Gendered Pathways in Science: Exploring the Impact of Researcher Gender on Academic Success

Team: Deniel Shumeiko, Zed Liu. **Project Mentor TA:** Yufei Wang. **Link to Notebook:**

https://colab.research.google.com/drive/1gWvLq8kUlsm2lYMKQLdy7byK5wUfys6b?usp=sharing

1) Abstract

In this project, we investigate the impact of researcher gender on academic success, using the H-index as our primary target metric. This study is crucial as it provides insights into the factors influencing a researcher's academic impact and examines potential gender bias in academic settings.

Our primary goal was to compile an extensive dataset of researchers' backgrounds, affiliations, and publications in the Human-Computer Interaction (HCI) domain. We then aimed to predict academic impact quantified by the H-index by using Machine Learning (ML) algorithms such as Linear Regression, AdaBoost Regression, Random Forest, and Feed Forward Neural Networks (FFNN).

We created a dataset with 3,453 rows and 17 columns, covering features like H-index, total citations, university ranking, and gender. Our models showed strong predictive power for the H-index. By toggling the gender feature in the testing set, we found no major gender bias in HCI research. Our investigation also found that out of all ML methods, custom-made FFNN showed the highest accuracy on the test dataset, making it the model of choice for the proposed problem.

2) Introduction

Problem Setup: We aim to predict the H-index of researchers in the Human-Computer Interaction (HCI) field using machine learning techniques, including Linear Regression, Random Forests, AdaBoost, and Feed Forward Neural Networks (FFNN). Our feature vector includes total citations, citations in the last 5 years, gender, title, university affiliation, overall university ranking, country of the university, total publications, and max citations per paper.

Data Collection: We use a custom script to scrape data from Google Scholar on HCI researchers. The features are chosen based on prior studies and include additional considerations such as university affiliation, ranking, researcher's position (e.g., PhD student, Senior Scientist), gender, and country of the university.

Motivation: Understanding the key factors behind academic success helps guide career decisions and enhance impactful research. This analysis identifies the strongest predictors of academic success measured by the H-index. By including gender as a feature, we assess if gender influences academic success, aiming to confirm that merit, rather than systemic biases, determines success.

3) Prior Work

1. Lucaweihs. (n.d.). *Impact-prediction: Predict author h-index and paper citation counts on the dataset underlying semantic scholar.* GitHub. Available at: https://github.com/Lucaweihs/impact-prediction
2. Acuna, D., Allesina, S., & Kording, K. (2012). *Predicting scientific success.* Nature, 489(7415), 201–202. https://doi.org/10.1038/489201a
3. Kong, X., Zhang, J., Zhang, D., Bu, Y., Ding, Y., & Xia, F. (2020). *The Gene of Scientific Success.* ACM Transactions on Knowledge Discovery from Data, 14(4), Article 41. https://doi.org/10.1145/3385530

4) Summary of Contributions

Apart from building our own dataset with custom-engineered features in an unexplored HCI field, none of the prior research mentioned above used neural networks to tackle the problem of H-index prediction. Therefore, we will try to see if our ways of implementing FFNN will yield a more accurate model with a more accurate prediction than more traditional methods of Machine Learning used in the prior research.

5) Detailed Description of contributions

5.1) Data Collection

(i) Automated Data Extraction for Human-Computer Interaction Research

Our project involves automating data extraction from Google Scholar to create datasets for machine learning (ML) training. We focus on researchers in the Human-Computer Interaction (HCI) domain. Using the `Scholarly` Python library, we efficiently gather information without triggering CAPTCHAs. The custom code begins by defining a function named `fetch_scholars`, which retrieves researcher details and stores them in CSV files. The Scholars DataFrame contains 8,800 entries with key information, such as Name, Affiliation, H-index, I10-Index, Total Citation, and Citations in the Last 5 years. A second dataset, the Publications DataFrame, holds 518,149 records, capturing Author Name, Year, Title, and Citations.

(ii) Data Cleaning and Preprocessing

After data extraction, we clean and preprocess the datasets to ensure they're ready for analysis and ML training. This step involves addressing inconsistencies, handling missing values, and organizing data into a structured format.

The `Affiliation` column presented a challenge because researchers often write their affiliations in inconsistent formats. To extract meaningful organization names, we used the `Spacy` library with the `en_core_web_lg` language model to extract patterns from the unstructured text. We focused on specific keywords like "university," "college," "institute," "school," "academy," "polytechnic," and "faculty" to maintain relevance to our research on academic affiliations.

