predicted negative), then the risk of losing large amounts of money in a financial institution would be very high. Precision and recall cannot fully describe a model's predictability. A model can have very high precision and very low recall, or vice versa.

F_1-score

Provides a combined measure between precision and recall. It corresponds to the harmonic mean between precision and recall. Thus, it is calculated through the following formula:

$$F_1\text{score}=2\frac{(\text{Recall x Precision})}{\text{Recall+Precision}}$$

Example: Consider a dataset containing 12975 positive samples and 38925 negative samples and the results of a prediction model are presented by a confusion matrix in Table 2, there is no value of true negative because it is not used for the metrics in question:

Precision= 12327/ (12327 + 7136) = - 0.63

Recall = 12327 / (12327 + 648) = - 0.95

F1-Score = (2* 0.63 * 0.95) / (0.63 + 0.95) = 0.757

From the previous measures, we could say that the model has low precision but excellent recall. Furthermore, the F-score balances precision and recall, providing sufficient information to describe the model's predictability when trained in an imbalanced dataset.

F_1-Score was chosen to assess the methods for the co-authorship prediction problem addressed in the UNAM datathon, because

Table 2: Confusion matrix for a dataset with 12975 positive and 38925 negative samples.

		Predicted			
		Negative		Positive	
Actual	Negative	TN	-	FP	7136
	Positive	FN	648	TP	12327

Table 3: Confusion matrix generated.

Outcome	Meaning
TP	Co-authors labeled as positive that do exist in the test set (2003-2004)
FP	Co-authors labeled as positive that do not exist in the test set (2003-2004)
TN	Co-authors labeled as negative that do not exist in the test set (2003-2004)
FN	Co-authors labeled as negative that do exist in the test set (2003-2004)

it was the most conveniently used metric in imbalanced classification issues. The steps carried out during the evaluation process are explained in the following section.

Evaluation process

The evaluation process goes through three stages:

Delivering results

The participant teams received an unlabeled version of the test set to evaluate their co-authorship prediction model. The predictions obtained by their models are sent to the competition judges.

Generating the Confusion Matrix

For each prediction file submitted by the participants, the judges generate a confusion matrix; the judges used the labeled version of the test set to count the number of positive and negative samples correctly classified taking into account the equivalences listed in Table 3.

Calculation of the evaluation metric

The F_1-Score is calculated for each prediction file of the participating teams, with the values of the confusion matrix.

Overview of the Challenge Results

Each team chose its data science strategy and working plan and competed through a long, cheerful programming night and day against their opponents. Sixteen teams made it to the end of the competition and uploaded the results to the file system. Nine teams predicted the link with a reasonable F_1-Score. Most teams approached the link prediction through hand-crafted rule-based methods and others programmed machine-learning-based methods. The rule-based methods focused on computing the similarities of disconnected pairs of nodes by analyzing the proximity of nodes, where every potential node pair would be assigned a score. A higher score means a higher probability of establishing a link in the future. The machine-learning-based methods focused on a binary classification task[4] using as feature set the same similarity metrics as the rule-based systems. If there is a potential link connecting a pair of nodes, this pair is labeled as positive, otherwise it is negative.

Table 4 shows for each participating team, the team number, the language in which they developed the solution, the methodology followed (rules or machine learning algorithm), the features computed (link prediction metrics) and the F_1-score value achieved in descending order, it can be observed the highest F_1-score value was 0.62 by Team #16. The submissions of the rest of the teams failed to be evaluated, for different reasons. Table 5 describes the reasons why these submissions failed. Team #3 provided numerical indexes instead of author names in their prediction files. Team #8 and Team #20 did not provide a prediction field. Team #9 did not provide any file for evaluation. Team #29 provided a file with a different character set. Team #15 provided a file with many predictions different from those expected.

The best-participating team (team #16) developed a rule-based system using three similarity methods and combined them to predict the co-authorship of two authors. The first method is to compute the common neighbors[12] between the pair of nodes. Nodes with more neighbors in common are more likely to form an edge in the future.

The second method calculates the preferential attachment,[12] which gives the probability of co-authorship of x and y by computing the product of the number of collaborators of x and y. The third method is to find the length of the shortest path between the pair of nodes; if there is no path between the nodes a large number was assigned. Once the three values were computed for each pair of nodes in the training and testing set, they performed a grid search on the training set and obtained the threshold for each metric. With the obtained threshold, a binary value was assigned to each metric as follows: a) the number of common neighbours should be greater or equal to 2, b) the preferential attachment probability should be greater or equal to 200 and c) the shortest path should be less than 5. Finally, if any of the pairs of nodes meet any of the three conditions, they are classified as possible future co-authors.

