UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

IMPACTO DE LAS MUJERES EN EQUIPOS DE TRABAJO DE DESARROLLO DE SOFTWARE

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS, MENCIÓN COMPUTACIÓN

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL EN COMPUTACIÓN

JAVIERA FERNANDA ROMERO DROGUETT

PROFESORA GUÍA:
MARÍA CECILIA BASTARRICA PIÑEYRO

MIEMBROS DE LA COMISIÓN:
JOCELYN SIMMONDS WAGEMANN
VALENTÍN MUÑOZ APABLAZA
CARLA VAIRETTI MARTIN

SANTIAGO DE CHILE
2024

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN CIENCIAS DE LA COMPUTACIÓN
Y MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL EN COMPUTACIÓN
POR: JAVIERA FERNANDA ROMERO DROGUETT
FECHA: 2024
PROF. GUÍA: MARÍA CECILIA BASTARRICA PIÑEYRO

IMPACTO DE LAS MUJERES EN EQUIPOS DE TRABAJO DE DESARROLLO DE SOFTWARE

La participación femenina en el ámbito STEM (Science, Technology, Engineering and Mathematics) ha exhibido históricamente índices significativamente inferiores en comparación al género masculino. Cuando se examina específicamente la participación de las mujeres en el campo de la computación, nos encontramos con una representación aún más reducida. Esta situación podría deberse a que existen estereotipos respecto al bajo rendimiento de las mujeres en estas áreas y a no ser un aporte en los grupos de trabajo.

En el área de ingeniería de software, el trabajo en equipo desempeña un papel fundamental en el desarrollo de proyectos. Sin embargo, no es fácil evaluar la calidad del trabajo en equipo ni determinar la contribución de cada uno de los miembros. Una de las formas de abordar este problema es analizando cómo se percibe el rendimiento de los miembros del equipo a través de coevaluaciones y autoevaluaciones. Estas evaluaciones miden el desempeño percibido tanto individualmente como el de los compañeros de grupo.

El objetivo de esta tesis es analizar y medir el impacto de la presencia de mujeres en equipos de desarrollo de software, particularmente si esta presencia modifica las percepciones sobre su rendimiento. Para llevar a cabo este análisis, se cuenta con una plataforma de coevaluaciones que ha sido utilizada dentro del departamento desde hace muchos años y se implementó una extensión para realizar también autoevaluaciones. Esta nueva herramienta se empleó en dos cursos obligatorios del área de ingeniería de software del Departamento de Ciencias de la Computación de la Universidad de Chile con el fin de recopilar datos relevantes.

Se definieron métricas para evaluar el trabajo en equipo, como la composición de género y promedios de coevaluaciones y autoevaluaciones. Se compararon estas métricas entre géneros y equipos con/sin mujeres para detectar diferencias significativas mediante pruebas estadísticas.

Los resultados obtenidos mostraron que no existe una diferencia estadísticamente significativa en la calidad del trabajo en equipo dependiendo de la presencia de las mujeres de acuerdo a la percepción de los alumnos respecto al trabajo de sus compañeros y de ellos mismos. En particular, este resultado encapsula el hecho de que tener mujeres en un equipo de desarrollo de software, mantiene y no empeora las dinámicas de trabajo. Se plantea además que la presencia de más de una mujer en los equipos mejora el ambiente de trabajo y la comodidad de los integrantes del equipo.

Este trabajo intenta ser un aporte para derribar estereotipos negativos acerca del impacto de la participación de mujeres en equipos de desarrollo de software.

Agradecimientos

Quiero agradecer a todas las personas que me han acompañado de distintas formas en estos años.

En primer lugar a mis papás por todo lo que me han dado. Todo lo que soy es por y para ustedes. Gracias por siempre incentivarme a hacer lo que me gusta, por ser un ejemplo para mi y, sobre todo, por nunca presionarme.

A mi hermano Ignacio, por entenderme y soportarme en los momentos de estrés. Gracias por ayudarme siempre que pudiste, escucharme y por apañarme en todo.

A toda mi familia, en especial a mi prima Viviana, que desde el primer día me abrió las puertas de su casa, y a mi primo Matías, que estuvo para guiarme en este nuevo mundo.

A mis amigas y amigos del colegio, por tantos años de amistad y carretes que necesitaba para distraerme de la U, en especial a Javi, Feña, Isi, Isa y Cote.

Infinitas gracias a Werner, por acompañarme todos estos años. Nunca voy a terminar de agradecerte lo mucho que me has ayudado, sobre todo en esta última etapa. Por ser mi mayor apoyo, estudiar conmigo siempre que lo necesité, darme el ánimo que me faltaba para avanzar y siempre creer en mí.

Gracias a Cata, Pascal, Coloma y Coni, por estar conmigo desde el primer semestre, y a pesar de dejar de tener ramos juntos, nunca nos separamos.

A cada uno de los amigos que fui haciendo en los siguientes semestres. A mi grupo Pichigang, gracias por tantas tardes de estudio, almuerzos y viernes de carretes en la U, viajes a la playa y miles de momentos que hicieron que esta etapa sea inolvidable.

A Rai y Mati, mis amigos DCC. Gracias a ustedes logré sobrevivir a la universidad en pandemia, por las miles de horas que pasamos en Discord estudiando, haciendo tareas y motivándonos con los ramos que se veían imposibles. Gracias por ayudarme a entender un poco más esta carrera.

Y finalmente, a mi profesora guía, Cecilia Bastarrica. Gracias por orientarme a lo largo de este proceso, por estar siempre disponible para responder mis dudas y reunirse conmigo, y por sus palabras de aliento en los momentos difíciles.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Motivación

La participación de las mujeres en carreras universitarias y puestos de trabajo en el ámbito de las ciencias, tecnología, ingeniería y matemáticas (STEM, por sus siglas en inglés) se ha convertido en un tema de interés en los últimos años. Históricamente, la representación de mujeres en estos campos ha sido significativamente menor en comparación con la de los hombres [1]. Esta disparidad es aún más evidente en áreas como la computación y el desarrollo de software.