(iii) Title Extraction and Encoding

Similarly, for the `title` column, we created the `define_title` function to identify professional titles from descriptions. This function assigns a value to each title based on its hierarchy, with "Professor" receiving the highest rank of 10, while roles like "Student" or "Lecturer" have a lower value. Titles that could not be classified are assigned a value of 0. This structured approach helps standardize titles for further analysis.

(iv) Incorporating Institutional and Country Data

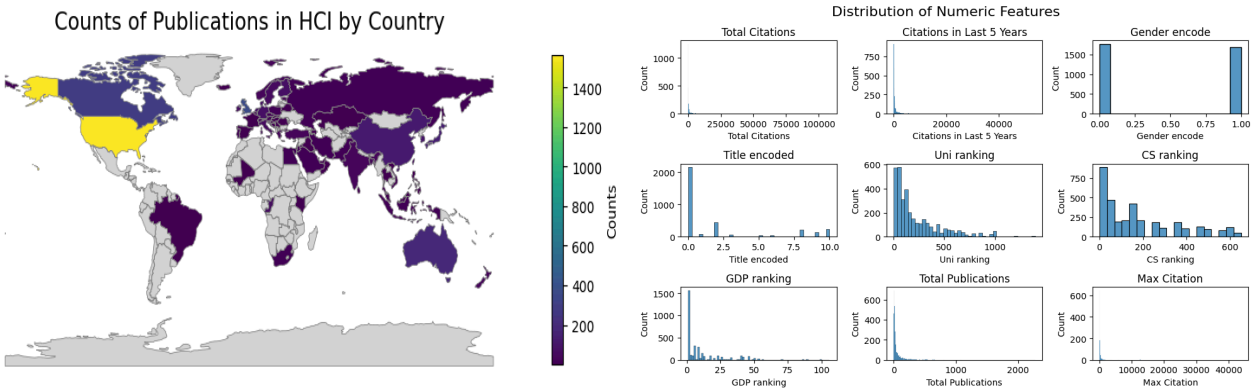
To understand the impact of institutional affiliations, we integrated the QS World University Ranking and the QS Subject Ranking for Computer Science into our dataset. We also added country information with GDP rankings to evaluate the economic context of academic productivity. These additions allow us to analyze the influence of institutional reputation and economic background on academic success.

(v) Final Dataset and Future Steps

The final dataset contains 16 feature columns and has been cleaned and structured for ML training. It holds 3,500 instances, providing a robust sample size for exploring trends in HCI research. Our next step is to consolidate the publication data with the scholars' data, focusing on total publications and maximum citations to understand

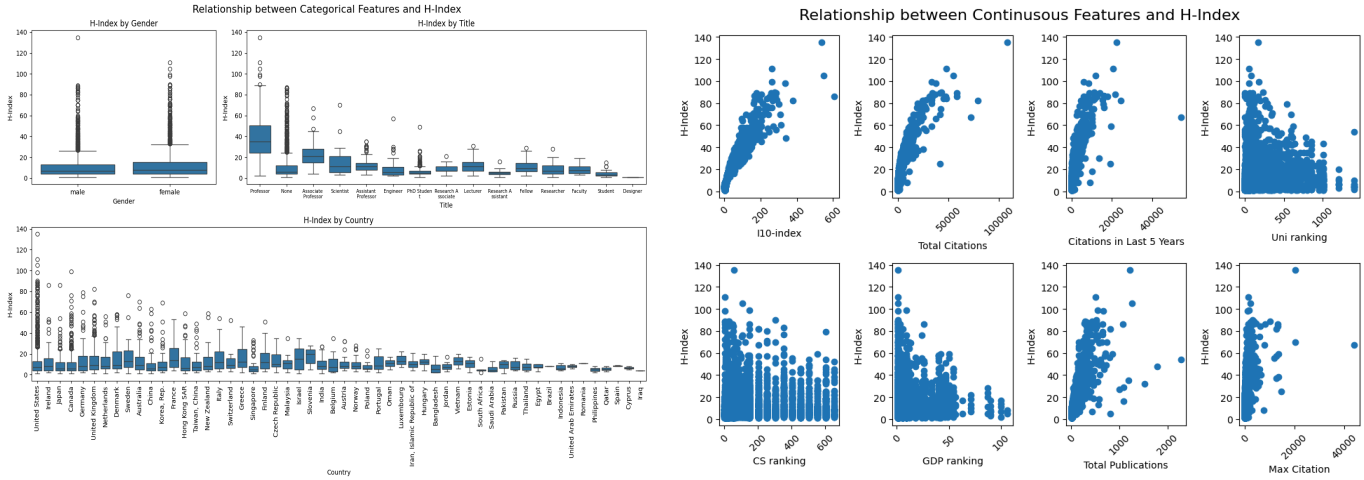
productivity and the impact of key papers. This comprehensive dataset sets the stage for deeper analysis and robust machine learning modelling.