The second-best approach (team #14) also used a similarity-based method with a rules hierarchy for predicting the nodes (authors) that will form a connection (co-authorship). The first rule is that there must be a path between a pair of nodes to form a relation in the future. If there is no path, the pair of nodes is automatically rejected to form a relationship. The second rule is that the

Table 4: Ranking of F_1-score values along with the details of submitted methods.

Team	Language	Method	Computed Metrics	F_1-score
16	Python	Rules	Common neighbours, preferential attachment and shortest path.	0.62
14	Python	Rules	Shortest path, secondary neighbours and Jaccard coefficient.	0.59
30	R	Log.Regr.	Jaccard coefficient and preferential attachment.	0.56
4	Python	Naive Bayes	Jaccard coefficient, resource allocation, Adamic adar index Soundarajan Hopcroft index and within inter cluster.	0.47
10	Python	XGboost	Common neighbours, Jaccard coefficient, resource allocation Index, Adamic adar index, preferential attachment, triadic closure left, triadic closure right, node centrality left, node centriality right and vector similarity.	0.35
5	Python	XGBoost	Common neighbours, jaccard coefficient, adamic adar index preferential attachment, eccentricity and shortest path length.	0.28
6	Python	Rules	Shortest path, node connectivity, minimum node cut, edge connectivity and minimum edge cut. Distance measures based on eccentricity: diameter, radius, periphery and center.	0.25

Table 5: Analysis of outcomes per team.

Team	F_1-score	Observations
3	0	The output format was incorrect (numerical indexes).
8	0	The output format was incorrect (no prediction field existed).
19	0	Not finished.
20	0	The output format was incorrect (no prediction field existed).
29	0	The character set must be UTF8.
15	0	Huge dataset with 43000000 million.

common neighbours between the nodes must be less than 4. If such a condition is fulfilled, the pair of nodes is automatically classified as a possible relationship. The third rule computes the Jaccard coefficient, which is added to the 1/400 of the common neighbor's measure. If this value exceeds a threshold, the pair of nodes is classified as a possible relationship.[13]

The third-best approach (team #30) used a machine-learning approach. They also computed similarity measures between nodes, but instead of programming fixed rules, they used these values as features to train a logistic regression algorithm. The computed similarity metrics were the Jaccard coefficient, preferential attachment and resource allocation index.[14] Regarding the programming languages used by the participants in the datathon, most of the teams chose Python to develop their solutions and one team developed its solution with R.

It is important to remember that the results obtained were the work of approximately 18 hr by students who had no experience in data science projects and who took three workshops in one day to identify and apply appropriate actions required to develop and implement a solution, allowing learners to try out new methods.

CONCLUSION

The IIMAS-UNAM data science datathon 2020 was organized by data scientists and researchers from this institution. The event aimed at strengthening and disseminating data science among the university community and was, therefore, raised as a training and development activity for future data scientists. The datathon brought together people working in the industry as data scientists or programmers and actuarial science, physics, mathematics and computing students. The planning and implementation of the datathon achieved all its objectives by integrating enthusiastic youth groups with a single academic purpose: the development of future data scientists.

Datathons have become a new way of creating predictive models, used by industries, scientists and now by universities to solve a wide variety of problems and to develop new strategies within a short period of time. The datathon format undertaken here, which focuses on analyzing a specific data set, is a good strategy for obtaining a deep data analysis and promotes learning based on problem-solving. The participants have the opportunity to put their knowledge into practice by working in a collaborative environment within a multidisciplinary team, allowing the combination of skills from young researchers, industry programmers and data scientists, who might not have had the chance to work closely together to build solutions through integrating small contributions from participants with different backgrounds.

As the original dataset contained many years and was in XML format, the organizers prepared the training and testing datasets for the teams to avoid wasting time preparing data, so the teams were focused on analysis. In this paper, we describe the data preparation performed for the co-authorship prediction problem, which is freely available in the repository of the event[1] in.[15] We

[1]https://github.com/pilarang/Dathaton.

explain the evaluation process and the metrics used to evaluate the models. The results of the developed models show that they can build a reasonable solution in a really short time period. Therefore, the proposed methodology for the environment configuration that covers the data collection, selection and preparation stages required for link prediction in a massive event was successful. The results allowed us to test some models and show the feasibility of automatically identifying potential collaboration in a research network as a link prediction task.

