



دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر



بررسی الگوی تبعیض و تنوع در دانشجویان تحصیلات تکمیلی دانشگاه‌های برتر آمریکای شمالی

پایان‌نامه برای دریافت درجه کارشناسی
در رشته مهندسی مهندسی کامپیوتر خوشه‌ی نرم‌افزار

سیده بهاران خاتمی

شماره دانشجویی

۸۱۰۱۹۵۳۸۷

استاد راهنما:

دکتر بهنام بهرک

مرداد ۱۴۰۰

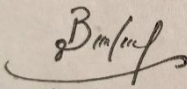
بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

تعهدنامه اصالت اثر
باسمه تعالی

اینجانب سیده بهاران خاتمی تأیید می‌کنم که مطالب مندرج در این پایان‌نامه حاصل تلاش اینجانب است و به دستاوردهای پژوهشی دیگران که در این نوشته از آنها استفاده شده است مطابق مقررات ارجاع گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم سطح یا بالاتر ارائه نشده است. کلیه حقوق مادی و معنوی این اثر متعلق به دانشکده فنی دانشگاه تهران می‌باشد.

نام و نام خانوادگی دانشجو : سیده بهاران خاتمی

امضای دانشجو :



۱۴۰۰ / ۵ / ۸

تقدیم به: خانواده‌ام که همواره در تمامی مراحل زندگی از من حمایت کردند.
همچنین تقدیم به تمامی افراد اقلیت که در هر کجا مورد تبعیض قرار گرفته‌اند.

.....

تشکر و قدردانی:

تشکر ویژه از دکتر بهنام بهرک که در مسیر این پروژه با دانش و صبوری خود مرا یاری کردند.

بدون کمک ایشان به پایان بردن این پژوهش ممکن نبود.

چکیده

بدون شک تحصیلات تکمیلی یکی از مهمترین عوامل تاثیرگذار در زندگی آکادمیک و شغلی افراد است. همچنین سطح علمی دانشگاه و استاد راهنما تاثیر بسزایی در کیفیت تحصیلات تکمیلی دارد. لذا مساله‌ی تبعیض^۱ در بحث آکادمیک یکی از مسائلی است که همواره مورد توجه قرار گرفته است. همچنین مساله‌ی تنوع و شمول^۲، از مسائلی است که اخیراً در دانشگاه‌ها و شرکت‌ها بسیار مورد توجه قرار گرفته است. در این تحقیق به دنبال پاسخ دادن به سوالات زیر هستیم:

- آیا در پذیرش‌های تحصیلات تکمیلی دانشگاه‌های برتر جانبداری جنسیتی و ملیتی وجود دارد؟ به این معنی که استادها تمایل به پذیرش دانشجویان با ملیت و جنسیت یکسان با خود را داشته باشند.
- در صورت وجود این جانبداری، آیا تاثیر منفی بر روی عملکرد علمی آن استاد و گروه تحقیقاتی خواهد داشت؟
- تاثیر تنوع و شمول ملیتی و جنسیتی بر این عملکرد به چه صورت است؟

نتایج حاصل از کار محاسباتی این تحقیق نشان می‌دهد جانبداری ملیتی در پذیرش تحصیلات تکمیلی وجود دارد اما وجود این جانبداری جنسیتی قابل اثبات با استفاده از داده‌ی موجود نیست. همچنین نشان خواهیم داد که این جانبداری ملیتی تاثیری منفی بر عملکرد علمی گروه تحقیقاتی خواهد داشت. همچنین نشان می‌دهیم که تنوع ملیتی در گروه منجر به افزایش کارایی علمی آن می‌شود. در ارتباط با تنوع جنسیتی ادعایی نمی‌توانیم بکنیم. در انتها نیز مسائل جالب قابل بررسی دیگری را بر روی داده‌ی موجود مصورسازی کرده و بررسی می‌کنیم.

نتایج کار این تحقیق قابل استفاده و استناد برای عادلانه‌تر کردن سیستم پذیرش دانشجویی و شمول بیشتر گروه‌های اقلیت در موقعیت‌های تحصیلی و شغلی است. همچنین منجر به افزایش کارایی فعالیت‌های علمی خواهد شد.

کلمات کلیدی: جانبداری و تبعیض - تنوع و شمول - استاد دانشجو - تحصیلات تکمیلی - ملیت - جنسیت - آزمون آماری - جامعه‌شناسی علم

¹ bias and discrimination

² diversity and inclusion

فهرست مطالب

فصل ۱: مقدمه و بیان مساله.....	۱
۱-۱- مقدمه و بیان مساله.....	۲
۱-۲- تاریخچه‌ای از موضوع تحقیق.....	۳
۱-۳- اهداف و آرمان‌های کلی تحقیق.....	۴
۱-۴- ساختار پایان‌نامه.....	۴
فصل ۲: معرفی داده.....	۵
۲-۱- معرفی دادگان.....	۶
۲-۲- طریقه‌ی جمع‌آوری داده و چالش‌ها.....	۷
۲-۳- محدودیت‌های جمع‌آوری داده.....	۸
۲-۴- پیش‌پردازش داده.....	۸
۲-۵- مصورسازی ابتدایی داده.....	۹
فصل ۳: بررسی الگوی جانبداری جنسیتی و ملیتی و ارتباط آن با عملکرد علمی.....	۱۳
۳-۱- بیان مساله.....	۱۴
۳-۲- بررسی وجود الگوی جانبداری ملیتی و جنسیتی.....	۱۴
۳-۳- مدل کردن عملکرد علمی.....	۱۷
۳-۴- بررسی ارتباط جانبداری با عملکرد.....	۱۸
فصل ۴: بررسی الگوی تنوع و شمول ملیتی و جنسیتی و ارتباط آن با عملکرد علمی.....	۲۲

۴-۱- بیان مساله.....	۲۳
۴-۲- مدل کردن تنوع.....	۲۳
۴-۳- بررسی ارتباط تنوع و شمول با عملکرد.....	۲۴
فصل ۵: بررسی سوالات جانبی در داده.....	۲۶
۵-۱- بررسی زیرشاخه‌های علوم کامپیوتر.....	۲۷
۵-۱-۱- آماره‌های زیرشاخه‌ها.....	۲۷
۵-۱-۲- بین رشته‌ای بودن یا متمرکز بودن؟.....	۲۹
۵-۱-۳- نسبت خانم و آقا در هر زیرشاخه.....	۳۱
۵-۱-۴- تنوع در هر زیرشاخه و دانشگاه.....	۳۴
۵-۲- مقایسه‌ی عملکرد اقلیت‌ها در برابر اکثریت‌ها.....	۳۵
۵-۲-۱- جنسیت.....	۳۵
۵-۲-۲- ملیت.....	۳۸
5-3- مصورسازی گرافی داده.....	۴۰
۵-۳-۱- ارتباط کشورها در روابط استاد-دانشجو.....	۴۰
۵-۳-۲- ارتباط زیرشاخه‌های علوم کامپیوتر.....	۴۳
فصل ۶: جمع‌بندی.....	۴۴
فصل ۷: مراجع.....	۴۷

فصل ۱

فصل ۱:

مقدمه و بیان مساله

در این فصل نخست به بیان مقدمات کار، تاریخچه‌ای کوتاه از مساله تحقیق و روش کلی تحقیق پرداخته، سپس مساله و موضوع مورد بررسی در این پایان نامه و اهداف و آرمان‌های کلی تحقیق را بیان می‌کنید و در نهایت به ساختار پایان نامه‌ی پیش رو اشاره خواهید کرد.

۱-۱- مقدمه و بیان مساله

مسالهی تبعیض در بحث آکادمیک یکی از مسائلی است که همواره مورد توجه قرار گرفته است. تبعیض در آکادمیا تبعیضی است که منجر می‌شود عقاید دانشمندان کار پژوهشی آن‌ها را تحت‌الشعاع قرار داده یا گروه پژوهشی آن‌ها را شکل دهد [1]. در این حیطه، تحقیق‌های متعددی انجام شده که نشان‌دهنده‌ی این تبعیض در ابعاد مختلف جنسیتی، نژادی، ملیتی و ... باشد که در بخش بعد به تفکیک به آن‌ها اشاره خواهیم کرد. شناسایی و رفع این عوامل که منجر به تبعیض می‌شوند از اهمیت بالایی برخوردار است چرا که هم از نظر اخلاقی مورد اهمیت است و علاوه بر تاثیر بر دانشجو، بر دانشگاه و استاد مربوطه نیز تاثیرگذار است زیرا اگر ملاک انتخاب بر اساس شایستگی افراد باشد و نه عوامل نامربوط مانند جنسیت یا ملیت، عملکرد دانشگاه نیز در درازمدت بهبود بسزایی خواهد داشت. در اینجا ما به بررسی این تبعیض می‌پردازیم که آیا استادها تمایل به پذیرش دانشجویان با ملیت و جنسیت مشابه خود را دارند؟ این نوع تبعیض در بخش تبعیض در شکل‌گیری گروه پژوهشی استاد جا می‌گیرد. همچنین بررسی می‌کنیم که تاثیر این جانبداری در صورت وجود، بر عملکرد علمی این استادها و گروه‌های تحقیقاتی آنان چیست.

مسالهی دیگری که در این پژوهش مورد بررسی است مسالهی تنوع و شمول است که اخیرا بسیار مورد توجه قرار گرفته است. تنوع به معنای هر بعد و شاخصه‌ای است که برای ایجاد تمایز و تفکیک بین گروه‌ها و افراد مورد استفاده قرار می‌گیرد. تنوع به معنای این است که گروه مورد نظر بازتابی از جامعه‌ای باشد که در آن است. این شاخصه‌ها انواع متعددی دارند نظیر جنسیت، ملیت، قومیت، مذهب، زبان، سن و تحصیلات. در این مساله ما دو پارامتر ملیت و جنسیت را به‌عنوان شاخصه‌ی تنوع مورد بررسی قرار می‌دهیم. شمول به معنای برابری و رفتار منصفانه و دسترسی یکسان اعضا با شاخصه‌های متنوع به منابع و موقعیت‌ها است. شمول به این معناست که به افراد متنوع یک گروه با وجود این تنوع‌ها احترام گذاشته و آن‌ها احساس مشارکت داده شدن در و متعلق بودن به گروه را داشته باشند [2]. در این تحقیق تنوع گروه‌های تحقیقاتی از نظر ملیتی و جنسیتی را به صورت کمی مدل کرده و تاثیر این تنوع بر عملکرد علمی گروه را بررسی می‌کنیم.

۲-۱- تاریخچه‌ای از موضوع تحقیق

در رابطه با تبعیض در آکادمیا پژوهش‌های متعددی انجام شده است؛ ولی هیچ کدام از این پژوهش‌ها به بررسی تبعیض استاد در پذیرش دانشجو با ملیت و جنسیت یکسان نپرداخته است. در تعدادی از این پژوهش‌ها بررسی شده که آیا تبعیض جنسیتی در پذیرش دانشجو وجود دارد یا خیر. به طور مثال آیا شانس آقایان در گرفتن این پذیرش بالاتر است یا خیر [3][4][5]. تعدادی از این پژوهش‌ها وجود این تبعیض را در زمینه‌ی نژادی و افراد اقلیت بررسی کرده‌اند. به طور مثال آیا افراد سیاه‌پوست شانس کمتری در دریافت پذیرش دانشجویی دارند یا خیر [4][5][6][7]. در دسته‌ای دیگر از این پژوهش‌ها اثرات پارامترهای مختلف نظیر معدل، آزمون GRE، توصیه‌نامه‌ها و ... بر منصفانه بودن پذیرش‌های تحصیلات تکمیلی بررسی شده است [8][9]. در مقاله‌ای دیگر، تاثیر همه‌گیری بر تبعیض جنسیتی در آکادمیا بررسی شده است [10]. در این مقاله نشان داده شده که پاندمی تبعیض جنسیتی در آکادمیا و ریسرچ را افزایش داده و در این دوره نسبت خانم‌هایی که مقاله منتشر می‌کنند به شدت نسبت به قبل کاهش پیدا کرده است. همچنین نشان داده شده که این فاصله جنسیتی در کشورهای فقیرتر بدتر شده با وجود اینکه در این کشورها اختلاف جنسیتی در تحقیقات پیش از همه‌گیری کمتر بوده باشد.

در رابطه با تنوع و شمول نیز کارهای پژوهشی صورت گرفته است. مقاله‌های متعددی اظهار داشته‌اند که تنوع به وجودآورنده‌ی نوآوری و خلاقیت است. [11][12][13][14][15]. این مقالات بیان می‌دارند افراد مختلف و علی‌الخصوص افراد اقلیت تجربه‌ها و دغدغه‌هایی دارند که با افراد غالب جوامع متفاوت است. بدین منظور تنوع در گروه سبب می‌شود مساله از زوایای متعددی مورد بررسی قرار گیرد و این موضوع احتمال رسیدن به راه‌حلی خلاقانه را افزایش می‌دهد. با توجه به دانش و بررسی‌های ما، کار محاسباتی اندکی در این زمینه انجام شده است و مساله مورد بررسی بعضاً متفاوت است. در یکی از این کارها بدنبال بررسی این تناقض است که افراد اقلیت کارهای علمی خلاقانه‌تری منتشر می‌کنند و تولید دانش نوآورانه منجر به شغل بهتر می‌گردد ولی این افراد اقلیت معمولاً موقعیت‌های شغلی پایین‌تری دارند و در نهایت این نتیجه‌گیری را می‌کند که کار خلاقانه‌ی این گروه کمتر مورد توجه قرار گرفته و دیده می‌شود [16].

۳-۱- اهداف و آرمان‌های کلی تحقیق

نتایج کار این تحقیق قابل استفاده و استناد برای عادلانه‌تر کردن سیستم پذیرش دانشجویی و شمول بیشتر گروه‌های اقلیت در موقعیت‌های تحصیلی و شغلی است. همچنین می‌تواند منجر به افزایش کارایی فعالیت‌های علمی شود.

۴-۱- ساختار پایان‌نامه

در فصل دوم، به معرفی داده‌ی مورد استفاده در مساله و طریقه‌ی جمع‌آوری و پیش‌پردازش آن خواهیم پرداخت.

فصل سوم در برگیرنده‌ی بررسی الگوی جانبداری جنسیتی و ملیتی در داده و تاثیر این جانبداری بر عملکرد علمی استادان است.

در فصل چهارم در مورد طریقه‌ی مدل کردن تنوع و بررسی اثر این تنوع بر عملکرد علمی گروه صحبت خواهیم کرد.

در نهایت، در فصل پنجم، تعدادی مساله‌ی جانبی جالب در داده معرفی و بررسی می‌کنیم و به جمع‌بندی موضوع می‌پردازیم.