En el año 2015, la representación de las mujeres en trabajos relacionados con STEM en los Estados Unidos apenas alcanzó el 24 % [2]. En Chile, en el año 2021, las mujeres representaron únicamente el 20 % de los estudiantes inscritos en carreras vinculadas a la tecnología, informática, mecánica, electrónica y construcción [3]. Esta disparidad se refleja también en la Facultad de Ciencias Físicas y Matemáticas (FCFM) de la Universidad de Chile, a pesar de sus esfuerzos por impulsar la admisión de mujeres a través del Programa de Ingreso Prioritario de Equidad de Género (PEG) [4]. Esta brecha persiste, como se puede apreciar en el gráfico de la figura 1.1. Estos datos subrayan la urgente necesidad de promover una mayor inclusión de las mujeres en estos campos.

La línea negra de la figura 1.1 representa el porcentaje de mujeres matriculadas en la FCFM desde 2006 hasta 2020, donde ha habido un aumento progresivo en el tiempo, llegando a un máximo de 32.89 % el 2018. Por otro lado, al observar la línea rosada, se expone el porcentaje de mujeres del total de estudiantes que ingresaron al Departamento de Ciencias de la Computación (DCC). Se infiere una gran disparidad en el porcentaje de hombres y mujeres ingresados en el DCC obteniendo un máximo de 23.76 % de mujeres el año 2020 lo que implica que aún dentro de la propia facultad, pocas mujeres eligen computación.

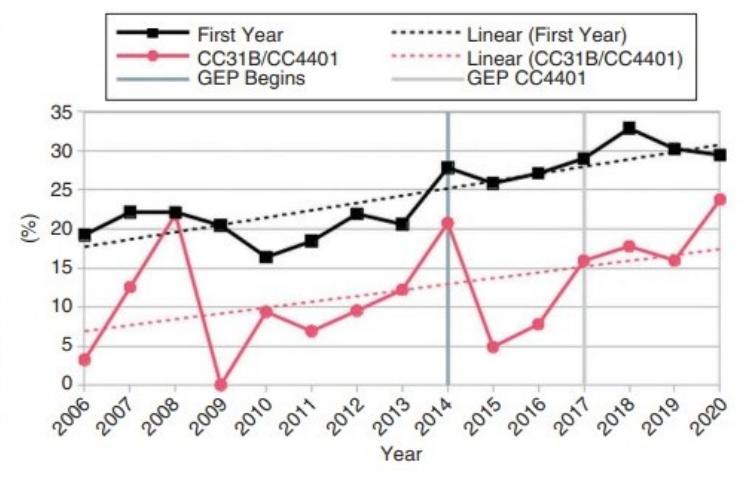


Figura 1.1: Porcentaje de mujeres matriculadas a nivel facultad (línea negra) y en el DCC (línea rosada) [5]

Además, a nivel nacional, el porcentaje de mujeres matriculadas en carreras de Ingeniería Civil en Computación e Informática en la universidades chilenas el 2023 fue de 11,4 %, lo que muestra una brecha aún mayor que a nivel de la facultad [6]. Este escenario de baja participación femenina en el desarrollo de software, tanto en el ámbito académico como laboral, ha llevado a la propagación de estereotipos. Se insinúa erróneamente que las mujeres podrían tener un rendimiento inferior en comparación con los hombres en estas áreas. Sin embargo, se ha planteado también la creencia de que tener mujeres en los equipos de trabajo no mejora el desempeño general del grupo, ya que se asume que las mujeres son más hábiles solamente en áreas menos técnicas [7] [8].

Lo anterior presenta un desafío, ya que es difícil evaluar y medir el trabajo en equipo en cualquier área y poder cuantificar el impacto que tiene la presencia femenina. Una posible metodología para evaluar el desempeño de un equipo de desarrollo de software en un entorno universitario es a través de las calificaciones obtenidas en las evaluaciones del curso. Además de esto, para determinar si estas calificaciones reflejan el esfuerzo conjunto del equipo o si se deben a unos pocos estudiantes destacados, se puede medir el desempeño percibido entre los miembros del equipo mediante coevaluaciones y autoevaluaciones [9]. La coevaluación es un instrumento de evaluación en el cual cada miembro del equipo evalúa a sus compañeros, lo que promueve un aprendizaje integral y fortalece los valores personales y profesionales [10]. Por otro lado, la autoevaluación es un método mediante el cual cada miembro del equipo se evalúa a sí mismo, lo que fomenta la responsabilidad y el compromiso de los estudiantes en su propio proceso de aprendizaje [11].

La carrera de Ingeniería Civil en Computación de la Universidad de Chile se extiende a lo largo de 11 semestres e incluye tres cursos obligatorios en el área de ingeniería de software: Ingeniería de Software en el séptimo semestre, Ingeniería de Software II[1] en el noveno semestre y Proyecto de Software en el décimo semestre. Estos cursos se basan en una metodología práctica en la que los estudiantes trabajan en equipo para desarrollar proyectos, con el objetivo de prepararlos para el entorno laboral. Además, en estos tres cursos, cada estudiante debe completar coevaluaciones dentro de su grupo mediante una plataforma del

[1] En la nueva malla curricular este curso es electivo.

DCC. El formulario de las coevaluaciones es idéntico para los tres cursos y su nota influye en las calificaciones que obtienen los estudiantes.