(vi) Understanding the Data Distribution



The world map on the left visualizes the distribution of Human-Computer Interaction (HCI) researchers by country, indicating that the majority are based in North America, primarily in the USA and Canada. However, a significant number of researchers come from other regions, highlighting the global scope of the HCI domain.

The histograms on the right display the distribution of various numeric features from our dataset. A key observation is that while most researchers have 0-200 publications, a small minority have over 2,000 publications. Additionally, the distribution of researchers' gender is nearly even, indicating a balanced representation between males and females in the dataset, which supports the assertion that HCI is a diverse field. The histograms also reveal patterns in other features, such as total citations, citations in the last five years, title encoding, and university ranking, offering insights into the dataset's complexity and diversity.



(vii) Inspecting Relationships between features and H-index

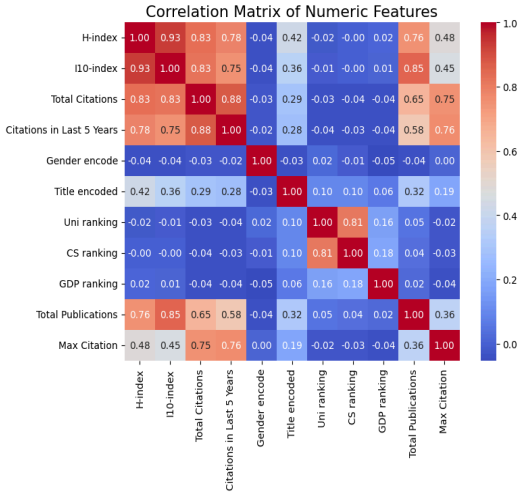
The composite figure illustrates the relationship between the H-index and various features, both categorical and continuous.

For categorical variables like gender, academic title, and country affiliation, the findings suggest that gender has minimal impact on the H-index, with male and female researchers showing similar medians. The academic title reveals that higher-ranking positions, such as 'Professor,' generally correspond to higher H-indices. The country-based analysis shows researchers from the United States typically have higher H-indices, likely due to a robust research network.

For continuous features, the i10-index and total citations have a positive correlation with the H-index, though the effect of total citations tapers off at high counts. Citations from the last five years follow a similar trend. University and CS rankings show a broader spread, indicating less consistent correlations with the H-index. GDP rankings also vary, suggesting other factors might play a role. Total publications and max citations generally align with higher H-indices, but these relationships are not absolute.

(viii) Confusion Matrix

A correlation heatmap revealed that the H-index has strong positive relationships with citation-based metrics like the i10-index, total citations, and citations in the last 5 years, highlighting the importance of citation counts for academic impact. Moderate correlations were observed between academic titles and total publications, as well as between max citation and the H-index, indicating that higher titles and high citation counts can boost the H-index, but other factors matter too. Institutional rankings, GDP ranking, and gender encoding showed low to negligible correlations with the H-index.

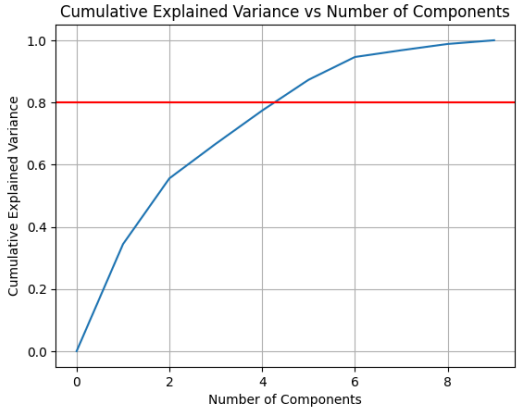


These insights indicate that while certain factors like academic title and total citations strongly correlate with the H-index, other variables show more complex or varied relationships, offering a deeper understanding of academic success in the Human-Computer Interaction domain.

5.2) Algorithm

(i) Principal Component Analysis

To enhance our model's performance, we assessed the impact of Principal Component Analysis (PCA). After splitting the data into training and testing sets, we standardized both using `StandardScaler()` due to PCA's sensitivity to scale. We then applied PCA, creating `X_train_pca` and `X_test_pca`. The Cumulative Explained Variance plot indicated that five components, explaining 80% of the variance, were optimal.