فصل ۲

فصل ۲: معرفی داده

در فصل پیش رو با داده‌ی مورد استفاده در این پروژه آشنا می‌شوید. شاخصه‌های موجود در داده، روش جمع‌آوری و چالش‌ها و محدودیت‌های جمع‌آوری داده معرفی می‌گردد. پیش‌پردازش‌های انجام شده برای تمیز کردن داده بیان می‌گردد. همچنین تعدادی مصورسازی از داده برای آشنایی بهتر با آن خواهیم دید.

۱-۲- معرفی دادگان

داده‌ی مورد استفاده در این پروژه، اطلاعات استادهای دانشگاه‌ها برتر و دانشجویان کارشناسی ارشد، دکتری و پسادکتری آنهاست. این دانشگاه‌ها با توجه به رنکینگ QS^۱ در رشته‌ی علوم کامپیوتر انتخاب شده‌اند و شامل دانشگاه‌های Massachusetts Institute of Technology (MIT)، Stanford، Cornell، Caltech از آمریکا و دانشگاه Waterloo از کانادا است که جزو ۲۰ دانشگاه برتر آمریکای شمالی هستند. از هر دانشگاه ۲۰ استاد به صورت رندوم انتخاب شده‌اند و داده‌ی مربوط به این استادها و دانشجویان آنها به صورت دستی جمع‌آوری شده است تا از جانبداری انتخاب^۲ جلوگیری شود. دادگان جمع‌آوری شده شامل ۲۳۱۵ سطر است. نمونه‌ای از داده در زیر آمده که شامل مجموعه‌ای از ویژگی‌ها برای استاد و مجموعه‌ای از ویژگی‌ها برای دانشجو است.

University	World Field Ranking	North US Field Ranking	Department	Advisor Role	Advisor Name	Advisor Field	Advisor CS Subfield	Advisor Nationality	Advisor Gender	Advisor Citation	Advisor h-index	Advisor Publication Number	Advisor First Paper Year	Advisor Website
MIT	1	1	EECS	Assistant Professor	Guy Bresler	information theory, statistics, theoretical comp...	Theoretical CS	Usa	M	1715	16	37	2006	http://www.mit.edu/~gbresler/

شکل ۱: ویژگی‌های مربوط به استاد در دادگان

Advisee Name	Advisee Degree	Advisee Nationality	Advisee Gender	Start Year	End Year	Advisee Previous School	Advisee Citation	Advisee h-index	Advisee LinkedIn or Website
Mina Karzand	PHD	Iran	F	2009.0	present	EPFL	101	3	https://www.linkedin.com/in/mina-karzand-10a63...

شکل ۲: ویژگی‌های مربوط به دانشجو در دادگان

اسم ویژگی‌ها گویاست و نیازی به توضیح نیست. تنها موردی که نیاز به توضیح دارد تفاوت دو ستون advisor field و advisor cs subfield است. مورد اول به طور کلی زیرشاخه‌هایی از علوم کامپیوتر است که استاد در آنها فعالیت می‌کند. این زیرشاخه‌ها ممکن است بعضاً بسیار ریز و

¹ <https://www.topuniversities.com/university-rankings/university-subject-rankings/2020/computer-science-information-systems>

² Selection bias

تخصصی باشند. در مورد دوم، ۱۱ زیرشاخه کلی^۱ در علوم کامپیوتر در نظر گرفته شده و باتوجه به حوزه‌ی فعالیت علمی استاد، تعدادی از این زیرشاخه‌ها به او نسبت داده شده است. توجه داشته باشید که این تعداد ممکن است یکی یا بیشتر باشد. در بخش‌های آتی موارد استفاده از این ویژگی را باهم خواهیم دید.

۲-۲-۲- طریقه‌ی جمع‌آوری داده و چالش‌ها

دادگان از پیش آماده‌ای شامل موارد مورد نیاز ما وجود نداشت. لذا بر آن شدیم که خودمان داده را جمع‌آوری کنیم. همانطور که در بخش قبل اشاره شد، داده‌ها از وبسایت استادها جمع‌آوری شده است. از آنجایی که قالب html وبسایت استادها با هم تفاوت داشت، امکان استفاده از خزنگر^۲ برای جمع‌آوری داده وجود نداشت و این داده باید بصورت دستی جمع‌آوری می‌شد که بسیار کار دشوار و وقت‌گیری بود و زمان زیادی صرف جمع‌آوری این داده شده است. اطلاعات مربوط به استادان و دانشجویان از وبسایت و linkedin افراد جمع‌آوری شده است. اطلاعات مربوط به citation، h-index، تعداد مقاله و سال اولین مقاله‌ی منتشرشده از google scholar افراد اخذ شده است. اطلاعات مربوط به ملیت، از ترکیب محل تحصیل دبیرستان و کارشناسی افراد (چرا که معمولاً افراد این مقاطع را در کشور خود سپری می‌کنند)، زبان‌هایی که فرد به آن در حد یک بومی مسلط است (افراد در linkedin خود زبان‌هایی که بلدند را با درجه‌ی تسلط خود به آن زبان ذکر می‌کنند) و وبسایتی^۳ که ملیت و جنسیت افراد را بر اساس نام و نام خانوادگی آن‌ها تشخیص می‌دهد حاصل شده است. در صورتی که از ترکیب این اطلاعات ملیت قابل تشخیص نبود، به این افراد ایمیل زدیم

1

AI, Communication & Security- Computer Architecture & Operating Systems- Computer Graphics- Parallel & Distributed Computing- Databases & Information Management- Scientific Computing- Software, Languages & Compilers- Hardware- Theoretical CS- HCI

² crawler

³ <https://quecst.qcri.org/tool/Name2GAN>

و از آن‌ها خواستیم تا این اطلاعات را با ما به اشتراک بگذارند. در صورتی که از این طریق هم نتیجه‌ای نمی‌گرفتیم ویژگی ملیت یا جنسیت این سطرها را خالی گذاشتیم.

۲-۳- محدودیت‌های جمع‌آوری داده

موارد محدودیت‌زایی در جمع‌آوری داده وجود داشت که در زیر به آن اشاره می‌کنیم:

- همانطور که پیش‌تر به آن اشاره شد، از هر دانشگاه به صورت رندوم ۲۰ استاد انتخاب کردیم و داده‌ی مربوط به آن‌ها را جمع‌آوری کردیم. بعضی از این استادها، لیست دانشجویان خود را در وبسایت‌شان درج نکرده بودند. به ناچار در این موارد یک عدد رندوم دیگر تولید کردیم و داده‌ی استاد جدید را به جای آن جمع کردیم.
- داده‌ها با توجه به وبسایت استادها جمع‌آوری شده است. این داده‌ها با توجه به آخرین به‌روز رسانی استاد از وبسایتش انجام شده است. اگر دانشجویی از وی در وبسایتش ذکر نشده باشد، ما اطلاعی از این مورد نداشتیم.
- بعضی از استادها، در طول کار خود دانشگاه محل کار خود را تغییر دادند. تمامی دانشجویان وی تحت عنوان آخرین آخرین دانشگاه استاد جمع‌آوری شده است.

۲-۴- پیش‌پردازش داده

از آنجایی که داده به صورت دستی جمع‌آوری شده است، امکان بروز اشتباهاتی در آن هست. لذا تمیز کردن دادگان از اهمیت بالایی برخوردار است. مجموعه‌ای از اقدامات در این مورد انجام شده که در زیر به آن‌ها اشاره می‌کنیم:

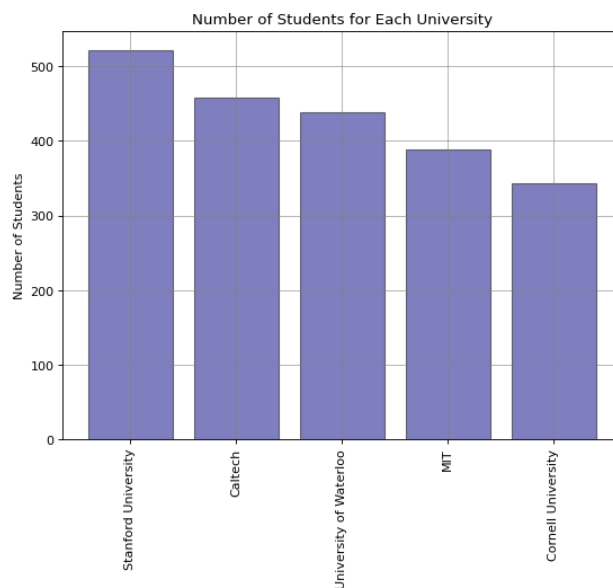
- همانطور که اشاره شد، برخی از سطرها مقداری برای جنسیت و ملیت ندارند. تعداد این سطرها ۱۶۵ تا از ۲۳۱۵ سطر داده است. از آنجایی که در ادامه تحلیل ما وابسته به این مقادیر است، این سطرها از داده حذف شدند.

- مقادیر رشته‌های مربوط به ملیت بسیار مهم است که یکدست باشد چرا که جلوتر تحلیل‌ها را بر اساس گروه کردن داده با توجه به این ویژگی انجام می‌دهیم. کوچک و بزرگ بودن حروف این ستون همسان‌سازی شد.
- تایپ متغیرها با توجه به مقادیر آن تنظیم و مقداردهی شد.

۵-۲- مصورسازی ابتدایی داده

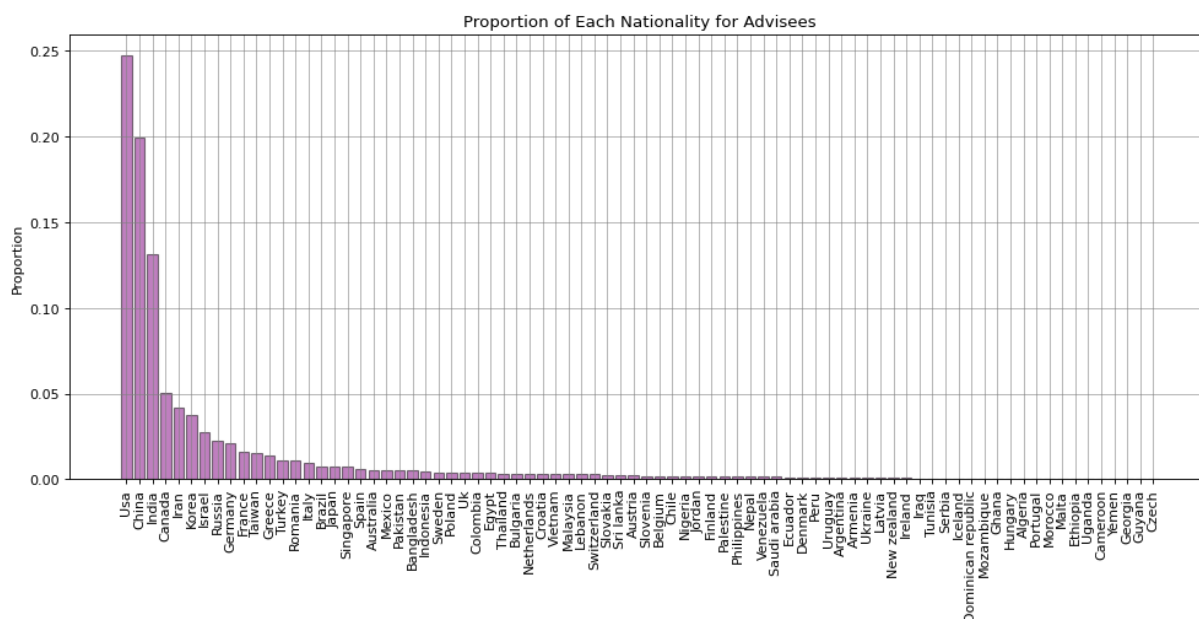
برای درک بهتر از مقادیر داده، تعدادی مصورسازی انجام شده که در زیر نمایش داده می‌شوند:

در هر دانشگاه ۲۰ استاد به صورت رندوم انتخاب شدند. تعداد دانشجویان این استادها با توجه به سال فعالیت آن‌ها متغیر بود. در زیر ترکیب کلی تعداد دانشجویان هر دانشگاه آورده شده است.

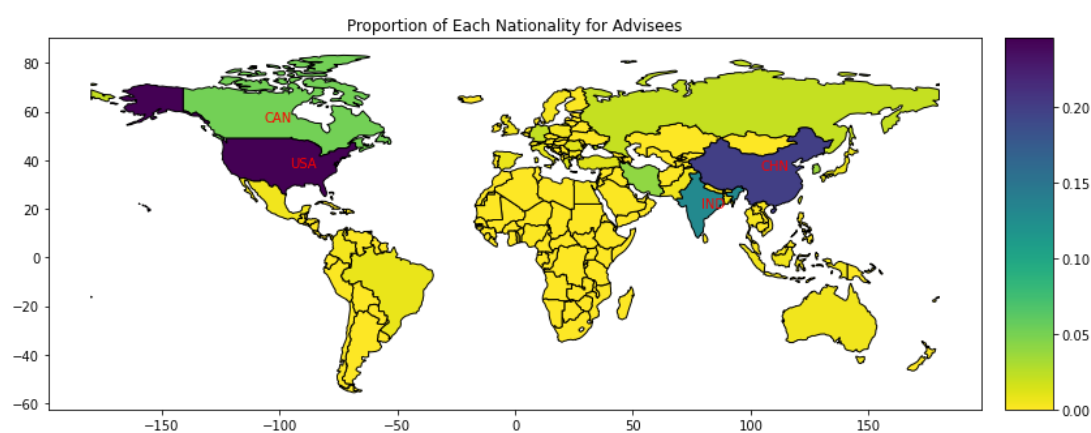


شکل ۳: تعداد دانشجوی هر دانشگاه در دادگان

در زیر نسبت ملیت‌های موجود در دادگان برای استاد و دانشجو نشان داده شده است. این نسبت هم بصورت نمودار میله‌ای و هم بر روی نقشه برای درک بهتر نمایش داده شده است:

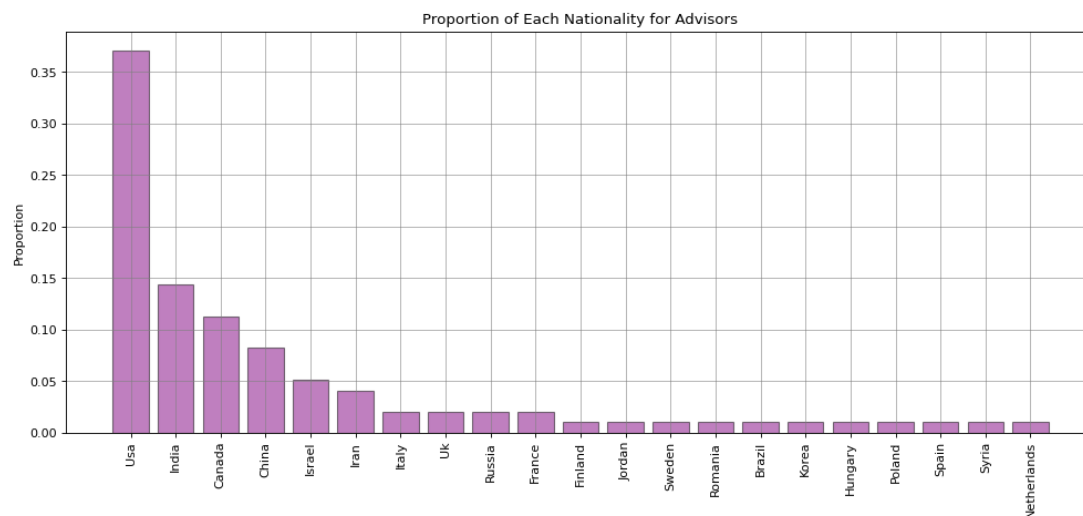


شکل ۴ الف: نمودار میله‌ای توزیع ملیت دانشجویان در دادگان

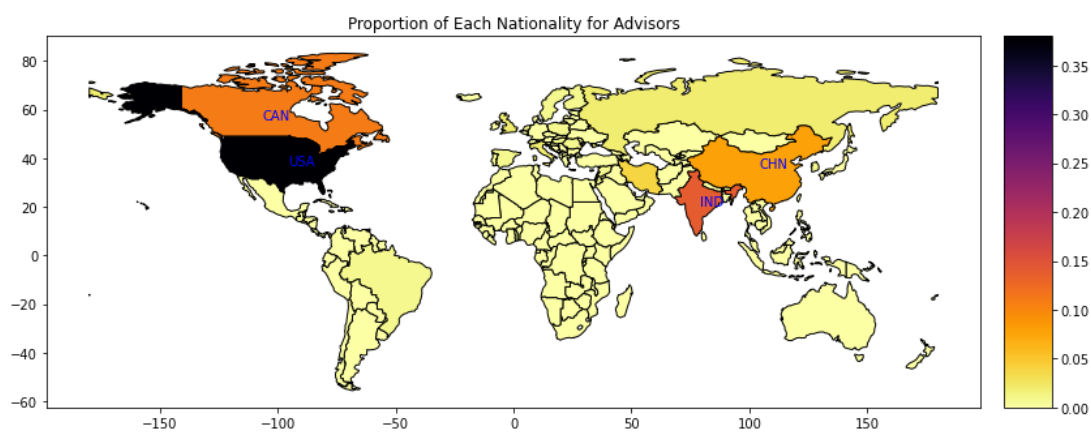


شکل ۴ ب: توزیع ملیت دانشجویان در دادگان بر روی نقشه

همانطور که از نمودارهای بالا مشخص است، نسبت افراد مهاجرت‌کننده برای ادامه‌ی تحصیل در کشورهای آسیایی بیشتر از اروپا و آفریقا است. کشورهای آمریکای شمالی نیز منطبقاً از نسبت بالایی برخوردارند چراکه دانشگاه‌های مورد بررسی در این کشورهاست.



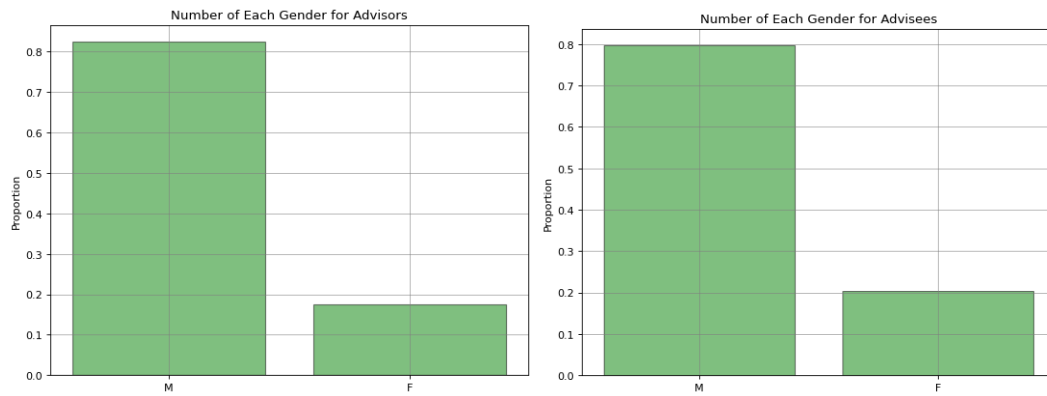
شکل ۴ ج: نمودار میله‌ای توزیع ملیت اساتید در دادگان



شکل ۴ د: توزیع ملیت اساتید در دادگان بر روی نقشه

نکته‌ی ذکر شده در مورد توزیع در قاره‌ها در مورد اساتید نیز در دادگان برقرار است.

در ادامه این توزیع را در رابطه با جنسیت استاد دانشجو نیز نمایش خواهیم داد:



شکل ۵: نمودار میله‌ای توزیع جنسیت دانشجویان و اساتید

فصل ۳

فصل ۳: بررسی الگوی جانبداری جنسیتی و ملیتی و ارتباط آن با عملکرد علمی

فصل سوم در برگیرنده‌ی بررسی الگوی جانبداری جنسیتی و ملیتی در داده و ارتباط این
جانبداری با عملکرد علمی استادان است.

۱-۳- بیان مساله

در این بخش به دنبال بررسی این سوال هستیم که آیا جانبداری‌ای از لحاظ تشابه جنسیت و ملیت استاد و دانشجو در دادگان وجود دارد؟ به این معنی که استادان تمایل بیشتری داشته باشند که دانشجویان هم‌ملیتی و هم‌جنسیتی خود را به عنوان دانشجوی خود بپذیرند. پس از حذف سطرهایی که ملیت یا جنسیت آنها مشخص نبود، ۲۱۵۰ سطر داده باقی می‌ماند که نسبت هم‌ملیت و هم‌جنسیت بودن استاد و دانشجو در دادگان ما به صورت زیر است:

۲۲.۶۰۴٪	درصد زوج استاد دانشجو با ملیت مشابه
۶۹.۴۸۸٪	درصد زوج استاد دانشجو با جنسیت مشابه

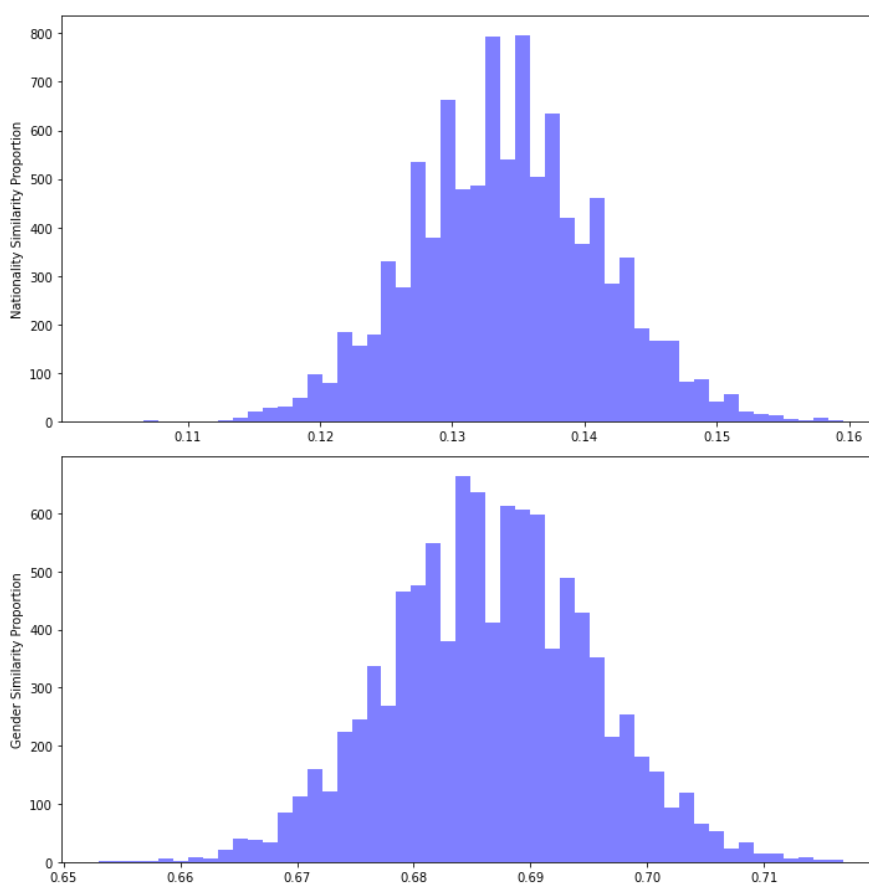
جدول ۱: نسبت زوج‌های استاد دانشجو با ملیت و جنسیت مشابه در دادگان

۲-۳- بررسی وجود الگوی جانبداری ملیتی و جنسیتی

- جهت بررسی اینکه آیا این نسبت معنادار و جانبدارانه است یا خیر، شبیه‌سازی به صورت زیر اجرا می‌کنیم:
- در این شبیه‌سازی، ستون ملیت و جنسیت استاد را تغییر نمی‌دهیم (چراکه تعداد دانشجوی هر استاد متغیر است و این تعداد باید لحاظ شود. به این معنی که اگر یک استاد با ملیت n_1 ، k_1 دانشجو و استادی دیگر با ملیت n_2 ، k_2 دانشجو داشته باشد و با حفظ کلیت مساله فرض کنیم $k_2 > k_1$ ، شانس هم‌ملیت شدن با n_2 بیشتر از n_1 خواهد بود و این مساله باید لحاظ شود تا عدد حاصل از شبیه‌سازی با اعداد جدول فوق قابل مقایسه باشند.
 - برای ستون‌های ملیت و جنسیت دانشجو، از توزیع احتمال ملیت و جنسیت دانشجویان در داده‌ی اصلی نمونه برمی‌داریم و عدد رندوم تولید می‌کنیم. به‌طور مثال اگر نسبت دانشجویان خانم در داده p و نسبت دانشجویان آقا $1-p$ باشد، برای هر سطر از ستون جنسیت دانشجو در داده‌ی شبیه‌سازی، یک مقدار رندوم تولید می‌کنیم که با احتمال p ، F باشد و با احتمال $1-p$ ، M باشد. همین‌طور برای ملیت.

(این توزیع همان نسبت‌هایی است که در شکل‌های ۴ و ۵ نشان داده شده است. این مقادیر را به تعداد سطرهای دادگان مان تولید می‌کنیم.

- حال در داده‌ی شبیه‌سازی شده نسبت زوج‌های دانشجو و استاد با جنسیت و ملیت یکسان را دوباره محاسبه می‌کنیم. در واقع با این کار داریم محاسبه می‌کنیم که اگر توزیع احتمال ملیت و جنسیت را در داده حفظ کنیم و این دانشجویان را به صورت رندوم بین استادها بازتوزیع کنیم، نسبت‌های هم‌ملیتی و هم‌جنسیتی بودن در داده‌ی شبیه‌سازی چقدر با نسبتی که در داده‌ی اصلی دیده‌ایم فرق خواهد داشت و آیا این تفاوت معنادار است یا خیر.
- فرآیند فوق یک‌بار اجرای شبیه‌سازی است. این فرآیند را $n = 10000$ بار تکرار می‌کنیم. ۱۰۰۰۰ عدد نسبت تشابه ملیت و ۱۰۰۰۰ عدد نسبت تشابه جنسیت به ما می‌دهد. توزیع این اعداد به صورت زیر است:



شکل ۶: توزیع نسبت زوج‌های استاد دانشجو با ملیت و جنسیت مشابه در داده‌ی شبیه‌سازی

میانگین نسبت‌های بدست آمده در شبیه‌سازی برای ملیت و جنسیت به صورت زیر است:

۱۳.۴۱۷٪	درصد زوج استاد دانشجو با ملیت مشابه
۶۸.۶۶۸٪	درصد زوج استاد دانشجو با جنسیت مشابه

جدول ۲: نسبت زوج‌های استاد دانشجو با ملیت و جنسیت مشابه در دادگان شبیه‌سازی

حال با توجه به توزیع فوق، باید بسنجیم که نسبت‌های دیده شده در داده‌ی اصلی (جدول ۱) آیا بصورت اتفاقی اتفاق افتاده است یا اختلاف نسبت‌های دادگان اصلی و دادگان شبیه‌سازی معنادار است؛ بدین معنی که فاصله زیادی از نسبت رندوم تولید شده در شبیه‌سازی که بدون جانبداری است داریم و جانبداری در دادگان اصلی وجود دارد. بدین منظور از آزمون فرض استفاده می‌کنیم. p نشان‌دهنده‌ی این نسبت در جامعه و \hat{p} نشان‌دهنده‌ی این نسبت در نمونه که همان دادگان ما است، می‌باشد. این آزمون برای ملیت به صورت زیر است:

$$H_0: p = 0.13417$$

$$H_a: p > 0.13417$$

فرض صفر بیان می‌دارد که جانبداری وجود ندارد و این نسبت در جامعه برابر با حالتی است که این بازتوزیع به صورت تصادفی اتفاق افتاده باشد. فرض مقابل بیان می‌دارد که جانبداری در داده است و این نسبت برابر حالت تصادفی نیست. آزمون فرض یک‌طرفه است چرا که به دنبال نشان دادن این هستیم آیا این نسبت بیشتر از حالت تصادفی و بدون جانبداری است یا جانبدارانه است. جانبدارانه بودن به معنای بزرگتر بودن این نسبت از حالت تصادفی است. برای بررسی فرض مقابل $pvalue$ محاسبه می‌کنیم. $pvalue$ بیان می‌دارد که احتمال دیده شدن داده‌ی مشاهده شده (\hat{p}) با فرض صحیح بودن فرض صفر چقدر است. یعنی چقدر محتمل است که داده‌ی مشاهده شده به صورت تصادفی و شانسی اتفاق افتاده باشد و جانبداری در کار نباشد. به عبارتی:

$$pvalue = P(a \geq \hat{p} \mid H_0)$$

پس از محاسبه‌ی $pvalue$ ، آن را با α که ضریب اتکاست مقایسه می‌کنیم. اگر از α کمتر بود یعنی احتمال شانسی بودن مشاهدات کم است و فرض صفر رد می‌شود. اگر از α کمتر نبود، نمی‌توان فرض صفر را رد کرد. برای α سه مقدار مرسوم ۰/۰۵، ۰/۰۱ و ۰/۰۰۱ در نظر گرفته می‌شود. حال برای محاسبه‌ی $pvalue$ در این مساله از دو روش زیر استفاده می‌کنیم:

روش اول) استفاده‌ی مستقیم از توزیع دیده شده در داده‌ی شبیه‌سازی (شکل ۶) است.

$$pvalue = \frac{\#simulated\ proportion > observed\ proportion}{\#simulation\ trials}$$

با استفاده از این روش، مقدار pvalue برای ملیت و جنسیت به صورت زیر است:

۰	pvalue ملیت
۰.۱۶۵۵	pvalue جنسیت

جدول ۳: pvalue های محاسبه شده با روش اول

روش دوم) توزیع شکل ۶ تا حد خوبی متقارن و نزدیک به توزیع نرمال است. از روش مرسوم z test برای محاسبه‌ی pvalue استفاده می‌کنیم:

۰	pvalue ملیت
۰.۲۰۴	pvalue جنسیت

جدول ۴: pvalue های محاسبه شده با روش دوم

نتیجه‌ی نهایی حال از دو روش یکسان است. pvalue حاصل از ملیت در هر دو روش صفر است و این به معنای رد کردن قاطع فرض صفر یا همان عدم وجود جانبداری در دادگان است. به عبارتی این جانبداری ملیتی وجود دارد و استادها متمایلند که دانشجویان با ملیت مشابه خودشان را بپذیرند. همچنین pvalue حاصل از هر دو روش از هر سه مقدار مرسوم α بزرگتر است و در مورد فرض صفر جنسیت با استفاده از این داده‌ها نمی‌توان حرفی زد و آن را رد کرد. به عبارتی، داده‌های مشاهده شده مدرک کافی برای وجود جانبداری و تمایل برای جذب دانشجوی همجنسیت از سمت استادان را ارائه نمی‌دهد.