En esta tesis, se propone realizar el análisis del impacto de las mujeres en los equipos de trabajo de los cursos "Ingeniería de Software" (CC4401) y "Proyecto de Software" (CC5402), considerando la composición de género de dichos equipos y utilizando las coevaluaciones y autoevaluaciones para medir su desempeño de acuerdo con las percepciones de los alumnos respecto a sus compañeros de equipo y de ellos mismos. Para lograr lo anterior los estudiantes deben responder las coevaluaciones publicadas en la plataforma del DCC. Esta plataforma está implementada sólo con las coevaluaciones por lo que parte del trabajo de esta tesis es añadir las autoevaluaciones también a la plataforma.

Dado el problema propuesto se plantean las siguientes preguntas de investigación:

RQ1 ¿Cómo se autoevalúan las mujeres respecto a los hombres?

RQ2 ¿En qué se diferencian los equipos de trabajo con mujeres y sin mujeres en relación a las autoevaluaciones y coevaluaciones?

RQ3 ¿Influye la cantidad de mujeres que hay en un equipo de trabajo en la manera que se coevalúan y autoevalúan los estudiantes?

1.2. Objetivos

1.2.1. Objetivo general

El objetivo general de esta tesis es analizar y cuantificar el impacto de la presencia de mujeres en los equipos de trabajo de desarrollo de software. En particular, se busca determinar si esta presencia cambia las percepciones de su desempeño, específicamente en los cursos CC4401 y CC5402.

1.2.2. Objetivos específicos

Para entender los objetivos específicos, se debe contextualizar brevemente que existe una herramienta de coevaluaciones para los cursos CC4401 y CC5402, la cual consiste en una plataforma web en la que los estudiantes responden una coevaluación de sus compañeros y compañeras de trabajo en relación a su desempeño en el equipo a lo largo del semestre.

Los objetivos específicos son:

1. **Implementar una extensión de autoevaluaciones en la herramienta de coevaluaciones**: consiste en las mismas preguntas de coevaluación con la diferencia que los estudiantes evalúan su propio desempeño.
2. **Definir métricas para evaluar el impacto de las mujeres en los equipos de trabajo**: estas métricas son estadísticas específicas calculadas a partir de las respuestas de autoevaluación y coevaluación.

3. **Analizar y encontrar patrones en los datos y las métricas definidas para responder las preguntas de investigación**.

4. **Realizar un análisis de sentimiento de las preguntas abiertas de fortalezas y debilidades:** el análisis de sentimiento se basa en procesar el texto de las preguntas abiertas para calificar el sentimiento del comentario como positivo o negativo a través de un puntaje.

1.3. Estructura del documento

Este informe contempla siete capítulos principales. El primero corresponde al presente y describe el contexto y el problema a abordar, así como también una idea general de la solución propuesta. En el segundo capítulo, "Trabajo Relacionado", se exploran estudios de los temas relacionados a la investigación como las mujeres en carreras y trabajos del área STEM, cómo trabajan las mujeres en equipo y las evaluaciones formativas como la coevaluación y autoevaluación.

El tercer capítulo, "Marco Teórico", establece las bases teóricas para entender los métodos de análisis utilizados en este proyecto. Posteriormente, el cuarto capítulo presenta el "Caso de estudio", detallando el contexto de este y explicando la metodología que se seguirá a lo largo informe.

El quinto capítulo, "Resultados y Análisis", muestra los resultados obtenidos en el estudio. El sexto capítulo corresponde la "Discusión" de la investigación realizada, finalizando con el séptimo y último capítulo, que resume los hallazgos y las conclusiones de la tesis basadas en el análisis realizado.

Capítulo 2

Trabajo Relacionado

En este capítulo se presentan estudios relacionados con la participación de mujeres en carreras vinculadas al ámbito STEM, específicamente en computación, así como en la industria. Además, se aborda el tema de la evaluación del trabajo en equipo mediante coevaluaciones y autoevaluaciones.

2.1. Mujeres en carreras STEM

La representación de mujeres en los campos STEM, es decir, Ciencia, Tecnología, Ingeniería y Matemáticas, ha sido históricamente limitada, lo que ha dado lugar a una marcada brecha de género. A pesar de los avances en la igualdad de género en diversas áreas, en numerosos países, la proporción de mujeres que optan por seguir estudios superiores en campos STEM sigue siendo notablemente baja.

En el año 2017, de acuerdo con un estudio realizado por la UNESCO [12] solo alrededor del 30 % de las mujeres que estaban inscritas en programas de educación superior se dedicaban a disciplinas STEM a nivel mundial. En este contexto, las áreas de salud y bienestar tenían la mayor representación de mujeres (15 %), seguidas por ingeniería, manufactura y construcción (8 %), ciencias naturales, matemáticas y estadísticas (5 %), e información, comunicación y tecnología (3 %).

Centrándonos en el caso de Chile, en 2017, solamente el 23.11 % de las mujeres se matricularon en carreras STEM durante su primer año en la universidad [13]. Esta disparidad plantea interrogantes sobre los factores socioculturales que influyen en esta brecha, tales como prejuicios, discriminación, estereotipos, sexismo, supuestas incapacidades o estructuras científicas predominantemente masculinas [14]. La carencia de modelos femeninos a seguir en el ámbito STEM también ha contribuido a esta brecha persistente. Se ha sugerido que la presencia de más profesoras en las clases introductorias de carreras universitarias podría aumentar la confianza de las estudiantes al ver a estas profesoras como modelos a seguir [15].

A nivel institucional, algunos programas buscan mitigar esta brecha. Por ejemplo, el Programa de Equidad de Género (PEG) de la FCFM, que otorga cupos adicionales a mujeres en la lista de espera para ingresar a la universidad. A pesar de que este enfoque ha llevado a un incremento de mujeres en las carreras de la facultad, la mayoría de las alumnas beneficiadas no eligen la carrera de Ciencias de la Computación [5]. Esto evidencia la necesidad de

estrategias efectivas específicamente para este tipo de áreas con baja representación feminina.