PCA identifies key factors distinguishing researchers by examining the top features of the first three principal components. The first principal component (PC1) is driven by citation metrics, including citations in the last five years (0.5347), total citations (0.5343), and maximum citation count (0.4671). The second principal component (PC2) highlights institutional factors like computer science rankings (0.6737), university rankings (0.6714), and GDP rankings (0.2433). The third principal component (PC3) focuses on demographic and institutional factors such as gender encoding (-0.9897), academic titles (0.1008), and GDP rankings (-0.0895).

In summary, PC1 emphasizes citation metrics, PC2 focuses on institutional rankings, and PC3 highlights gender and institutional factors, providing a comprehensive view of researcher diversity.

(ii) Linear Regression

We instantiated the LinearRegression model, fitted it to the scaled training data, and calculated R^2 values for both training and testing sets. Predictions were made on the scaled testing data, and the Mean Squared Error (MSE)

was computed to evaluate accuracy. PCA did not enhance prediction performance, suggesting the initial features were optimal.

In coefficient analysis, Total Citations (8.0697) and Citations in the Last 5 Years (5.3121) significantly impacted the H-index, highlighting academic impact. Max Citation (-5.2126) showed a negative relationship, indicating complexity. Total Publications (3.2360) and Title Encoded (1.9859) positively influenced the H-index. Institutional factors like Uni ranking (-0.7871) had a negative impact, while CS ranking and GDP ranking had minor positive effects. Gender Encode (0.1064) had minimal influence, indicating an inclusive evaluation framework.

In summary, citation metrics, publication volume, academic titles, and institutional factors are key determinants of scholarly impact.

(iii) AdaBoost Regression

We optimized an AdaBoost Regressor using grid search, exploring 36 hyperparameter combinations with 3-fold cross-validation. The best hyperparameters were a learning rate of 0.1, an exponential loss function, and 200 estimators. The optimized model achieved an R^2 of 0.9664 on the training set and 0.9415 on the testing set, with a testing MSE of 12.7068, indicating strong generalization.

Feature importance analysis revealed that Total Citations (0.7691) was the most critical factor, followed by Total Publications (0.1547). Max Citation (0.0460) and Citations in the Last 5 Years (0.0233) had moderate impacts. University Ranking (0.0040), CS Ranking (0.0028), Title Encoded (0.00006), and Gender Encode (0.0) had minimal to no influence. GDP Ranking (0.0) had no impact.

In summary, the AdaBoost model effectively captures patterns, with citation history being the most significant predictor of academic success, while institutional prestige, titles, gender, and economic conditions have minimal impact.

(iv) Random Forest Regression

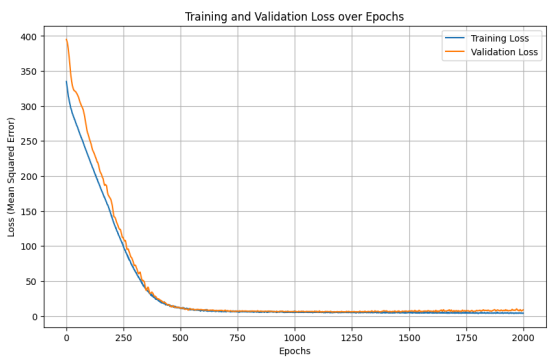
For the Random Forest Regressor, we optimized the model using a grid search across 216 parameter combinations, leading to 648 fits. The optimal hyperparameters were bootstrap=True, max_depth=10, min_samples_leaf=1, min_samples_split=2, and n_estimators=100. The model achieved an R^2 of 0.9950 on the training set and 0.9640 on the testing set, with a testing MSE of 7.8255, indicating strong predictive performance and generalization.

Feature importance analysis showed that Total Citations (0.9374) was the most critical factor. Total Publications (0.0365) and Max Citation (0.0148) had notable impacts. University Ranking (0.0042), Citations in the Last 5 Years (0.0036), CS Ranking (0.0017), GDP Ranking (0.0010), Title Encoded (0.0006), and Gender Encode (0.0003) had minimal influence.

In summary, citation metrics, particularly total citations, are the strongest predictors of academic success, while institutional prestige, economic conditions, and personal titles have minimal impact.

(v) Feedforward Neural Network

We trained a feedforward neural network with a four-layer architecture using batch normalization, dropout, and ReLU activation. Hyperparameters were tuned using three-fold cross-validation, and the model achieved an R^2 of 0.97 on the testing set.