۳-۳- مدل کردن عملکرد علمی

برای مدل کردن عملکرد و موفقیت علمی از سه شاخص citation (مجموع تعداد ارجاعات مقالات دیگر

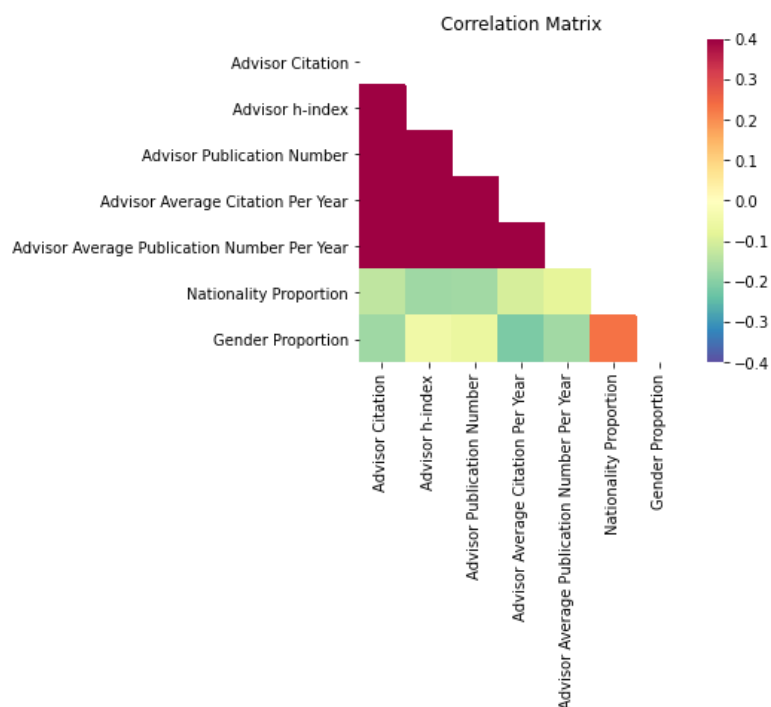
به مقالات فرد)، تعداد مقالات چاپ شده توسط فرد و h-index که شاخص استوار تری در مقایسه با citation است چراکه citation مجموعه‌ی مقالات را در نظر می‌گیرد. در citation زیاد بودن ارجاعات به یک مقاله می‌تواند پارامتر را بطور ناگهانی افزایش دهد ولی h-index اینطور نیست و کیفیت و citation همه مقالات و همچنین تعداد مقالات فرد مهم می‌شود.

در کنار این سه پارامتر، متوسط آن‌ها را نیز در طول مدت زمانی که آن شخص فعالیت پژوهشی انجام داده در نظر می‌گیریم؛ چراکه ممکن است یک استادی تازه استاد شده باشد و citation و تعداد مقاله وی به طور کلی کمتر باشد ولی عملکرد او در همین بازه‌ای که استاد شده بسیار عالی باشد. citation و تعداد مقاله پارامترهایی خطی هستند و به صورت خطی برای هر مقاله جمع زده می‌شوند اما پارامتر h-index با این دو پارامتر متفاوت است. وقتی h-index استادی برابر با k است، یعنی k بزرگترین عددی است بطوریکه وی حداقل k مقاله دارد که به حداقل k بار به آن ارجاع داده شده است. وقتی این مقدار $k+1$ می‌شود یعنی تمامی آن مقالات یک ارجاع بیشتر دریافت کردند و یک مقاله‌ی دیگر نیز به جمع k مقاله قبل افزوده شده است. لذا رفتار این پارامتر خطی نیست و متوسط آن در تعداد سال پژوهش خیلی معنادار نیست.

لذا در کل، ۵ شاخص citation، h-index، تعداد مقاله و متوسط citation و تعداد مقاله در طول زمانی که فرد پژوهش انجام داده (تعداد سالی که از زمان انتشار اولین مقاله‌ی وی گذشته) را در نظر می‌گیریم.

۴-۳- بررسی ارتباط جانبداری با عملکرد

برای بررسی ارتباط جانبداری با عملکرد، همبستگی نسبت یکسان بودن ملیت و جنسیت استاد و دانشجو را با ۵ متریک سنجش عملکرد که در بخش قبل معرفی شد محاسبه می‌کنیم:



شکل ۷: هیت‌مپ همبستگی شاخص‌های جانبداری با شاخص‌های عملکرد

			متوسط تعداد مقاله به سابقه پژوهش	متوسط citation به سابقه پژوهش	متوسط تعداد مقاله به سابقه پژوهش
	citation	h-index	تعداد مقاله	تعداد مقاله	تعداد مقاله
نسبت تشابه ملیت	-0.132	-0.174	-0.168	-0.102	-0.077
نسبت تشابه جنسیت	-0.172	-0.048	-0.064	-0.214	-0.170

جدول ۵: همبستگی شاخص‌های جانبداری با شاخص‌های عملکرد

مقادیر همبستگی منفی است. این به این معناست که این دو شاخص نسبت عکس دارند؛ یعنی در صورت بیشتر شدن شاخص‌های جانبداری، شاخص‌های عملکرد کاهش می‌یابد. اما صرفاً از روی عدد این همبستگی نمی‌توان نتیجه‌ی خاصی گرفت و باید دوباره از آزمون فرض بهره بگیریم تا ببینیم آیا این همبستگی مشاهده شده معنا دار است یا خیر.

$$H_0: corr = 0$$

$$H_a: corr < 0$$

برای محاسبه‌ی ضریب همبستگی از روش spearman استفاده می‌کنیم [17]. برخلاف روش pearson این روش فرض نمی‌کند که دادگان به صورت نرمال توزیع شده‌اند. مقادیر همبستگی آن مانند سایر روش‌ها بین ۱ و -۱ قرار دارد و صفر به معنای عدم وجود همبستگی است. مقادیر pvalue ی متناظر با این ضرایب همبستگی به صورت زیر است:

متوسط تعداد مقاله به سابقه‌ی پژوهش	متوسط citation به سابقه‌ی پژوهش	تعداد مقاله	h-index	citation	نسبت تشابه ملیت
0.019	0.025	0.011	0.025	0.015	نسبت تشابه جنسیت
0.025	0.071	0.217	0.414	0.151	

جدول ۶: pvalue متناظر با ضرایب همبستگی بین شاخص‌های جانبداری با شاخص‌های عملکرد

با $\alpha = 0.05$ که یکی از مقادیر مرسوم برای α است، همبستگی بین شاخص جانبداری ملیت و شاخص‌های عملکرد معنادار بوده و نشان می‌دهد که این دو پارامتر رابطه‌ی عکس با یکدیگر دارند. در مقابل همبستگی بین شاخص جانبداری جنسیت و شاخص‌های عملکرد معنادار نبوده و مقدار pvalue برای آن بزرگتر از α است که با توجه به نتایج بخش قبل که نشان دادیم برای ملیت نمی‌توان نتیجه گرفت که جانبداری وجود دارد، نتیجه‌ی

این بخش نیز با آن همخوانی دارد و نمی‌توان نشان داد که این نسبت تأثیری بر شاخص‌های عملکرد دارد. اما در رابطه با ملیت که قبل‌تر وجود این جانبداری را نشان دادیم، تأثیر آن بر پارامترهای عملکرد قلیل اثبات است.

فصل ۴

فصل ۴: بررسی الگوی تنوع و شمول ملیتی و جنسیتی و ارتباط آن با عملکرد علمی

فصل چهارم حاوی بررسی الگوی تنوع و شمول از لحاظ ملیتی و جنسیتی در داده و بررسی ارتباط این تنوع در گروه‌های تحقیقاتی با عملکرد علمی اساتید است.

۱-۴- بیان مساله

در این بخش، ابتدا تنوع ملیتی و جنسیتی را در گروه‌های تحقیقاتی اساتید به صورت کمی مدل می‌کنیم. سپس همانند بخش ۳-۴ به بررسی رابطه‌ی بین تنوع و شاخص‌های عملکرد معرفی شده در بخش ۳-۳ می‌پردازیم.

۲-۴- مدل کردن تنوع

برای مدل کردن تنوع از مفهوم آنتروپی شانون استفاده می‌کنیم. می‌دانیم آنتروپی یک متغیر تصادفی بیانگر متوسط میزان اطلاعات یا عدم قطعیتی است که در پیشامد آن متغیر تصادفی خوابیده است. به طور مثال زمانی که یک سکه را می‌اندازیم، اگر خروجی با احتمال p شیر و با احتمال $1-p$ خط باشد، این عدم قطعیت زمانی بیشینه می‌شود که p برابر با 0.5 باشد یا بعبارتی توزیع احتمال یکنواخت باشد.

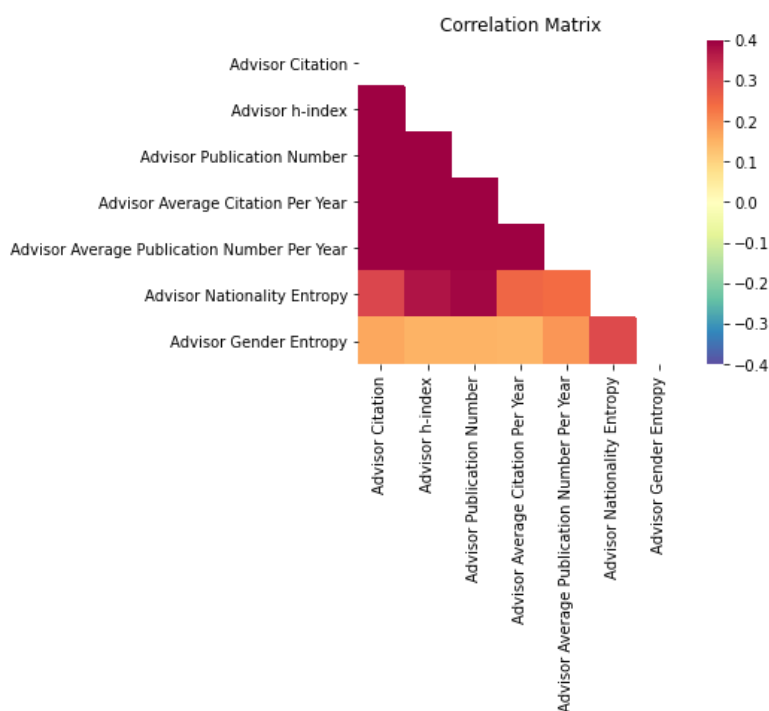
اگر X یک متغیر تصادفی گسسته با پیشامدهای x_1, \dots, x_n باشد، که با احتمال $P(x_1), \dots, P(x_n)$ رخ دهند، آنتروپی متغیر تصادفی X از طریق رابطه‌ی زیر محاسبه می‌گردد:

$$H(X) = - \sum_{i=1}^n P(x_i) \log P(x_i)$$

حال برای محاسبه‌ی تنوع در هر گروه، ابتدا برای هر استاد دو بردار محاسبه می‌کنیم که نشان‌دهنده‌ی توزیع احتمال ملیت و جنسیت در آن گروه است. هر المان بردار ملیت نشان‌دهنده‌ی نسبت دانشجویان آن ملیت در کل گروه است. با توجه به اینکه المان‌ها نامنفی بوده و جمع‌شان برابر با یک است، یک توزیع احتمال هستند و می‌توان آنتروپی آن را محاسبه کرد. هرچقدر این توزیع، به توزیع یکنواخت نزدیک‌تر باشد، تنوع در آن گروه بیشتر است و از ملیت‌ها و جنسیت‌های مختلف به یک میزان در گروه مشارکت دارند. همانطور که کمی قبل‌تر بیان شد، هرچقدر این توزیع احتمال به یکنواخت نزدیک‌تر باشد، عدم قطعیت آن بیشتر بوده و آنتروپی بیشتر است. لذا آنتروپی بیشتر معادل تنوع بیشتر است و بالعکس.