En cuánto al rendimiento académico, en un estudio que se llevó a cabo en el DCC se hizo un análisis de los resultados académicos en cursos de computación [16]. A pesar de que en la mayor parte de los cursos no hay una diferencia significativa, los resultados revelaron que, en promedio, las mujeres superan a los hombres en algunos cursos obligatorios. Específicamente, destacaron en asignaturas como "Bases de Datos", "Ingeniería de Software II" y "Diseño y Análisis de Algoritmos", tanto en términos de mediana de calificaciones como en el porcentaje de aprobación. Además en un informe publicado por la FCFM en 2021 [17] se ha observado que las mujeres, en promedio, logran titularse en las carreras de ingeniería de esta facultad en un plazo menor que los hombres, con una diferencia de hasta un año. Ambos estudios no solo exponen el buen rendimiento académico de las mujeres en áreas técnicas específicas, sino que también desafían los estereotipos existentes sobre las capacidades de género en las disciplinas STEM, demostrando que las mujeres no solo igualan, sino que también superan al género masculino en rendimiento y eficiencia en sus estudios de ingeniería.

Existen estudios en otros países donde en general se observan patrones parecidos al caso de Chile y la FCFM. Por ejemplo en un estudio realizado por la Universidad de Costa Rica [18], se observó que las mujeres presentan menos tasas de abandono en la carrera y logran un rendimiento similar o incluso superior al de los hombres. Sin embargo, muestra que las mujeres enfrentan dificultades adicionales en los cursos introductorios de ingeniería en computación, ya que, en general, los hombres tienden a acercarse a esta área desde una edad más temprana. Otros estudios han encontrado que, en algunos casos, las mujeres no solo igualan, sino que incluso superan a los hombres en su desempeño en áreas de STEM [19].

Por último, en contraste a los estudios anteriores, otras investigaciones han demostrado que estas diferencias existen pero son mínimas y, al no tener diferencias significativas, no se ha hallado que el género sea un factor determinante en relación a las calificaciones [20] [21]. No obstante, resulta interesante destacar que en países con sociedades caracterizadas por una mayor desigualdad de género, como Malasia, ha habido un aumento en el número de mujeres que optan por la ingeniería de software. Esto podría indicar que la brecha de género en este campo puede deberse en gran medida a elecciones personales más que a discriminación o estereotipos [22]. Además, se reconoce que la disparidad de género en la inscripción en estas disciplinas académicas desempeña un papel fundamental en la explicación de la reducida presencia de mujeres en el mercado laboral [23].

2.2. Mercado laboral

En el año 2016, la Asociación Chilena de Empresas de Tecnologías de la Información (ACTI) [24] realizó un análisis que reveló que la participación de las mujeres en el mercado laboral de tecnologías de la información (TI) apenas alcanzaba el 5 %. Esta estadística pone de manifiesto una marcada brecha de género también en el sector laboral en Chile.

Por otro lado, un informe procedente del Reino Unido [25] arrojó datos preocupantes en el período comprendido entre 2019 y 2020: tan solo el 24 % de la fuerza laboral en el ámbito STEM estaba compuesto por mujeres. Esto subraya una vez más la persistencia de la des-

igualdad de género en estas disciplinas en un contexto internacional.

Un estudio llevado a cabo el 2018 en el sector de tecnología de la información en Alemania [7] se centró en investigar la brecha de género y sugirió que los gerentes tienden a preferir la contratación de mujeres para tareas que implican habilidades de orientación social, como la gestión de proyectos o la gestión de calidad. Los resultados revelaron que, en muchas ocasiones, se sigue identificando a las mujeres en función de sus habilidades blandas y un perfil menos técnico en comparación con los hombres en el mismo ámbito.

Estas cifras y hallazgos ponen de manifiesto la importancia de abordar la desigualdad de género en el campo de la tecnología y las disciplinas STEM a nivel global. Superar los estereotipos de género y promover una mayor inclusión de las mujeres en estas áreas es fundamental para aprovechar todo el potencial y la diversidad de talentos disponibles en la fuerza laboral y fomentar la igualdad de oportunidades en el mercado laboral tecnológico.

2.3. Trabajo en equipo

Un proyecto de ingeniería de software depende en gran medida del rendimiento del equipo [26]. Los proyectos en grupo ofrecen numerosos beneficios, ya que fomentan que los estudiantes aprendan a trabajar en equipo, haciendo hincapié en la cooperación, la colaboración y la negociación. También promueven un aprendizaje más efectivo, ya que los estudiantes se motivan mutuamente y dependen los unos de los otros, trabajando en horarios acordados, entre otros aspectos. Además, los proyectos abordados en equipo permiten a los estudiantes desarrollar sistemas más grandes y complejos de lo que sería posible de otra manera gracias a la división del trabajo. También les brindan la oportunidad de adquirir experiencia trabajando en equipo, una habilidad fundamental en la industria [27].

La inclusión de mujeres en equipos de trabajo en el ámbito STEM ha demostrado ser de un valor innegable en términos de creatividad, toma de decisiones y resolución de problemas. La diversidad de perspectivas y experiencias enriquece la calidad de los productos y servicios tecnológicos, lo que, a su vez, impulsa avances notables en la sociedad y en la economía [7]. Varios estudios han destacado que la proporción de mujeres en los equipos de trabajo está positivamente relacionada con el rendimiento de dichos equipos. Esta mejora en el rendimiento se logra a través de una mayor cohesión en el grupo, que se ve influenciada por la frecuencia de las interacciones entre líderes y miembros [28]. Además, la participación de mujeres en equipos de ingeniería de software ha demostrado conducir a una mayor coordinación y eficiencia en las tareas [29].