The plot of loss curves over 2000 epochs shows trends in model convergence and generalization. Initially, both training (blue) and

validation (orange) loss curves decline rapidly, indicating effective learning. Around 300 epochs, the loss curves stabilize, suggesting the model has learned most patterns. However, after 1000 epochs, the validation loss increases while the training loss decreases, indicating overfitting.

Overall, the model learned patterns well and performed effectively on testing data but showed signs of overfitting with extended training.

5.3) Gender Bias Experiment

To examine the influence of gender on academic impact, we conducted an experiment by modifying the gender attribute within the testing dataset and reassessing the H-index using pretrained models, including a linear regression model, a random forest model, an AdaBoost model, and a feedforward neural network.

The predictive accuracy, evaluated using the R^2 metric, remained high despite these changes to the gender data. This consistency suggests that altering the gender attribute of researchers did not significantly impact the predicted H-index values, indicating minimal gender bias in terms of research influence within the Human-Computer Interaction (HCI) domain. This experiment underscores the potential for a gender-neutral assessment of academic influence, reinforcing the objectivity of the H-index as a measure in this field.

6) Compute/Other Resources Used

In our investigation, we used the Colab-Pro version to train all ML models. The choice of Colab-Pro is a necessity as the training process of FFNN requires a lot of GPU power, and our current computers do not possess that. Apart from that, it allowed us to run training of multiple models at the same time, significantly reducing the time required to obtain the desired results for the project. There is also an alternative to use the General Purpose Cluster (GPC) provided for Penn PhD students; however, Colab-Pro turned out to be a more user-friendly option.

7) Conclusion

After conducting our investigation, several notable findings emerged. One key insight is that Linear Models, as used by previous researchers, are the least effective method for H-index prediction. In the CIS 5190 class, we learned that while Linear Regression is a simple and powerful method, it often fails to accurately predict target features. This highlights a trade-off between computational demand and model accuracy. In the context of investigating gender effects on H-index prediction, this trade-off is significant due to the high potential societal impact of inaccurate modelling. Therefore, incorporating more sophisticated methods like Random Forests, AdaBoost, and Feedforward Neural Networks is not only ethically justified but necessary to avoid costly mistakes. Our main contribution is developing more reliable methods for H-index prediction than linear regression, providing frameworks for other researchers to ensure robust and trustworthy findings in this field.

Another achievement of this project is the development of web-scraping tools that facilitate data collection, a major bottleneck in social science research. Our code is general and customizable for various related uses. Future research can leverage this web-scraping code to investigate H-index prediction in different scientific fields such as chemistry, biology, mathematics, etc. However, one unresolved issue is the inability to automate the systematic incorporation of researcher gender into the dataset, which remains a challenge for full automation. Future work could address this using more advanced web-scraping tools.

Finally, our key takeaway is that no statistically significant gender bias was found in predicting the H-index of researchers. All machine learning methods used consistently indicated the absence of gender bias. This finding has significant societal implications, confirming that academic success is not predetermined by gender but by factors within the researcher’s control, such as institutional choice and the quantity and quality of their publications. This discovery should encourage researchers to strive for excellence, reinforcing the idea that "hard work pays off" without the concern of systematic discrimination in academia.

Broader Dissemination Information:

Your report title and the list of team members will be published on the class website. Would you also like your pdf report to be published? **NO**

Midway report:

Gendered Pathways in Science: Exploring the Impact of Researcher Gender on Academic Success

Team: Deniel Shumeiko, Zed Liu. **Project Mentor TA:** Yufei Wang. **Link to Notebook:**

https://colab.research.google.com/drive/1gWvLq8kUlsm2lYMKQLdy7byK5wUfys6b?usp=sharing

1) Introduction

Set up the Problem: We aim to predict the H-index of researchers in the Human-Computer Interaction (HCI) field using machine learning (ML) techniques, including linear regression, random forests, AdaBoost, and customized feed-forward neural networks (FFNN). To create our feature vector, we gather data from Google Scholar, focusing on the following features: total citations, citations in the last 5 years, gender, title, university affiliation, overall university ranking, country of the university, total publications, and max citations per paper.

Data Collection: We use a custom script to scrape data from Google Scholar on HCI researchers due to the field's interdisciplinary nature. The focus on these diverse features is based on prior studies in other fields, with additional consideration given to university affiliation, university ranking, the researcher's position (e.g., PhD student, Senior Scientist, Assistant Professor), gender, and country of the university.