۳-۴- بررسی ارتباط تنوع و شمول با عملکرد

مشابه بخش ۳-۴، ضرایب همبستگی و pvalue متناظر با آن را بین آنالیزهای و شاخصه‌های عملکرد محاسبه می‌کنیم:



شکل ۸: هیت‌مپ همبستگی شاخص تنوع با شاخص‌های عملکرد

	متوسط تعداد مقاله به سابقه‌ی پژوهش	متوسط citation به سابقه‌ی پژوهش	تعداد مقاله	h-index	citation	آنالیزهای ملیتی
آنالیزهای ملیتی	0.242	0.252	0.392	0.372	0.305	آنالیزهای ملیتی
آنالیزهای جنسیتی	0.187	0.149	0.152	0.152	0.166	آنالیزهای جنسیتی

جدول ۷: همبستگی شاخص تنوع با شاخص‌های عملکرد

مقادیر همبستگی مثبت است. این به این معناست که این دو شاخص نسبت مستقیم دارند؛ یعنی در صورت بیشتر شدن شاخص تنوع، شاخص‌های عملکرد افزایش می‌یابد. اما صرفاً از روی عدد این همبستگی نمی‌توان نتیجه‌ی خاصی گرفت و باید دوباره از آزمون فرض بهره بگیریم تا ببینیم آیا این همبستگی مشاهده شده معنا دار است یا خیر.

$$H_0: corr = 0$$

$$H_a: corr > 0$$

مقادیر pvalue ی متناظر با ضرایب همبستگی با استفاده از روش spearman به صورت زیر است:

	citation	h-index	تعداد مقاله	متوسط citation به سابقه‌ی پژوهش	متوسط تعداد مقاله به سابقه‌ی پژوهش
آنتروپی ملیت	2.24×10^{-5}	7.70×10^{-6}	1.70×10^{-7}	2.91×10^{-3}	8.14×10^{-5}
آنتروپی جنسیت	0.073	0.262	0.325	0.034	0.146

جدول ۸: pvalue متناظر با ضرایب همبستگی بین شاخص تنوع با شاخص‌های عملکرد

با $\alpha = 0.01$ (و حتی بجز در یک مورد متوسط citation به سابقه‌ی پژوهش با $\alpha = 0.001$)، همبستگی بین شاخص تنوع ملیت و شاخص‌های عملکرد معنادار بوده و نشان می‌دهد که این دو پارامتر رابطه‌ی مستقیم با یکدیگر دارند. در مقابل همبستگی بین شاخص تنوع جنسیت و شاخص‌های عملکرد معنادار نبوده و مقدار pvalue برای آن بزرگتر از α است و نمی‌توان نشان داد که تنوع جنسیتی تاثیری بر شاخص‌های عملکرد دارد. اما در رابطه با ملیت، تاثیر آن بر پارامترهای عملکرد قابل اثبات است.

فصل ۵

فصل ۵: بررسی سوالات جانبی در داده

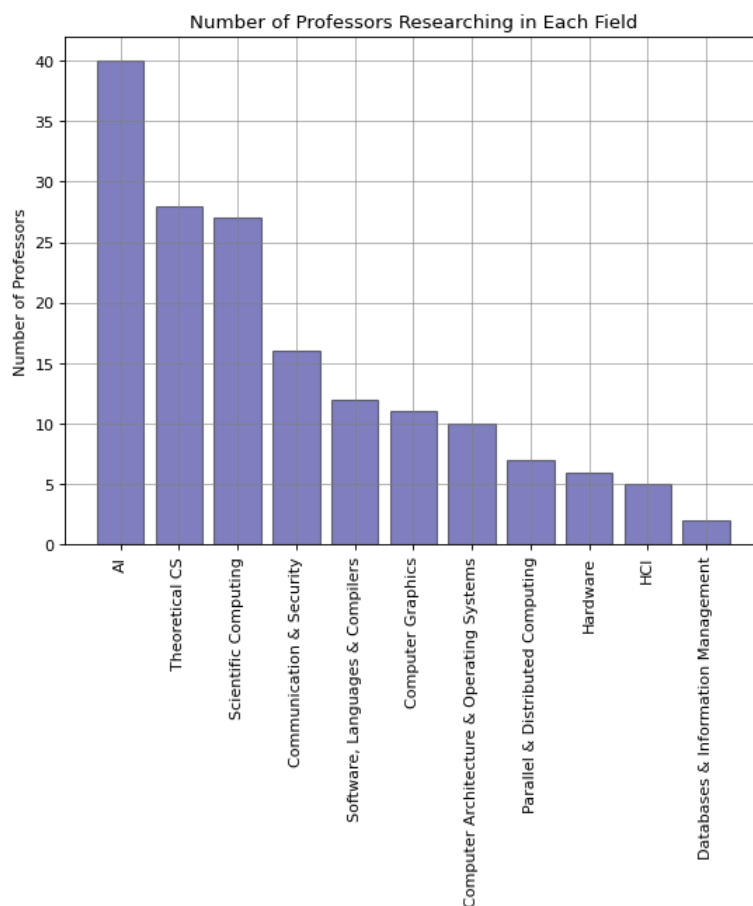
در بخش پایانی این پژوهش، به بررسی سوالات جالب دیگری در این دادگان و مصورسازی‌های بیشتر داده می‌پردازیم.

۱-۵- بررسی زیرشاخه‌های علوم کامپیوتر

همانطور که در فصل دوم به آن اشاره شد، ۱۱ زیرشاخه کلی علوم کامپیوتر را در نظر گرفتیم و به هر استاد تعدادی از این زیرشاخه‌ها را با توجه به حیطه‌ی فعالیت‌شان نسبت دادیم. با توجه به این مشخصه، بررسی‌هایی را در ادامه انجام خواهیم داد.

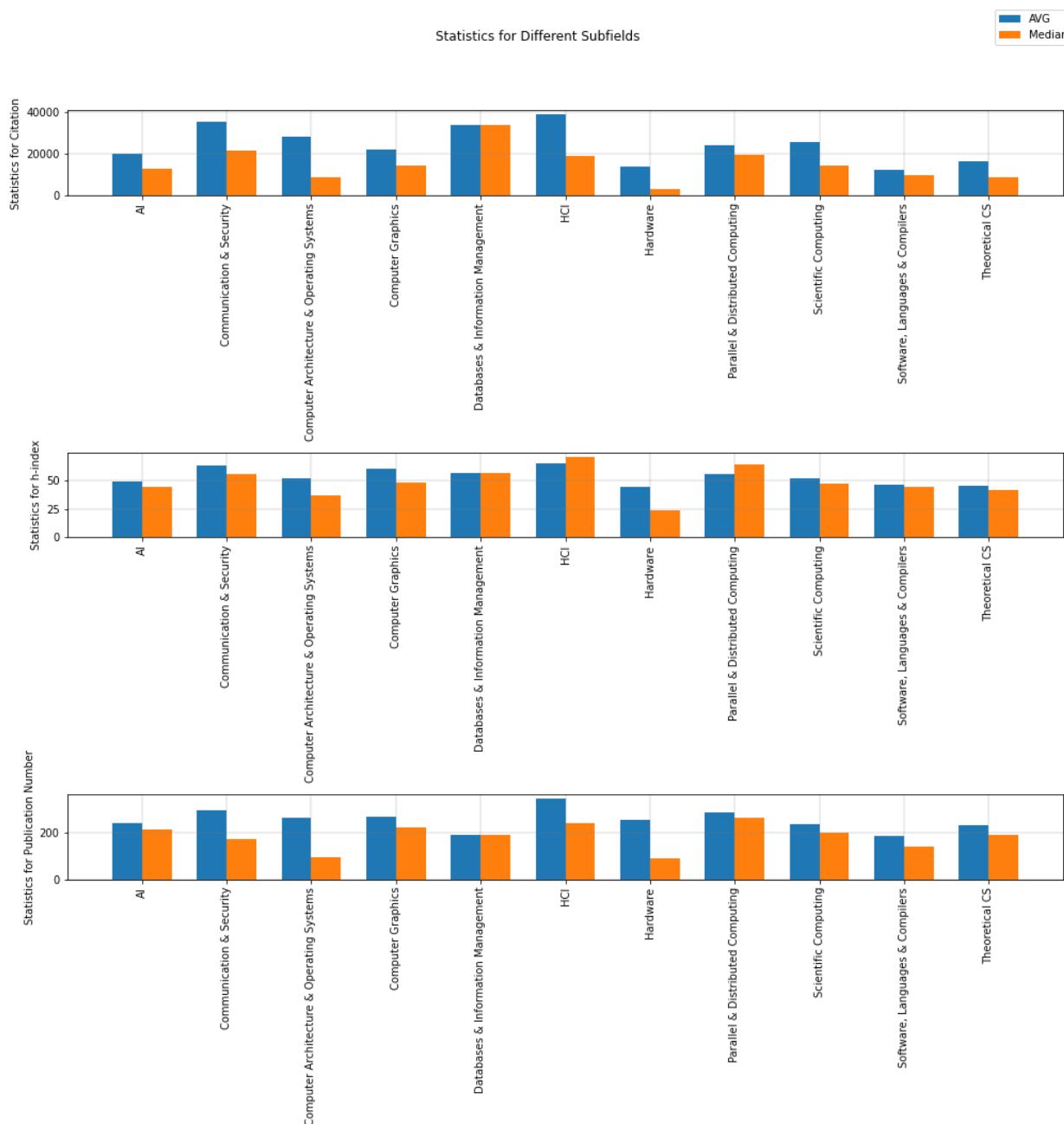
۱-۱-۵- آماره‌های زیرشاخه‌ها

تعداد استادی که در هر زیرشاخه فعالیت می‌کند به شرح زیر است:



شکل ۹: نمودار میله‌ای تعداد استاد هر زیرشاخه‌ی علوم کامپیوتر در دادگان

همچنین میانگین و میانه‌ی سه پارامتر citation، h-index و تعداد مقاله‌ی استادهای هر زیرشاخه به صورت زیر است:



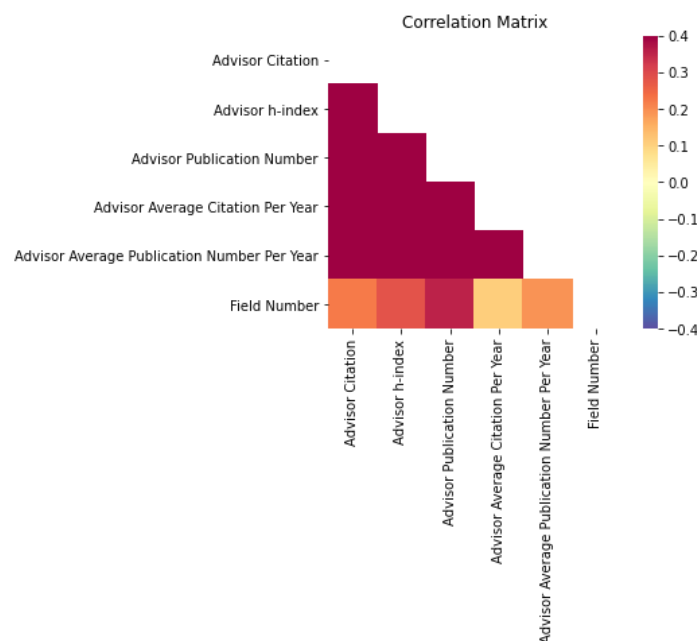
شکل ۱۰: نمودار میله‌ای آماره‌های شاخص‌های عملکرد مربوط به هر زیرشاخه علوم کامپیوتر

البته از روی نمودار فوق نمی‌توان نتیجه‌گیری داشت چرا که برخی از این زیرشاخه‌ها تعداد استاد کمی در دادگان دارند و ممکن است به صورت تصادفی این پارامترهای عملکرد برای همان تعداد کم استاد بالا باشد.

برخی از زیرشاخه‌ها نیز جدیدتر هستند و مدت کمتری از فعالیت‌شان می‌گذرد و یا اخیراً خیلی ترند شده‌اند و استادهای جوان بیشتری سمت آن رفته‌اند. لذا ممکن است متوسط این پارامترها برای این زیرشاخه‌ها کمتر باشد.

۲-۱-۵- بین رشته‌ای بودن یا متمرکز بودن؟

در این بخش بدنبال بررسی جواب این سوال هستیم که ارتباط بین تعداد زیرشاخه‌ای که هر استاد در آن فعالیت می‌کند با شاخصه‌های موفقیت چیست. عبارتی افرادی که فعالیت بین‌رشته‌ای دارند موفق‌ترند یا افرادی که در یک شاخه متمرکز دارند؟ مشابه روندی که در بخش‌های قبل برای محاسبه‌ی ضریب همبستگی و $pvalue$ متناظر با آن انجام می‌دادیم را طی می‌کنیم.



شکل ۱۱: هیت‌مپ همبستگی تعداد زیرشاخه فعالیت اساتید با شاخص‌های عملکرد

متوسط تعداد مقاله به سابقه پژوهش	متوسط citation به سابقه پژوهش	تعداد مقاله	h-index	citation	تعداد زیرشاخه فعالیت اساتید
0.191	0.106	0.354	0.283	0.224	

جدول ۹: همبستگی شاخص تعداد زیرشاخه فعالیت اساتید با شاخص های عملکرد

مقادیر pvalue ی متناظر با ضرایب همبستگی با استفاده از روش spearman به صورت زیر است:

متوسط تعداد مقاله به سابقه پژوهش	متوسط citation به سابقه پژوهش	تعداد مقاله	h-index	citation	تعداد زیرشاخه فعالیت اساتید
0.060	0.106	0.005	0.016	0.017	

جدول ۱۰: pvalue متناظر با ضرایب همبستگی بین شاخص تعداد زیرشاخه فعالیت اساتید با شاخص های عملکرد

با توجه به اینکه مقدار pvalue برای سه متغیر citation، h-index و تعداد مقاله از α کمتر است و برای دوتای دیگر نیست، به نظرم نمی توان بین رشته ای بودن و موفق بودن را لزوماً به هم مرتبط دانست چراکه متوسط عملکرد در زمان ارتباط معناداری با بین رشته ای بودن از خود نشان نداده است. تحلیلی که می توان داشت این است که در اساتیدی که به تنهایی خود citation و h-index بالایی دارند ولی می توانند متوسط این پارامترها برایشان پایین متوسط یا بالا باشد، که یعنی معمولاً استادهایی که قدیمی تر هستند و مدت زمان زیادی استاد هستند، بین دسته از استادها و بین رشته ای بودن ارتباط معنادار هست یعنی استادهای قدیمی تر معمولاً بیشتر بین رشته ای بوده اند یا رشته تخصص کاری شان را تغییر داده اند و استادهای جدیدتر بیشتر به صورت متمرکز روی شاخه های کمتری فعالیت دارند. گویی هرچه پیش برویم فضای علمی تخصصی تر می شود.