2.4. Coevaluaciones y autoevaluaciones

Las evaluaciones formativas, como la coevaluación y la autoevaluación, desempeñan un papel fundamental en el fortalecimiento y el estímulo de la responsabilidad y la autonomía de los estudiantes en su proceso de aprendizaje. Estos enfoques permiten a los estudiantes ver la evaluación como una oportunidad de aprendizaje [30].

La evaluación entre pares o coevaluación fomenta habilidades como el pensamiento crítico, la resolución de conflictos a través de diversas estrategias, la capacidad de debate y negociación, así como competencias relevantes en el entorno laboral y profesional. Esto conlleva la adquisición de herramientas para un aprendizaje continuo. Los estudiantes asumen una gran responsabilidad en su proceso de aprendizaje y participan activamente en una reflexión constante sobre su educación [31].

El fomento de la autoevaluación en el entorno universitario es un pilar esencial para promover el aprendizaje autónomo. Esta práctica permite a los estudiantes valorar su nivel de conocimientos y tomar medidas concretas para mejorarlo. Además, alineando sus percepciones con las de sus compañeros y profesores, la autoevaluación se convierte en una herramienta valiosa que tiene un carácter más formativo que sumativo. En otras palabras, se orienta hacia la conciencia de la situación de aprendizaje del estudiante y hacia la toma de decisiones para la mejora, en lugar de centrarse únicamente en calificar un resultado [32]. Esto también sirve para comparar la percepción que tiene cada alumno de sí mismo, es decir su autoconcepto. En las materias STEM, las mujeres suelen tener en principio un autoconcepto menor que los hombres, incluso si en realidad tienen las mismas notas y logros. Dado que las matemáticas son un filtro crucial para la inscripción y la permanencia en la educación STEM, un bajo autoconcepto matemático puede ser perjudicial y también puede reducir la motivación para el aprendizaje. En general, un autoconcepto demasiado crítico en matemáticas y STEM es un factor importante que influye en por qué las mujeres están menos motivadas en las materias STEM y por qué rara vez consideran una carrera en STEM [33].

Un estudio llevado a cabo el 2019 en la Universidad Técnica de Graz en Austria [34] examinó un curso introductorio de informática y se centró en la autoevaluación de los estudiantes antes y después del curso. El objetivo era abordar cuestiones relacionadas con las habilidades de programación en relación con las diferencias de género y si estas diferencias tenían algún impacto en el rendimiento académico. Los resultados del estudio revelaron que ni la educación previa ni el rendimiento de los estudiantes diferían significativamente en función del género. Sin embargo, se observó que las alumnas tendían a calificarse a sí mismas como menos capaces que los hombres en este ámbito.

También, un estudio realizado el 2021 en la Universidad Andrés Bello de Chile [35], en el que se analizaron 110 alumnos de las carreras de Ingeniería Civil en Computación e Ingeniería Civil Industrial, demostró que el uso de diferentes estrategias o técnicas en el aula, como dividir la clase en pequeños equipos, participar en evaluaciones por pares, realizar exposiciones y actividades de interés particular del estudiante, y trabajar en colaboración con estudiantes, en particular con mujeres, contribuye positivamente a aumentar el autoconcepto, al igual que mejora las evaluaciones formativas.

Capítulo 3

Marco Teórico

Este capítulo comienza estableciendo las bases teóricas pertinentes para los temas abordados en esta tesis. Se ofrece una exposición del marco teórico en dos áreas: el valor p (p-value) dentro del contexto de pruebas de hipótesis, como también el tipo de test usado en esta tesis: PERMANOVA (*Permutational Multivariate Analysis of Variance*) y la técnica de análisis de sentimiento.

3.1. Test de hipótesis y p-value

Basado en [36], la estadística se centra en la tarea de realizar inferencias sobre los parámetros desconocidos de una población en función de la información recopilada de una muestra de datos. En términos generales, estas inferencias se pueden interpretar de dos maneras: estimación de los parámetros de la distribución poblacional o como **test de hipótesis** de sus valores.

En una prueba estadística, como el test de hipótesis, el objetivo es evaluar una hipótesis sobre los valores de los parámetros de una población. Se formula una *hipótesis de investigación* o *hipótesis alternativa* H_1 sobre los parámetros que se busca respaldar y verificar. En un test de hipótesis, se respalda la hipótesis alternativa demostrando que su contraparte, conocida como *hipótesis nula* H_0, es falsa.

Para comprobar que la hipótesis alternativa es cierta, se debe demostrar que la hipótesis nula es falsa. Para lograr lo anterior es necesario calcular el **p-value**, el cual corresponde a la probabilidad de obtener el resultado observado al menos tan extremos como los que se obtienen en el estudio, bajo la suposición de que H_0 es cierta. Si se define T como la estadística de prueba calculada a partir de los datos y t es el valor observado de esta estadística, el p-value se expresa matemáticamente como $\mathbb{P}(T \geq t \mid H_0)$.

El objetivo del test es encontrar evidencia que respalde que la hipótesis nula es incorrecta. Como el p-value se calcula considerando H_0 como válida, cuanto menor sea la probablidad de que se cumpla la observación $T \geq t$ bajo esta suposición, mayor será la evidencia en contra de H_0. En otras palabras, un p-value más bajo indica una mayor evidencia en contra de la hipótesis nula.