Motivation: Understanding the key factors behind academic success helps guide career decisions and maximize impactful research. Our analysis identifies which factors most strongly predict academic success as measured by the H-index. Considering gender as a feature, this study seeks to assess if gender plays a role in academic success, given historical gender-based disparities. The broader goal is to confirm that success is determined by merit rather than systemic biases related to gender.

2) How We Have Addressed Feedback From the Proposal Evaluations

Our TA raised two key questions during the proposal evaluation: (i) How would we collect data? (ii) Would we use ML techniques other than linear regression and FFNN?

To address our TA's feedback, we created a custom script using `Scholarly` to scrape Google Scholar and extract the required data into a CSV file. We also employed the `Spacy` library to process researcher descriptions, collecting information on affiliations, countries, and titles. In response to suggestions for additional ML techniques, we incorporated Random Forests and AdaBoost alongside Linear Regression and Feed-Forward Neural Networks (FFNN). This comprehensive approach, informed by CIS 5190 lectures, allows us to explore a broader array of ML methods for predicting the H-index.

3) Prior work: Citations:

1. Lucaweihs. (n.d.). *Impact-prediction: Predict author h-index and paper citation counts on the dataset underlying semantic scholar.* GitHub. Available at: https://github.com/Lucaweihs/impact-prediction
2. Acuna, D., Allesina, S., & Kording, K. (2012). *Predicting scientific success.* Nature, 489(7415), 201–202. https://doi.org/10.1038/489201a
3. Kong, X., Zhang, J., Zhang, D., Bu, Y., Ding, Y., & Xia, F. (2020). *The Gene of Scientific Success.* ACM Transactions on Knowledge Discovery from Data, 14(4), Article 41. https://doi.org/10.1145/3385530

4) What We are Contributing

Apart from building our own dataset with custom-engineered features in an unexplored field of Human Computer Interaction, none of the prior research mentioned above used neural networks to tackle the problem of H-index prediction. Therefore, we will try to see if our ways of implementing FFNN will yield a more accurate model with a more accurate prediction than more traditional methods of Machine Learning used in the prior research.

5) Detailed Description of Each Proposed Contribution, Progress Towards It, and Any Difficulties Encountered So Far

Automated Data Extraction for Human-Computer Interaction Research

Our project involves automating data extraction from Google Scholar to create datasets for machine learning (ML) training. We focus on researchers in the Human-Computer Interaction (HCI) domain. Using the `Scholarly` Python library, we efficiently gather information without triggering CAPTCHAs. The custom code begins by defining a function named `fetch_scholars`, which retrieves researcher details and stores them in CSV files. The Scholars DataFrame contains 8,800 entries with key information, such as Name, Affiliation, H-index, I10-Index, Total Citation, and Citations in the Last 5 years. A second dataset, the Publications DataFrame, holds 518,149 records, capturing Author Name, Year, Title, and Citations.

Data Cleaning and Preprocessing

After data extraction, we clean and preprocess the datasets to ensure they're ready for analysis and ML training. This step involves addressing inconsistencies, handling missing values, and organizing data into a structured format.

The `Affiliation` column presented a challenge because researchers often write their affiliations in inconsistent formats. To extract meaningful organization names, we used the `Spacy` library with the `en_core_web_lg` language model to extract patterns from the unstructured text. We focused on specific keywords like "university," "college," "institute," "school," "academy," "polytechnic," and "faculty" to maintain relevance to our research on academic affiliations.

Title Extraction and Encoding

Similarly, for the `title` column, we created the `define_title` function to identify professional titles from descriptions. This function assigns a value to each title based on its hierarchy, with "Professor" receiving the highest rank of 10, while roles like "Student" or "Lecturer" have a lower value. Titles that could not be classified are assigned a value of 0. This structured approach helps standardize titles for further analysis.