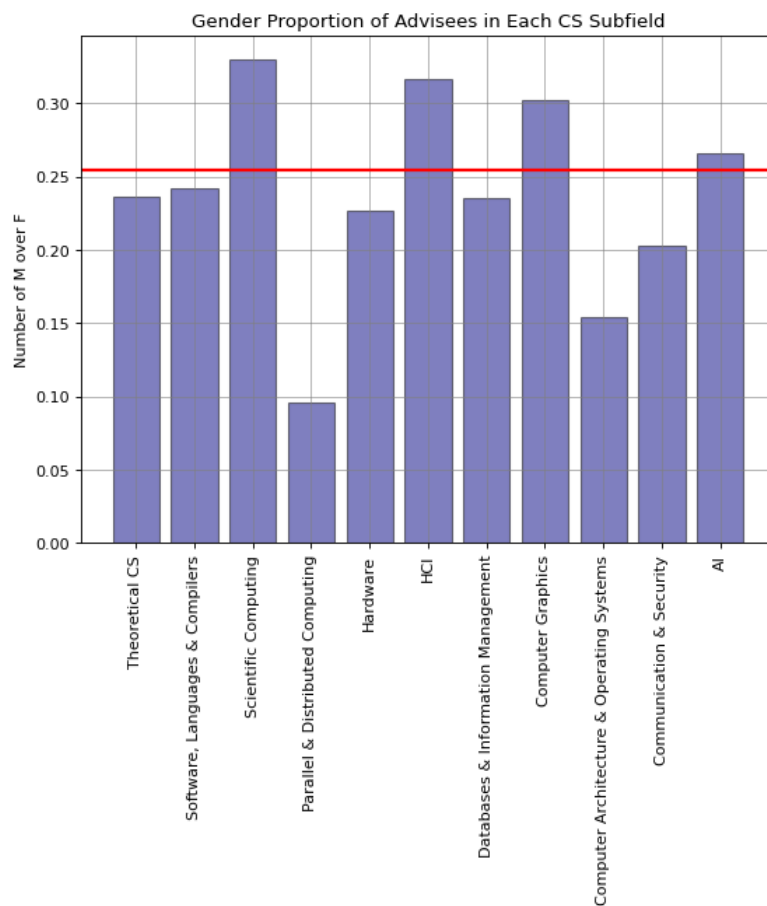
۳-۱-۵- نسبت خانم و آقا در هر زیرشاخه

در دادگان اصلی، در دانشجویان ۴۱۹ خانم و ۱۶۴۷ آقا داریم و نسبت خانم به آقا ۰.۲۵۴ است. حال این نسبت را در هر زیرشاخه می‌بینیم تا بررسی کنیم در هر شاخه کدام جنسیت حضور پررنگ‌تر یا کم‌رنگ‌تری نسبت به نسبت ۰.۲۵۴ دارند.

Subfields	Advisee Gender	size
Theoretical CS	M	457
Theoretical CS	F	108
Software, Languages & Compilers	M	227
Software, Languages & Compilers	F	55
Scientific Computing	M	476
Scientific Computing	F	157
Parallel & Distributed Computing	M	157
Parallel & Distributed Computing	F	15
Hardware	M	106
Hardware	F	24
HCI	M	117
HCI	F	37
Databases & Information Management	M	17
Databases & Information Management	F	4
Computer Graphics	M	192
Computer Graphics	F	58
Computer Architecture & Operating Systems	M	195
Computer Architecture & Operating Systems	F	30
Communication & Security	M	330
Communication & Security	F	67
AI	M	703
AI	F	187

شکل ۱۲: تعداد دانشجوی هر زیرشاخه به تفکیک جنسیت

این نسبت برای زیرشاخه‌ها به صورت زیر است:



شکل ۱۳: نسبت تعداد دانشجوی خانم به آقا در هر زیرشاخه

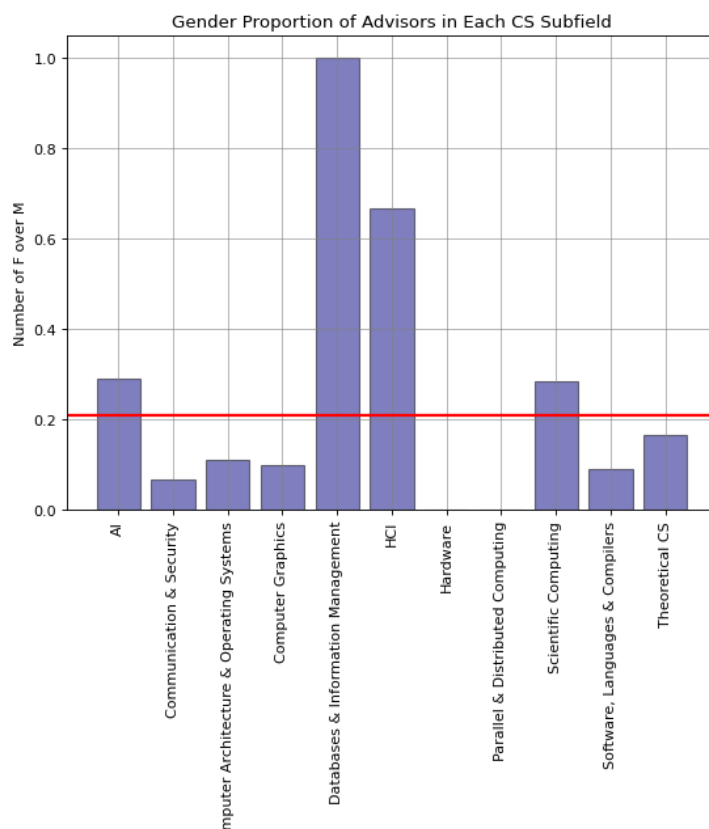
خط قرمز نشان‌دهنده‌ی نسبت تعداد دانشجوی خانم به آقا در کل دادگان است که بالاتر گفتیم ۰.۲۵۴ است.

نسبت خانم به آقا در بین اساتید ۰.۲۰۸ است. تعداد اساتید در هر زیرشاخه به صورت زیر است:

Subfields	Advisor Gender	size
Theoretical CS	M	24
Theoretical CS	F	4
Software, Languages & Compilers	M	11
Software, Languages & Compilers	F	1
Scientific Computing	M	21
Scientific Computing	F	6
Parallel & Distributed Computing	M	7
Hardware	M	6
HCI	M	3
HCI	F	2
Databases & Information Management	F	1
Databases & Information Management	M	1
Computer Graphics	M	10
Computer Graphics	F	1
Computer Architecture & Operating Systems	M	9
Computer Architecture & Operating Systems	F	1
Communication & Security	M	15
Communication & Security	F	1
AI	M	21

شکل ۱۳: تعداد استاد هر زیرشاخه به تفکیک جنسیت

این نسبت برای زیرشاخه‌ها به صورت زیر است:

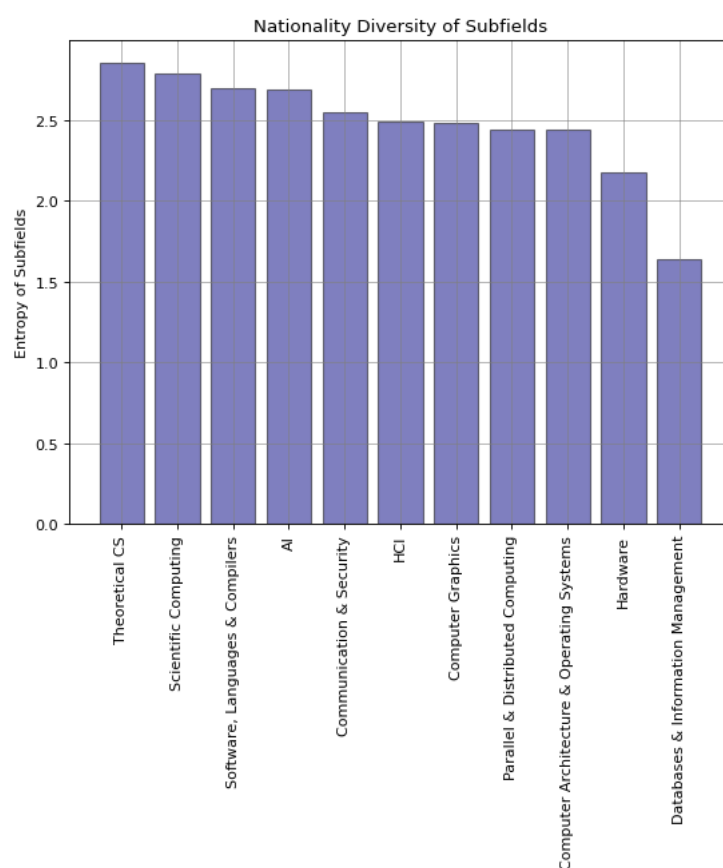


شکل ۱۳: نسبت تعداد استاد خانم به آقا در هر زیرشاخه

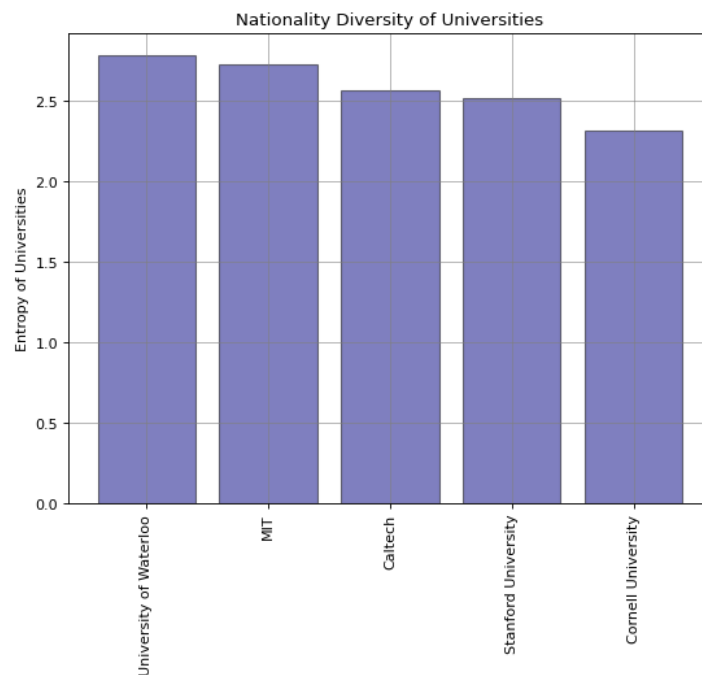
خط قرمز نشان‌دهنده‌ی نسبت تعداد استاد خانم به آقا در کل دادگان است که بالاتر گفتیم ۰.۲۰۸ است.

۴-۱-۵- تنوع در هر زیرشاخه و دانشگاه

همانطور که پیش‌تر اشاره شد، تنوع را با آنتروپی مدل کردیم. مقدار این آنتروپی بعنوان شاخصی برای تنوع ملیت در دانشگاه‌ها و زیرشاخه‌های مختلف در زیر نمایش داده شده است:



شکل ۱۲: نمودار میله‌ای تنوع ملیتی در هر زیرشاخه از علوم کامپیوتر



شکل ۱۳: نمودار میله‌ای تنوع ملیتی در هر دانشگاه

توجه داشته باشید که تابع آنترופی یک تابع لگاریتمی است و اختلاف مقادیر آنترופی کوچک است ولی می‌تواند نشان‌دهنده تنوع نه چندان کوچکی باشد.

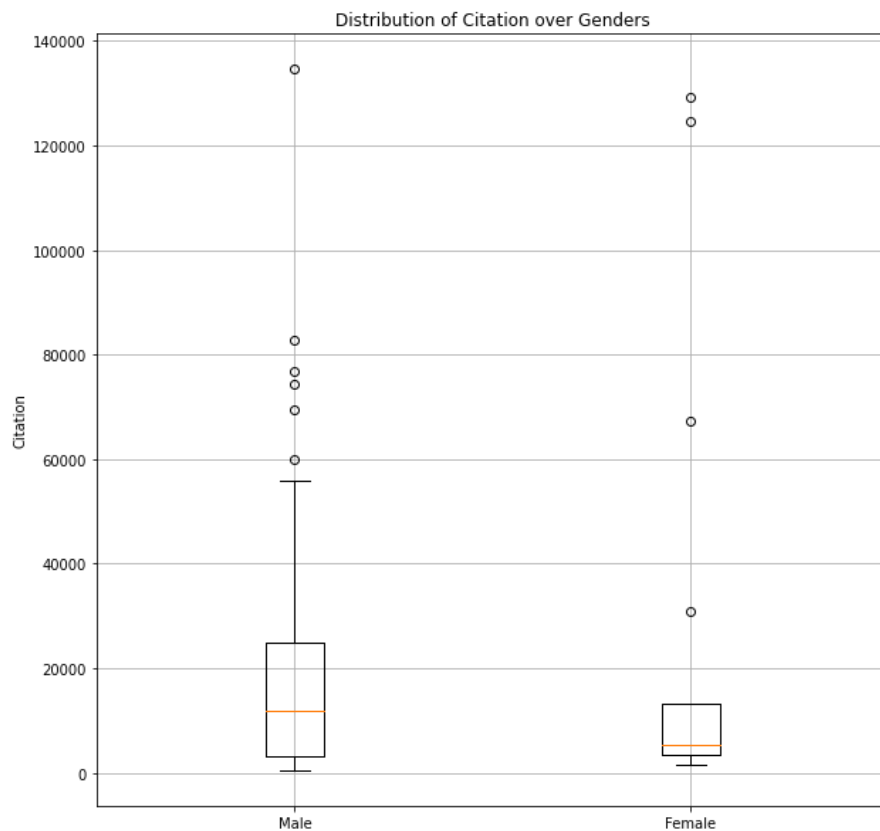
۵-۲- مقایسه‌ی عملکرد اقلیت‌ها در برابر اکثریت‌ها

۵-۲-۱- جنسیت

تعداد اساتید خانم در دادگان ۱۷ و تعداد اساتید آقا ۸۰ تاست. آماره‌های مربوط به citation:

بیشینه	کمینه	میانه	میانگین	
۱۲۹۲۷۹	۱۵۹۱	۵۲۳۹	۲۴۴۸۳	خانم
۱۳۴۶۷۴	۳۹۶	۱۱۹۰۷	۱۹۰۵۹	آقا