Calcular un p-value que es pequeño puede significar una de dos cosas: que la hipótesis nula es falsa, o se encontró un resultado poco común que ocurre con baja probabilidad. Para evitar

esta ambigüedad, se define la **significancia del test** α al umbral para el p-value en cual se rechaza el test. De esta manera, se rechaza la hipótesis nula si p-value $< \alpha$. Generalmente se usa un valor del 1 % o del 5 %. Cuando se elige un nivel de significancia y se cumple que p-value $< \alpha$ se dice que hay suficiente evidencia para rechazar H_0 al $100 \cdot \alpha\,\%$ (para una mayor intuición ver Anexo A).

En un test de hipótesis existen dos tipos de errores: error de Tipo I y error de Tipo II. El error de Tipo I es en el que H_0 es rechazada a pesar de que es verdadera, que tendrá una probabilidad de ocurrir de α (mismo valor del nivel de significancia). Tiene sentido ya que el nivel de significancia α se define como el umbral para el p-value en cual se rechaza la hipótesis nula, o en otras palabras, la probabilidad máxima a la que se está dispuesto a rechazar la hipótesis nula cuando es verdadera. Por otro lado, el error Tipo II es en el que H_0 no se rechaza cuando en realidad es falsa. Se define que este error tiene una probabilidad β.

Cuando se quiere acotar la probabilidad de cometer error tipo I, se debe achicar el valor α, pero esto conlleva a aumentar la probabilidad β de cometer error tipo II por definición. Por esto en un test de hipótesis, dependiendo del contexto del experimento, se debe tomar en cuenta este trade-off entre α y β.

Por ejemplo, si se considera un test basado en un estadístico de prueba X que sigue una distribución normal con varianza σ^2, pero no se conoce su media μ. Se da el siguiente test de hipótesis:

- $H_1: \quad \mu = \mu_1$ (3.1)
- $H_0: \quad \mu = \mu_0 < \mu_1$ (3.2)

Suponiendo que H_0 se rechaza cuando X es mayor a un valor crítico x_c ($X > x_c$), se determina x_c definiendo un valor de nivel de significancia α. De esta manera, usando la definición de error de tipo I, la probabilidad de rechazar H_0 ($X > x_c$) cuando H_0 es verdadera, matemáticamente $\mathbb{P}(X > x_c \mid H_0) = \alpha$, se puede despejar el valor de x_c. Esta región donde se rechaza la hipótesis nula y se comete error tipo I queda ejemplicada en la zona azul en la figura 3.1.

Por otro lado, como H_0 se rechaza cuando $X > x_c$, este no se rechaza cuando $X \leq x_c$. Recordando que el error tipo II es cuando no se rechaza H_0 cuando es falsa, o en otras palabras, H_1 es verdadera, la probabilidad de error tipo II se define como $\mathbb{P}(X \leq x_c \mid H_1) = \beta$. Así, β queda definido por el valor de x_c encontrado anteriormente. La región donde no rechaza la hipótesis nula y se comete error tipo II queda ejemplicada en la zona naranja en la figura 3.1.

Con este ejemplo se puede explicar el trade-off entre α y β. Se puede notar que en este ejemplo, las distribuciones de X bajo H_0 y H_1 son fijas, entonces si se achica α, x_c se desplaza a la derecha, aumentando β, y de la misma forma ocurre el efecto inverso al agrandar α.

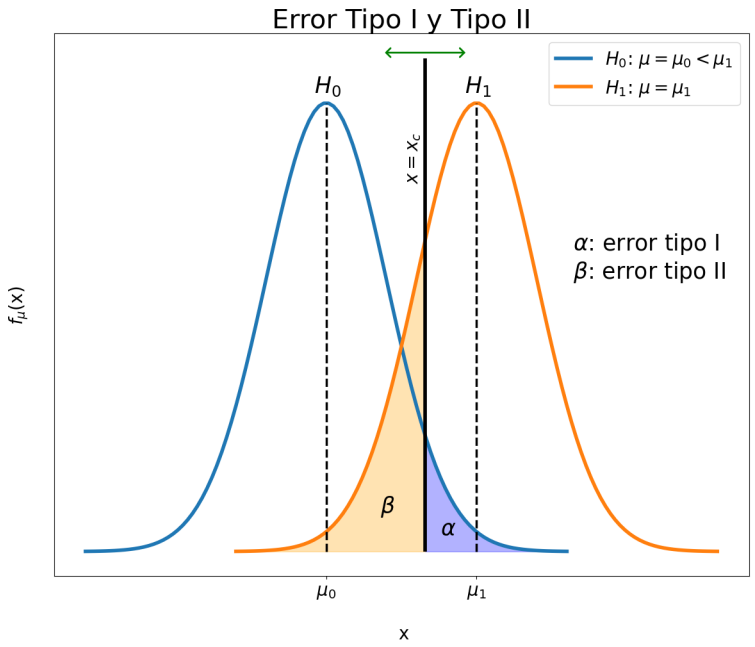


Figura 3.1: Ilustración que ejemplifica las regiones donde se comete error tipo I y tipo II

3.2. Correcciones de Bonferroni

Las correcciones de Bonferroni son un método de la estadística utilizado para ajustar el nivel de significancia α al hacer múltiples test de hipótesis. Su función es controlar la tasa de error de tipo I; existe una probabilidad α de cometer este tipo de error al hacer un test de hipótesis, pero al hacer más de uno, aumenta el riesgo de encontrar un resultado significativo por casualidad.

Por ejemplo, si $\alpha = 0.05$, existe una probabilidad de 0.05 de cometer un error tipo I o de rechazar H_0 cuando es verdadera al hacer un solo test de hipótesis. Si se hacen dos tests, la probabilidad de cometer al menos un error tipo I, es 0.05+0.05=0.1; si se hacen N tests, la probabilidad de cometer al menos un error tipo I aumenta a $0.05 \cdot N$.