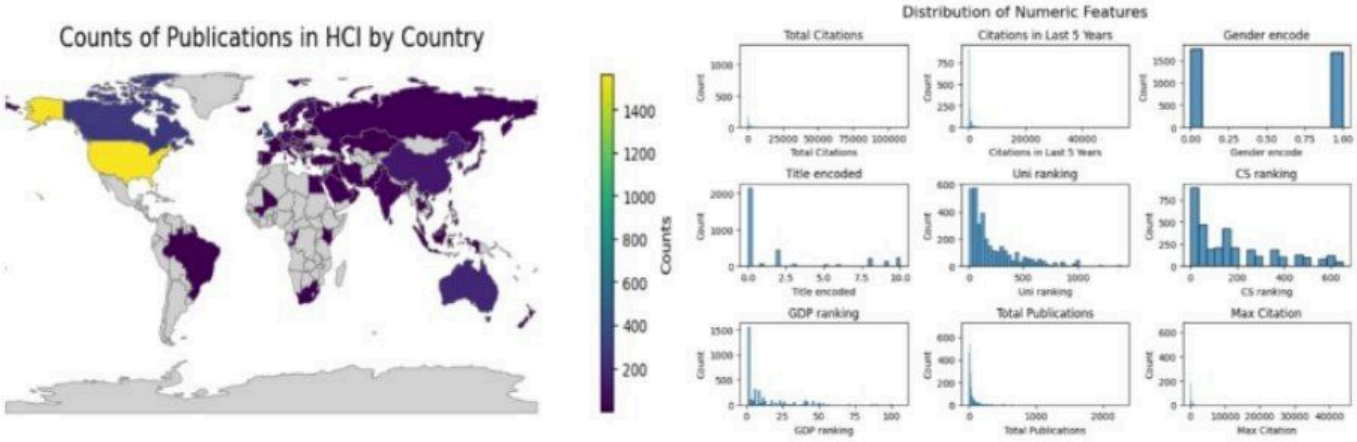
Incorporating Institutional and Country Data

To understand the impact of institutional affiliations, we integrated the QS World University Ranking and the QS Subject Ranking for Computer Science into our dataset. We also added country information with GDP rankings to evaluate the economic context of academic productivity. These additions allow us to analyze the influence of institutional reputation and economic background on academic success.

Final Dataset and Future Steps

The final dataset contains 16 feature columns and has been cleaned and structured for ML training. It holds 3,500 instances, providing a robust sample size for exploring trends in HCI research. Our next step is to consolidate the publication data with the scholars' data, focusing on total publications and maximum citations to understand productivity and the impact of key papers. This comprehensive dataset sets the stage for deeper analysis and robust machine learning modelling.

Understanding the Data Distribution



The world map on the left visualizes the distribution of Human-Computer Interaction (HCI) researchers by country, indicating that the majority are based in North America, primarily in the USA and Canada. However, a significant number of researchers come from other regions, highlighting the global scope of the HCI domain.

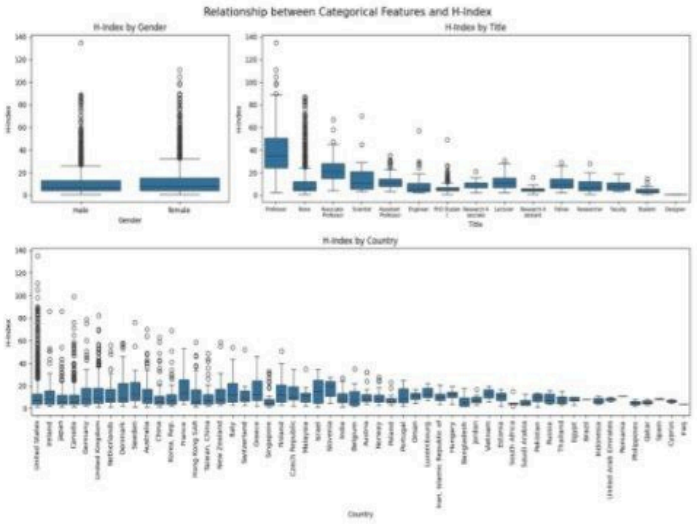
The histograms on the right display the distribution of various numeric features from our dataset. A key observation is that while most researchers have 0-200 publications, a small minority have over 2,000 publications. Additionally, the distribution of researchers' gender is nearly even, indicating a balanced representation between males and females in the dataset, which supports the assertion that HCI is a diverse field. The histograms also reveal patterns in other features, such as total citations, citations in the last five years, title encoding, and university ranking, offering insights into the dataset's complexity and diversity.

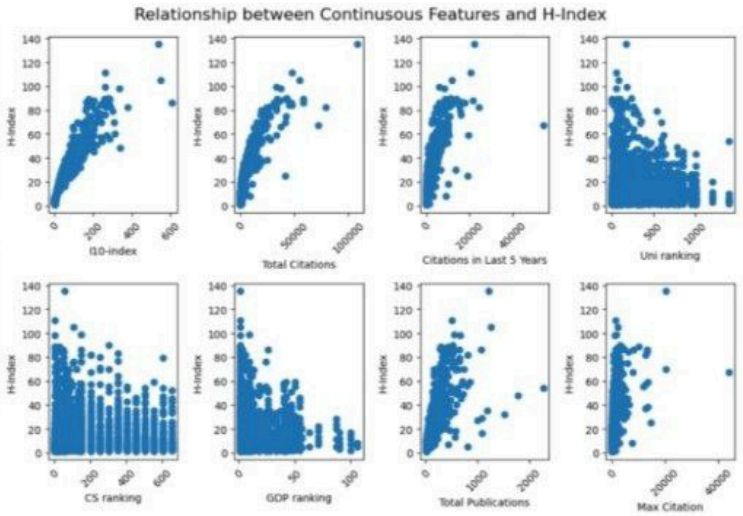
Inspecting Relationships between features and H-index

The composite figure illustrates the relationship between the H-index and various features, both categorical and continuous.