جدول ۱۱: مقایسه‌ی آماره‌های مربوط به citation به تفکیک جنسیت



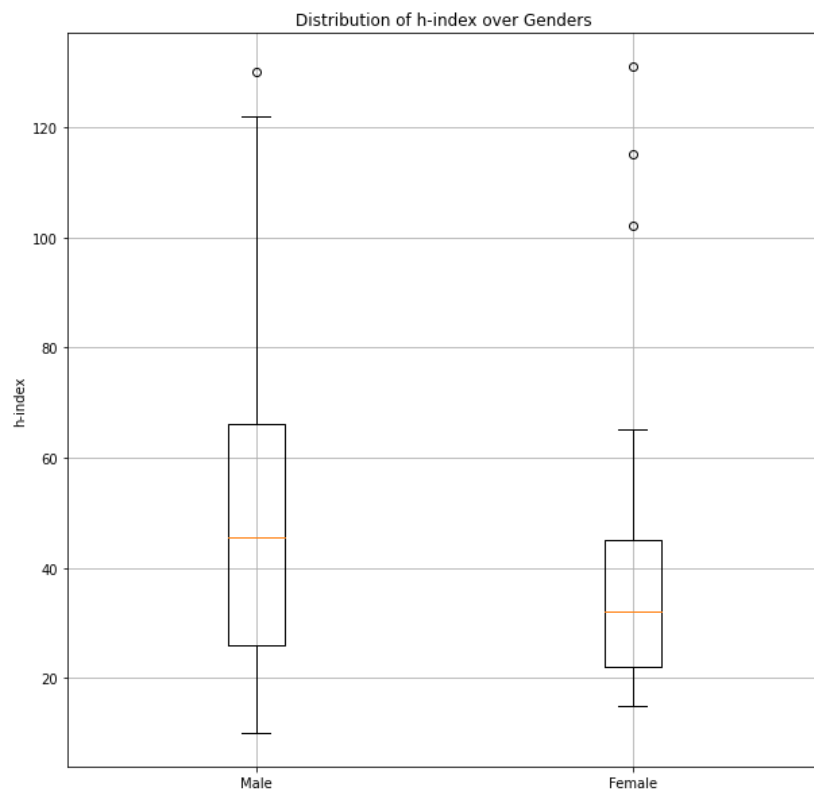
شکل ۱۴: نمودار جعبه‌ای citation به تفکیک جنسیت

نمودار جعبه‌ای و جدول فوق نشان می‌دهد که میانگین citation خانم‌ها بیشتر است ولی میانه‌ی citation آقایون بیشتر است. همچنین این شاخص در آقایان رنج پخش تری دارد ولی در خانم‌ها متمرکز تر است. (مقدار IQR در آقایان بزرگتر از خانم‌هاست)

آماره‌های مربوط h-index:

	میانگین	میانه	کمینه	بیشینه
خانم	۴۵.۹۴	۳۲	۱۵	۱۳۱
آقا	۴۹.۰۶	۴۵.۵	۱۰	۱۳۰

جدول ۱۲: مقایسه‌ی آماره‌های مربوط به h-index به تفکیک جنسیت

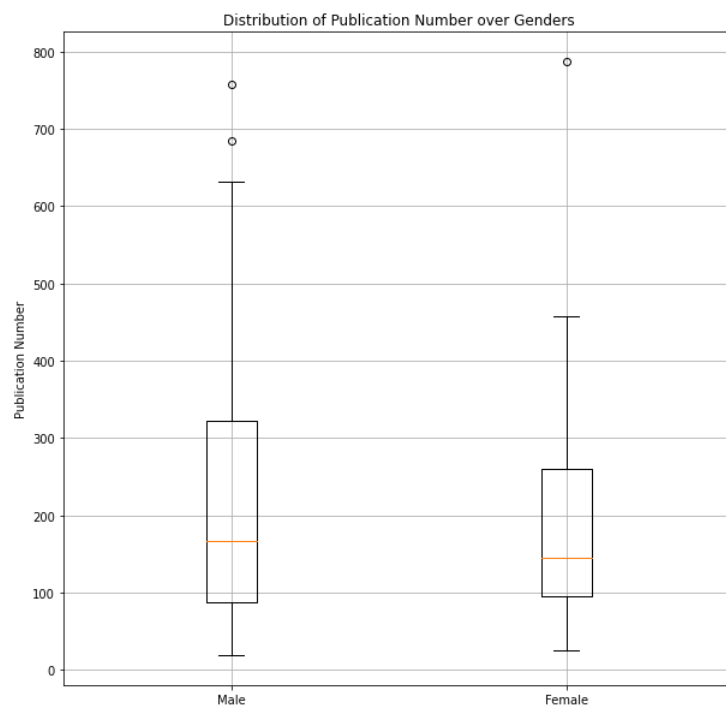


شکل ۱۵: نمودار جعبه‌ای h-index به تفکیک جنسیت

آماره‌های مربوط به تعداد مقاله:

	میانگین	میانه	کمینه	بیشینه
خانم	۲۱۸.۹۴	۱۴۵	۲۵	۷۸۷
آقا	۲۲۳.۰۹	۱۶۷	۱۹	۷۵۷

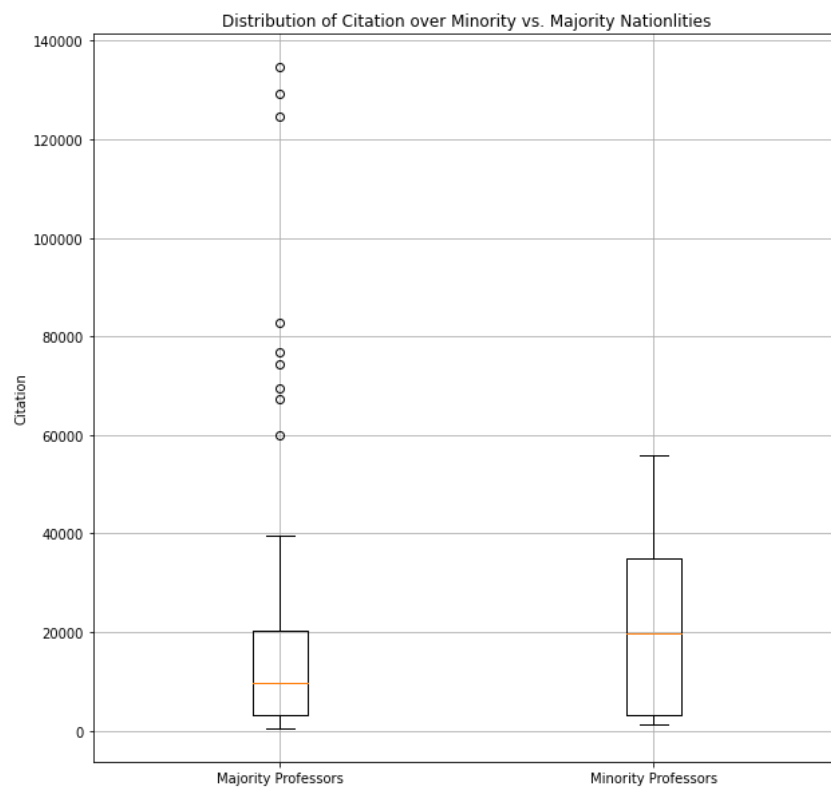
جدول ۱۲: مقایسه‌ی آماره‌های مربوط به h-index به تفکیک جنسیت



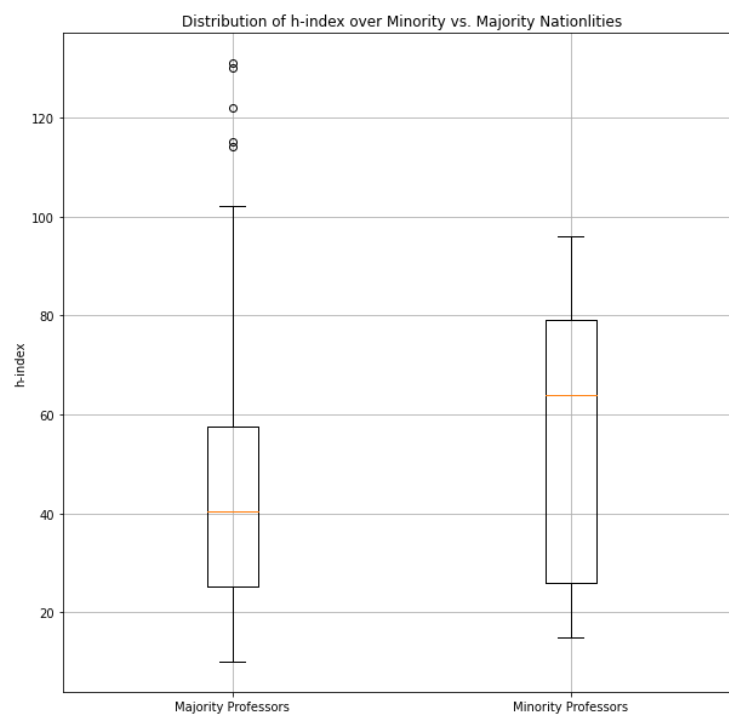
شکل ۱۶: نمودار جعبه‌ای تعداد مقاله به تفکیک جنسیت

۲-۲-۵- ملیت

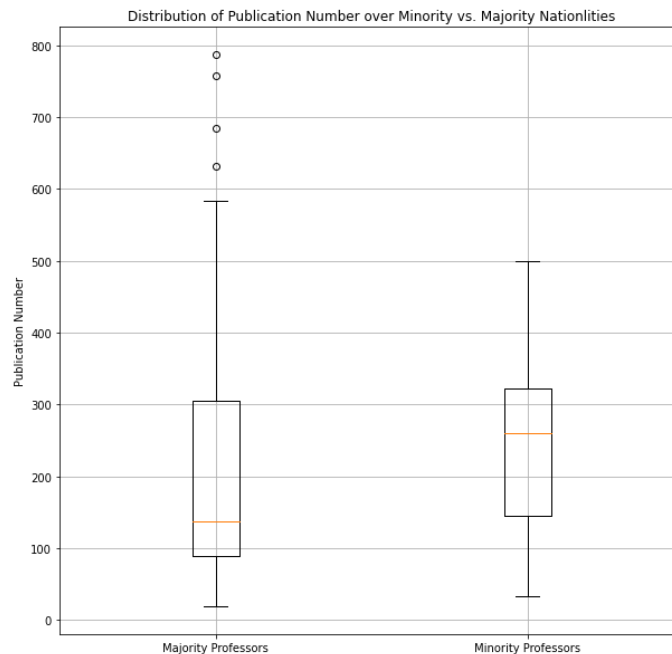
ملیت اساتید را بر حسب تکرار در دادگان به دو دسته‌ی اقلیت و اکثریت تقسیم کردیم. حال به بررسی نمودار جعبه‌ای سه شاخص عملکرد برای این دو دسته می‌پردازیم.



شکل ۱۶: نمودار جعبه‌ای citation به تفکیک اقلیت و اکثریت ملیتی



شکل ۱۷: نمودار جعبه‌ای h-index به تفکیک اقلیت و اکثریت ملیتی



شکل ۱۸: نمودار جعبه‌ای تعداد مقاله به تفکیک اقلیت و اکثریت ملیتی

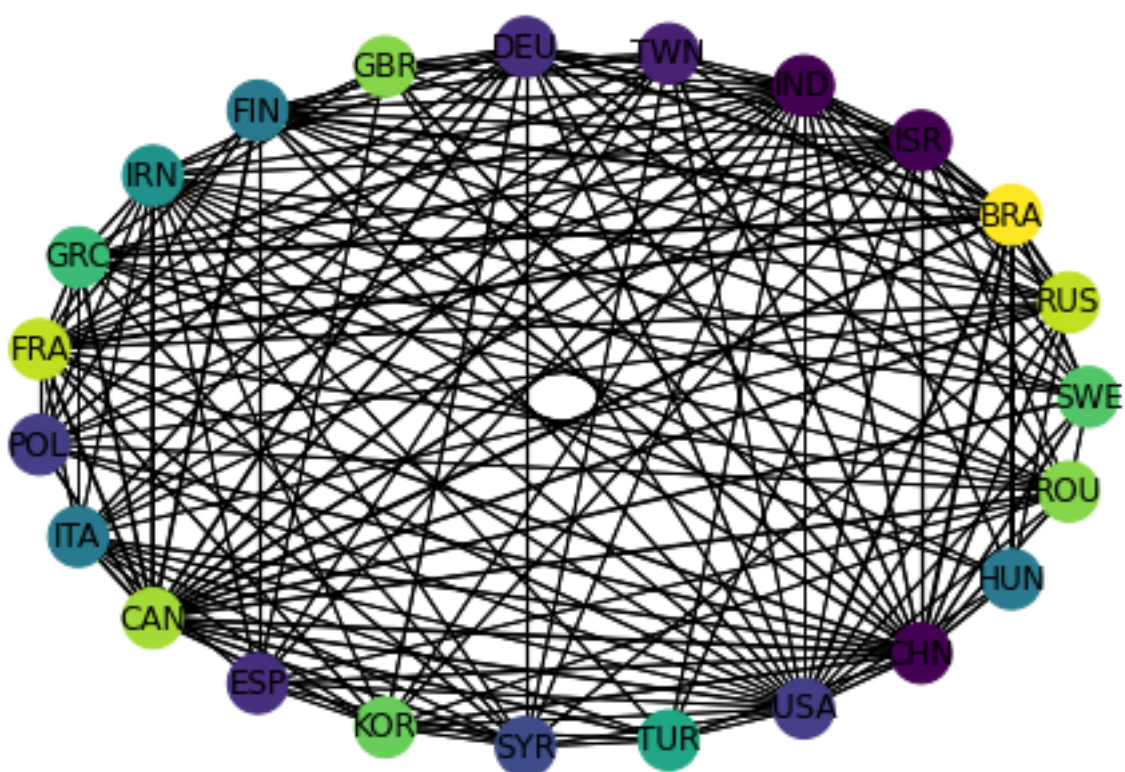
میانه اقلیت‌ها در هر سه‌ی این نمودارها بیشتر از اکثریت‌هاست.

۳-۵- مصورسازی گرافی داده

برای درک بهتر از داده، تعدادی مصورسازی بر روی گراف انجام دادیم تا روابط بین مولفه‌های موجود در داده بهتر درک شود.

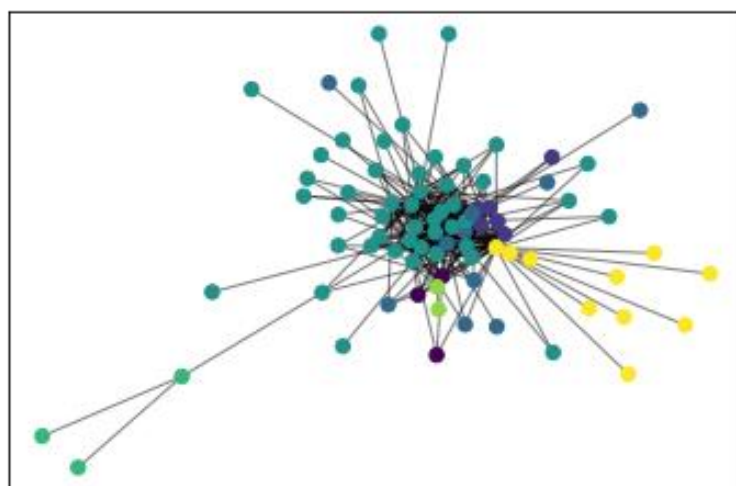
۱-۳-۵- ارتباط کشورها در روابط استاد-دانشجو

ابتدا گراف ملیت استاد دانشجو را تشکیل دادیم. رئوس این گراف کشورهاست. این گراف را بدون جهت و وزندار در نظر گرفتیم. به ازای هر زوج استاد دانشجو با ملیت‌های $n1$ و $n2$ ، بین دو راس مربوطه حال اضافه کردیم. سبک‌شده‌ی گراف را در زیر می‌بینید (رئوس با درجه‌ی کمتر از ۹ را حذف کردیم تا گراف شلوغ نشود)



شکل ۱۹: گراف ارتباطات ملیتی استاد دانشجو

بر روی گراف الگوریتم تشخیص انجمن louvin را اجرا می‌کنیم تا بررسی کنیم آیا در کمیته‌های گراف ارتباط ملیتی معناداری وجود دارد یا خیر.