El método de Bonferroni ajusta el nivel de significancia α para cada test individual de tal forma que controle la probabilidad de cometer un error tipo I. Bonferroni logra lo anterior dividiendo α global por la cantidad de tests realizados. De esta manera, si se realizan k tests, el nivel de significancia de cada test invididual es $\alpha' = \alpha/k$.

Estas correcciones son efectivas y simples para reducir los errores de tipo I al hacer múltiples tests de hipótesis. Aún así, tiene el contra que al reducir α, aumenta la probabilidad de error tipo II o β (sección 3.1). El error tipo II es cuando H_0 no se rechaza cuando en realidad es falsa, por lo que el complemento de β, $1 - \beta$, llamado **potencia del test**, es la probabilidad de rechazar H_0 cuando efectivamente es falsa. De esta manera, disminuir α también se traduce en disminuir la potencia del test.

3.3. *Permutational Multivariate Analysis of Variance* (PERMANOVA)

El análisis de varianza multivariado permutacional (PERMANOVA) es un test de hipótesis paramétrico multivariado que busca comparar g poblaciones o grupos para investigar si las medias poblacionales son iguales, y si no, encontrar cuales medias difieren significativamente. Es una variación del MANOVA (Anexo C) con la gran diferencia de ser **no-paramétrico**, en particular es un **test multivariado de permutación estadístico no-paramétrico**. En otras palabras, se asume que el vector de observación X_{jl}, que se asume gaussiano en test ANOVA o MANOVA, no sigue ninguna distribución conocida y/o parametrizada.

Este test de hipótesis fue propuesto en [37] el cual toma una perspectiva geométrica dada la naturaleza no-paramétrica del problema. La esencia del análisis de varianzas es comparar la variabilidad dentro de los grupos versus la variablidad entre diferentes grupos usando el ratio del estadístico F de Fisher. En [37] se usa una matriz de distancia euclidiana D de las observaciones. El elemento d_{ij} de la matriz D es la distancia entre el vector de observación x_i y el vector x_j. De esta manera se calcula:

$$SS_{cor} = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \tag{3.3}$$

$$SS_{res} = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{1}{N(x_i)} d_{ij}^2 \epsilon_{ij} \tag{3.4}$$

donde SS_{cor} es la suma de cuadrados total corregido, SS_{tr} es la suma de cuadrados **entre muestras de distintos grupo o tratamiento** y SS_{res} sería la suma de cuadrados **entre muestras del mismo grupo o tratamiento**. En otras palabras SS_{tr} y SS_{res} se relacionan con la varianza, donde el primero mide **la variación debido al tratamiento**, y el segundo **la variación dentro de cada tratamiento**. N es la cantidad total de observaciones, $N(x_i)$ es la cantidad de observaciones del grupo al que pertenece la observación x_i y ϵ_{ij} es 1 si la observación x_i y la observación x_j pertenecen al mismo grupo. Por ultimo $SS_{tr} = SS_{cor} - SS_{res}$.

Es importante notar que los valores SS_{cor}, SS_{tr} y SS_{res} son escalares a diferencia de MANOVA que son matrices. Con esto se toma el enfoque del ANOVA y se calcula un pseudo F-ratio:

$$F = \frac{SS_{tr}/(g-1)}{SS_{res}/\left(\sum_{l=1}^{g} n_l - g\right)}$$

donde g son la cantidad de grupos de la población, y n_l es la cantidad de muestras del grupo l. En este caso este no es un F-ratio de Fisher propiamente tal ya que las observaciones no distribuyen normal.

La parte permutacional viene en el cálculo del p-value. Si se supone que la hipótesis nula es verdadera y los grupos no son realmente diferentes, los vectores de observación debiesen ser "intercambiables" (*exchangeable*). Cada observación posee una etiqueta indicando a qué grupo pertenece, entonces si la hipótesis nula es verdadera, permutar las etiquetas de cada

observación no debiese mostrar ninguna diferencia estadística. Al hacer esta permutación se calcula un nuevo valor de F llamado F^π. Probando distintas permutaciones posibles al azar de las etiquetas y recalculando F^π y comparando con el valor F original se puede calcular el p-value como:

$$\text{p-value} = \frac{\text{n}^\text{o} \text{ de veces que } F^\pi \geq F}{\text{n}^\text{o} \text{ total de permutaciones}}$$

3.4. Análisis de sentimiento

El análisis de sentimiento es una técnica usada en distintas áreas de la industria como el marketing, redes sociales, investigación de mercados, entre otros [38]. Se utiliza para analizar una opinión en forma de texto y clasificar el sentimiento de la opinión como positivo, negativo o neutral. Se pueden utilizar distintos enfoques: 1) métodos basados en reglas, o 2) métodos basados en modelos de aprendizaje automático. Generalmente se usa este último ya que ofrece mejores resultados frente a textos más complejos o con uso de sarcasmo e ironía.

El enfoque basado en reglas se basa en reglas lingüísticas específicas para clasificar los sentimiento y se define manualmente. Funciona buscando palabras o frase claves en el texto que se asocien con un sentimiento negativo o positivo. El sentimiento de la palabra o frase luego es ponderado según una puntuación de sentimiento previamente asignada y la puntuación general de todo el texto se calcula sumando estas ponderaciones. Es importante también considerar negaciones y modificadores. Por ejemplo, la frase "no es bonito" sería una negación. Un modificador por el otro lado cambia la intensidad del sentimiento, por ejemplo la frase "muy triste" intensificaría el sentimiento encontrado. Las mayores ventajas de este tipo de enfoque es la facilidad de implementarlo en casos sencillos y no requiere datos de entrenamiento, pero es menos efectivo cuando se enfrenta a textos más complejos y requiere actualización constante de las reglas.