From an educational point of view, we can conclude that the promotion of data science among students and young researchers and the knowledge transfer from experienced data scientists were important parts of the event, which was enabled by working with peers across domains.

Two main alternatives have been considered to lay the groundwork for future work: the first corresponds to the reutilization of the data source and the preparation process of the training and testing samples explained throughout this paper, but widening the current span to cover a broader year range, adding other bibliographical sources such as IEEE, Web of Science, SCOPUS, etc. The second alternative is to increase the complexity of the datathon challenge by asking the competitors to design a prediction model to automatically learn the network's topological characteristics using various machine learning and deep learning models. Furthermore, a wide range of topological features can provide information regarding emergent properties of a social network through their predictive importance.[4] Regarding the data source to be utilized, some keywords mentioned in the academic articles can be incorporated to predict collaborations based on the topics covered in the articles.

ACKNOWLEDGEMENT

This research was partially funded by DGAPA-UNAM through PAPIIT projects TA101722, IN104424 and IN100719.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ABBREVIATIONS

DGAPA: Dirección General de Asuntos del Personal Académico; **UNAM:** Universidad Nacional Autónoma de México; **PAPIIT:** Programa de Apoyos de Investigación e Innovación Tecnológica.

REFERENCES

1. Flus M, Hurst A. Design at hackathons: new opportunities for design research. Des Sci. 2021;7. doi: 10.1017/dsj.2021.1.
2. Piza FM, Celi LA, Deliberato RO, Bulgarelli L, de Carvalho FR, Filho RR, *et al.* Assessing team effectiveness and affective learning in a datathon. Int J Med Inform. 2018;112:40-4. doi: 10.1016/j.ijmedinf.2018.01.005, PMID 29500020.
3. Lane C. Top universities. Retrieved; 2001. https://www. Available from: http://topuniversities.com/university-rankings-articles/world-university-rankings/top-universities-world-2021.
4. Hasan. M.A., Zaki, M.J.: A survey of link prediction in social networks. Soc Netw Data Anal. 2011;9:243-75. doi: 10.1007/978-1-4419-8462-3.
5. Namata, Getoor G, L.: Link prediction. Encyclopedia of machine learning and data mining. Springer, ??? (2017). p. 753-8. doi: 10.1007/978-1-4899-7687-19486. https://doi.org/10.1007/978-1-4899-7687-1 486.
6. Huang H, Tang J, Liu L, Luo J, Fu X. Triadic closure pattern analysis and prediction in social networks. IEEE Trans Knowl Data Eng. 2015;27(12):3374-89. doi: 10.1109/TKDE.2015.2453956.
7. Ley M. DBLP: some lessons learned. Proc VLDB Endow. 2009;2(2):1493-500. doi: 10.14778/1687553.1687577.
8. The Dblp Team: DBLP Computer Science Bibliography. Available from: https://dblp.org/.
9. Brownlee J. Imbalanced classification; 2020. Making developers awesome at machine learning. [retrieved Mar 19, 2021 from]. Available from: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/Computer.
10. Koo PS. Accuracy, precision, recall or F1? Towards data science Canada; 2018. accuracy-precision-recall-or-f1-331fb37c5cb9. Available from: https://towardsdatascience.com/.
11. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5(4):221-32. doi: 10.1007/s13748-016-0094-0.
12. Newman ME. Clustering and preferential attachment in growing networks. Phys Rev E Stat Nonlin Soft Matter Phys.Ping, S. K.(2018,march 15). Towards Data Science. Retrieved from:. 2001 accuracy-precision-recall-or-f1-331fb37c5cb9;64(2 Pt 2):025102. doi: 10.1103/PhysRevE.64.025102, PMID 11497639.
13. Salton G, McGill M. Introduction to modern information retrieval. New York: McGraw-Hill; 1983.
14. Zhou T, Lü L, Zhang YC. Predicting missing links via local information. Eur Phys J B. 2009;71(4):623-30. doi: 10.1140/epjb/e2009-00335-8.
15. Angeles MdP, Gomez-Adorno H, Corza-Vargas V, Hernandez-Guevara S. Current State of the graph, training and testing datasets for Link prediction problem; 2020. Available from: https://github.com/pilarang/Dathaton.

Cite this article: Angeles MDP, Adorno HG, Hernández-Guevara SD, Corza-Vargas VM. Predicting Collaborations among Research Scientists: A Datathon Experience. J Scientometric Res. 2025;14(1):300-8.