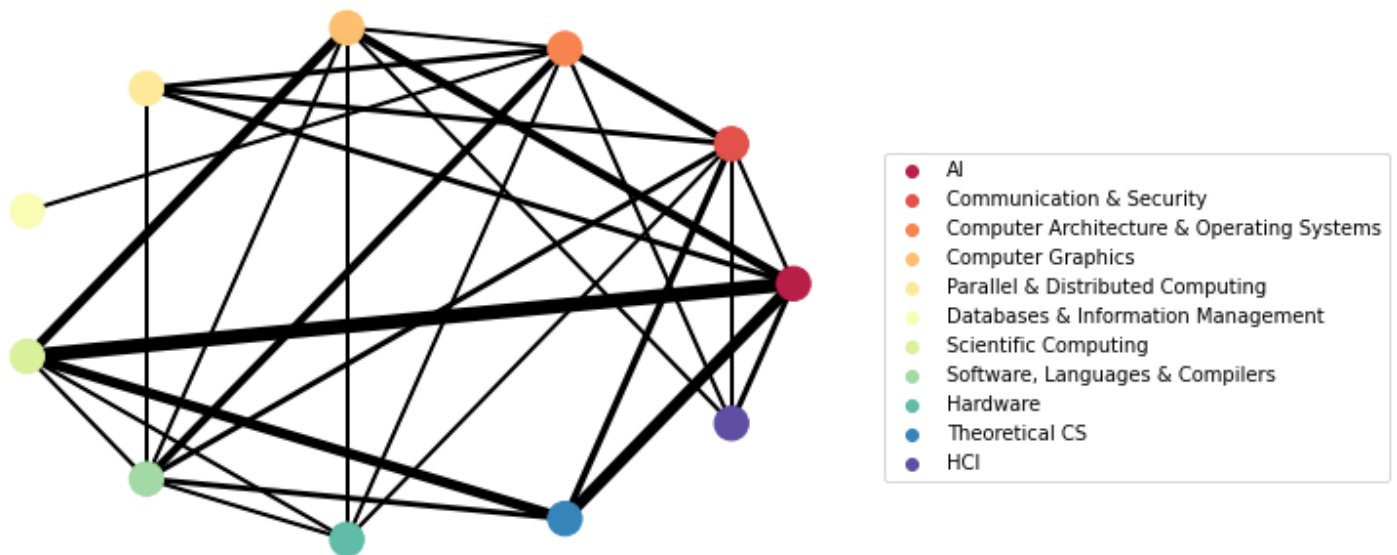
شکل ۲۰: گراف انجمن‌های تشخیص داده شده در گراف شکل ۱۹

-
- ['Egypt', 'Malaysia', 'Spain']
 - ['Sweden', 'United Kingdom', 'Ecuador', 'Iran', 'South Korea', 'Australia']
 - ['Nigeria', 'Argentina', 'Ethiopia', 'Italy', 'Vietnam', 'Iraq', 'Colombia', 'China', 'Tunisia', 'Cameroon']
 - ['Latvia', 'Russia', 'Slovakia', 'Bulgaria', 'Bangladesh', 'Brazil', 'Thailand', 'Israel', 'Switzerland', 'India', 'Portugal', 'Taiwan', 'Germany', 'Philippines', 'Serbia', 'Iceland', 'Finland', 'Czech Republic', 'Pakistan', 'Chile', 'Slovenia', 'Greece', 'Nepal', 'France', 'Poland', 'Lebanon', 'New Zealand', 'Singapore', 'Netherlands', 'Mexico', 'Canada', 'Belgium', 'Croatia', 'Sri Lanka', 'Syria', 'Ireland', 'Ukraine', 'Peru', 'Guyana', 'Mozambique', 'Turkey', 'Saudi Arabia', 'Morocco', 'Japan', 'Romania', 'Venezuela', 'Indonesia']
 - ['Yemen', 'Palestine', 'Jordan']
 - ['Armenia', 'Hungary']
 - ['Algeria', 'Austria', 'Uruguay', 'Uganda', 'Dominican Republic', 'Ghana', 'Georgia', 'United States', 'Denmark', 'Malta']

الگوهای از نزدیکی جغرافیایی در بین کشورهای هر انجمن قابل مشاهده است.

۲-۳-۵- ارتباط زیرشاخه‌های علوم کامپیوتر

ابتدا گراف دو بخشی استاد زیرشاخه را تشکیل دادیم به این صورت که بخش اول شامل رئوس مربوط به اساتید است و بخش دوم شامل رئوس زیرشاخه‌های اصلی علوم کامپیوتر که تعیین کردیم. بین راس استاد و زیرشاخه یال اضافه می‌کنیم اگر استاد در آن زیرشاخه فعالیت علمی داشته باشد. یک استاد می‌تواند به بیش از یک زیرشاخه وصل باشد. سپس گراف project شده‌ی این گراف دوبخشی را بر روی رئوس زیرشاخه محاسبه می‌کنیم. در این گراف وزندار بی‌جهت، دو زیرشاخه به هم متصل می‌شوند اگر استادی باشد که به صورت بین‌رشته‌ای در هر دو زیرشاخه فعالیت داشته باشد. این گراف نشان‌دهنده‌ی این است که زیرشاخه‌ها چقدر نزدیکی علمی دارند و از دانش هر کدام در دیگری استفاده می‌شود.



شکل ۲۱: گراف ارتباطات زیرشاخه‌های علوم کامپیوتر

فصل ۶

فصل ۶: جمع‌بندی

در این پروژه به بررسی و تحلیل داده‌ی استاد دانشجو در دانشگاه‌های برتر آمریکای شمالی و در مقطع تحصیلات تکمیلی پرداختیم. سوالات زیر را در داده بررسی کردیم:

- آیا جانبداری ملیتی و جنسیتی در داده وجود دارد؟ به این معنی که استادها تمایل داشته باشند دانشجویان هم ملیتی و هم جنسیتی با خود را بپذیرند؟ با استفاده از شبیه‌سازی تصادفی و آزمون فرض و محاسبه‌ی $pvalue$ دریافتیم که این جانبداری ملیتی در داده وجود دارد و داده شواهد کافی برای وجود این جانبداری جنسیتی نشان نمی‌دهد.

- تاثیر این جانبداری بر عملکرد استادها چیست؟ نشان دادیم که جانبداری ملیتی با شاخصه‌های عملکرد همبستگی منفی معناداری دارد ولی برای جانبداری جنسیتی شواهد کافی برای معنادار بودن همبستگی منفی دیده شده وجود نداشت.

- تاثیر تنوع بر شاخصه‌های عملکرد چیست؟ نشان دادیم این دو پارامتر ارتباط و همبستگی مثبت بسیار زیادی با یکدیگر دارند. (با $pvalue$ بسیار کوچک)

تعدادی سؤالات جانبی نیز در داده بررسی شد و مصورسازی‌های دیگری نیز از داده برای درک بهتر آن پیاده‌سازی کردیم.

سؤالاتی نظیر:

- اساتیدی که شاخه کاری‌شان بین رشته‌ای است موفق ترند یا کسانی که روی تک‌شاخه تمرکز دارند؟ نشان دادیم که همبستگی مثبت معناداری بین بین‌رشته‌ای بودن و موفقیت وجود دارد.
- نسبت دانشجویان پسر به دختر در هر زیرشاخه از علوم کامپیوتر را بررسی کردیم و این نسبت را با نسبت کلی پسر به دختر در کل دادگان مقایسه کردیم.

مصورسازی‌هایی نظیر:

- مصورسازی گرافی داده برای نمایش میزان تکرار رابطه‌ی کشورها در نقش استاد دانشجو و شناسایی انجمن^۱ در این گراف برای کشف ارتباطات معنادار جغرافیایی
- مصورسازی گرافی داده برای درک بهتر از میزان ارتباط زیرشاخه‌های علوم کامپیوتر
- مصورسازی تعداد استاد و شاخصه‌های آماری موفقیت در هر زیرشاخه از علوم کامپیوتر
- مصورسازی شاخصه‌های موفقیت در گروه‌های اقلیت در برابر اکثریت جنسیتی و ملیتی
- مقایسه‌ی تنوع ملیتی زیرشاخه‌های مختلف علوم کامپیوتر

¹ community detection

البته بایست توجه کرد که در تمامی بخش‌هایی که همبستگی محاسبه می‌کنیم، با تست آماری نشان می‌دهیم ارتباط قوی‌ای بین دو متغیر وجود دارد که با صفر اختلاف معناداری دارد. منتها نمی‌توان از آن رابطه‌ی علی نتیجه گرفت.

برخی از مواردی بررسی شده نیز نتیجه‌گیری در ارتباط با آن انجام نشده و تنها به صورت کاوشی داده مصورسازی و بررسی شده است که می‌تواند مسیری برای پژوهش‌های آینده باشد.

تلاش این پروژه بر این بود که علاوه بر بررسی سوالات اساسی مطرح‌شده در پروپوزال، به بررسی کاوشی داده برای درک بهتر از آن بپردازد. نتایج کار این تحقیق قابل استفاده و استناد برای عادلانه‌تر کردن سیستم پذیرش دانشجویی و شمول بیشتر گروه‌های اقلیت در موقعیت‌های تحصیلی و شغلی است. همچنین می‌تواند منجر به افزایش کارایی فعالیت‌های علمی شود.

فصل ٧

فصل ٧: مراجع

- [1] Derrida, Jacques (1998). *Of Grammatology*. The Johns Hopkins University Press. pp. 11–12
- [2] Caprice D. Hollins and Ilsa Govan, Diversity, Equity, and Inclusion: Strategies for Facilitating Conversations on Race.
- [3] Bickel, P. J., Hammel, E. A. & O’Connell, J. W. (1975) Sex bias in graduate admissions: data from Berkeley, *Science*, 187(4175), 398–404.
- [4] Attiyeh, G., & Attiyeh, R. (1997). Testing for Bias in Graduate School Admissions. *The Journal of Human Resources*, 32(3), 524-548. doi:10.2307/146182
- [5] White, S. W., Xia, M., & Edwards, G. (2020). Race, gender, and scholarly impact: Disparities for women and faculty of color in clinical psychology. *Journal of Clinical Psychology*. <https://doi.org/10.1002/jclp.23029>
- [6] Pruitt, A., & Isaac, P. (1985). Discrimination in Recruitment, Admission, and Retention of Minority Graduate Students. *The Journal of Negro Education*, 54(4), 526-536. doi:10.2307/2294713
- [7] Pruitt, J. (2020). The Roles of Racial Bias in Graduate Admission Decisions. *Long Island University, The Brooklyn Center. ProQuest Dissertations Publishing*.
- [8] Woo, S. E., LeBreton, J., Keith, M., & Tay, L. (2020, August 18). Bias, Fairness, and Validity in Graduate Admissions: A Psychometric Perspective. <https://doi.org/10.31234/osf.io/w5d7r>
- [9] Lauren A. (2018, June 7). Admitting Bias in Doctoral Programs, <https://doi.org/10.1177/1536504218776965>
- [10] Muric, Goran, Lerman, Kristina, Ferrara, Emilio. 2020. “COVID-19 Amplifies Gender Disparities in Research.” arXiv. Retrieved March 24, 2021. <http://arxiv.org/abs/2006.06142>.
- [11] M. S. Granovetter, The strength of weak ties. *Am. J. Sociol.* 78, 1360–1380 (1973). 3.
- [12] R. S. Burt, Structural holes and good ideas. *Am. J. Sociol.* 110, 349–399 (2004). 4.
- [13] M. W. Nielsen et al., Opinion: Gender diversity leads to better science. *Proc. Natl. Acad. Sci. U.S.A.* 114, 1740–1742 (2017). 5.

-
- [14] S. E. Page, *The Diversity Bonus: How Great Teams Payoff in the Knowledge Economy* (Princeton University Press, Princeton, NJ, 2009). 6.
- [15] S. T. Bell, A. J. Villado, M. A. Lukasik, L. Belau, A. L. Briggs, Getting specific about demographic diversity variable and team performance relationships: A meta-analysis. *J. Manage.* 37, 709–743 (2011).
- [16] Hofstra, B. et al. The diversity–innovation paradox in science. *Proc. Natl Acad. Sci. USA* 117, 9284–9291 (2020)
- [17] Zwillinger, D. and Kokoska, S. (2000). *CRC Standard Probability and Statistics Tables and Formulae*. Chapman & Hall: New York. 2000. Section 14.7

Abstract:

It is needless to mention that graduate study plays a paramount role in the academic and career opportunities of people. Besides, the scientific level of university and advisor are among contributing factors in the quality of the graduate study. Therefore, bias and discrimination in academia are among the problems that have been investigated carefully. Furthermore, diversity and inclusion have become hot topics in university admissions and job recruitments. In the following research, we aim to answer these questions:

- Are there any patterns of bias and discrimination in terms of nationality and gender in graduate admission in top universities? e.g Are professors more inclined toward accepting students of the same gender and nationality of their own?
- If so, how does this bias affect the scientific performance of the research group and the professor?
- How do diversity and inclusion affect performance?

The results of our computational research have shown that there exists a bias in terms of nationality. However, our data suggests inadequate evidence for proving gender bias. We also showed that this bias has a significant negative correlation with performance metrics. We have also shown that nationality diversity is strongly correlated with performance metrics, but we could not conclude the same result for gender diversity. At last, we explored some interesting questions in our dataset that can be a path for future researches. The results of this work can be employed to make the educational systems fairer and include more minorities in educational and job opportunities. It can also lead to an increase in performance in scientific activities.

Keywords: bias and discrimination- diversity and inclusion- advisor advisee- graduate study- nationality- gender- statistical testing- sociology of science



University of Tehran



College of Engineering

School of Electrical and Computer Engineering

Investigating the patterns of discrimination and diversity & inclusion in graduate study at top universities of North America

A thesis submitted to the Undergraduate Studies Office

In partial fulfillment of the requirements for

The degree of bachelor in

Computer Engineering

By:

Seyedeh Baharan Khatami

Supervisor:

Dr. Behnam Bahrak