For categorical variables like gender, academic title, and country affiliation, the findings suggest that gender has minimal impact on the H-index, with male and female researchers showing similar medians. The academic title reveals that higher-ranking positions, such as 'Professor,' generally correspond to higher H-indices. The country-based analysis shows researchers from the United States typically have higher H-indices, likely due to a robust research network.

For continuous features, the i10-index and total citations have a positive correlation with the H-index, though the effect of total citations tapers off at high counts. Citations from the last five years follow a similar trend. University and CS rankings show a broader spread, indicating less consistent correlations with the H-index. GDP rankings also vary, suggesting other factors might play a role. Total publications and max citations generally align with higher H-indices, but these relationships are not absolute.

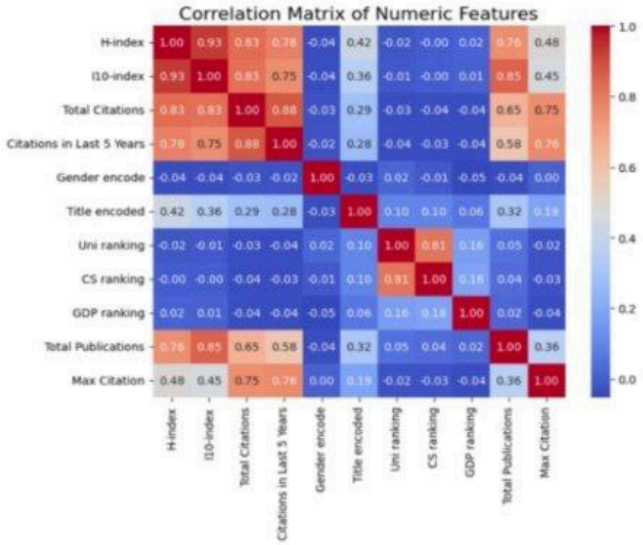




These insights indicate that while certain factors like academic title and total citations strongly correlate with the H-index, other variables show more complex or varied relationships, offering a deeper understanding of academic success in the Human-Computer Interaction domain.

Confusion Matrix

A correlation heatmap revealed that the H-index has strong positive relationships with citation-based metrics like the i10-index, total citations, and citations in the last 5 years, highlighting the importance of citation counts for academic impact. Moderate correlations were observed between academic title and total publications, as well as between max citation and the H-index, indicating that higher titles and high citation counts can boost the H-index, but other factors matter too. Institutional rankings, GDP ranking, and gender encoding showed low to negligible correlations with the H-index.



6) Risk Mitigation Plan

Several potential risks might impact our project. In terms of data collection, some researchers may have multiple affiliations with different universities. In these cases, we will select the affiliation with the highest title according to our code (e.g., "Professor Emeritus" is the highest, "PhD student" is the lowest). If a researcher changed institutions, we'll use their most recent affiliation or the one with the higher title.

Another risk is gathering gender information, as Google Scholar doesn't provide this data. To address this, we will manually look up the gender of researchers but only include those with clear gender information to avoid ambiguity.

Training our models could also be computationally intensive, but we plan to use resources like PENN GPC and Colab Pro to mitigate this issue. Lastly, we need to conduct thorough feature selection to avoid including irrelevant or highly correlated features with our target labels, ensuring model accuracy and efficiency.

CIS 5190 Final Project Checkin Feedback

Inbox ×

Yufei Wang <yufwang@seas.upenn.edu> Wed, May 1, 9:48 PM (12 days ago)

to Zed, me

Hi all,

Good job with your current progress! As we talked about before, creating the dataset itself is a large contribution of your project and you have done a good job at it and your data visualization is really nice in the report. It seems like the feature distributions are pretty right-skewed. You might want to take that into account when you are selecting model families to use, as some of them have the assumption of a normal distribution. Feel free to reach out if you have any questions!

Best regards,
Yufei Wang