El enfoque basado en modelos de aprendizaje automático por el otro lado utiliza algoritmos de aprendizaje automático para aprender a identificar los sentimientos [39]. En general dentro de estos tipos de modelos se encuentran los modelo supervisados y no supervisados. Los modelos supervisados se entrenan con un conjunto de datos previamente etiquetados. Estos datos son ejemplos de textos que tienen asociado un etiqueta del sentimiento que lo representa. Por otro lado los modelos no supervisados buscan identificar agrupaciones y patrones en los datos que no tienen etiquetas predefinidas. Dentro de este enfoque también existe lo que es el procesamiento de lenguaje natural (PLN) que se basa en procesar y transformar el texto en un formato que el modelo pueda entender. Estas transformaciones generalmente son vectores de palabras o *embeddings*. Las técnicas más avanzadas en el análisis de sentimiento se encuentran dentro del campo del *deep learning* en los que se usan redes neuronales profundas. Un modelo muy conocido es BERT (Bidirectional Encoder Representation from Transformers) [40] que logró resultados de vanguardia en el ámbito del PLN.

En cuanto al enfoque basado en aprendizaje automático, se puede decir que las mayores ventajas de estos tipos de modelos es que son más precisos y flexibles que los basado en reglas. Son capaces de aprender diferentes contextos y estilos lingüísticos, pero igualmente tienen desventajas. Estos modelos requieren conjuntos de datos etiquetados grandes para

poder entrenarlos y son más complejos y costosos en términos de recursos computacionales.

Un modelo potente que existe es el análisis de sentimiento de Cloud Natural Language API de Google ®. Es posible usarlo a través de Python y permite recibir texto para retornar dos valores numéricos: puntuación (*score*) y magnitud (*magnitude*). El score mide la inclinación emocional dentro del texto y es un valor real entre -1 y 1. Un score cercano a 1 significa que el texto muestra un sentimiento positivo, mientras que un valor de -1 indica un sentimiento negativo. Cuando es cercano a 0 se puede decir que el sentimiento es neutral. La magnitud por el otro lado mide la fuerza del sentimiento, independiente del signo que indica el score. Este valor varía entre 0 e infinito y no está normalizada, así un valor de magnitud más alto se traduce en emociones más fuertes encontradas en el texto.

Capítulo 4

Caso de estudio

El caso de estudio de esta tesis es el análisis de los equipos de trabajo formados en los cursos "Ingeniería de Software" (CC4401) y "Proyecto de Software" (CC5402) del Departamento de Ciencias de la Computación (DCC) de la Universidad de Chile, enfocándose específicamente en cómo la presencia de mujeres en los equipos afecta la percepción del rendimiento tanto propia del estudiante como de sus compañeros, esta se mide a través de autoevaluaciones y coevaluaciones.

A lo largo de este capitulo se presenta el contexto de los cursos dentro del área del desarrollo de software, se explica la plataforma existente del DCC para que los estudiantes respondan las coevaluaciones, y se expone la metodología que se siguió para llevar a cabo el trabajo de esta tesis. En esta última se presenta cómo se modificó la plataforma de coevaluaciones para incluir una sección de autoevaluaciones, cómo se recopilan los datos y cómo estos son procesados para obtener las métricas pertinentes con el objetivo de responder las preguntas de investigación.

4.1. Contexto de los cursos

El caso de estudio se basa en la recopilación y análisis de datos procedentes de las coevaluaciones y autoevaluaciones de dos cursos de la carrera de Ingeniería Civil en Computación del DCC de la Universidad de Chile, CC4401 y CC5402. En los próximos párrafos se describe el objetivo y propósito de ambos cursos en esta carrera dando a entender su relación con el desarrollo de software.

Ingeniería de Software (CC4401) representa el primer contacto de los estudiantes con el trabajo en equipo dentro de la carrera. Se trata de un curso teórico-práctico de séptimo semestre en el que se alternan clases de cátedra con tareas individuales donde se aplican los conceptos presentados para poder desarrollar un proyecto. Este curso tiene como objetivo fundamental que los estudiantes, colaborando en equipos de aproximadamente 5 alumnos, conformados por el cuerpo docente de acuerdo a la compatibilidad y complementaridad de competencias técnicas y de trabajo en equipo de los mismos estudiantes, para que construyan un sistema web con perfiles de usuario. Los proyectos consideran la creación de interfaces gráficas, la comprensión del dominio de negocio y el uso de frameworks para el desarrollo de aplicaciones web. Los equipos deben aprender a trabajar de manera estratégica y colaborativa, participando en diversas actividades formativas. Esto implica la autogestión personal y

la interacción con los demás, alternando roles de líder y colaborador. El curso se divide en dos sprints, en los cuales los equipos deben entregar informes, realizar presentaciones y llevar a cabo coevaluaciones para evaluar a sus compañeros.

Proyecto de Software (CC5402), por otro lado, se imparte en el décimo semestre de la carrera y representa la culminación de los conocimientos adquiridos durante la formación académica. En este curso, los estudiantes se organizan en grupos de 5 a 8 integrantes, determinados por su compatibilidad horaria (también formados por el cuerpo docente), ya que deben cumplir con una cantidad fija de horas trabajadas por semana. A cada equipo se le asigna un proyecto de desarrollo de software para una organización externa al departamento de computación. Los equipos deben establecer relaciones efectivas con los clientes, planificar y adaptar su trabajo según las circunstancias, evaluar los riesgos asociados al proyecto y, en última instancia, desarrollar una solución de software práctica y real. El curso se divide en tres iteraciones, y en cada una de ellas se evalúa la gestión del proyecto, el valor de la solución, la calidad del software, la presentación de los avances y se realizan coevaluaciones internas para evaluar el desempeño del equipo.

4.2. Plataforma de coevaluaciones

La herramienta de coevaluaciones del DCC (https://coevaluaciones.dcc.uchile.cl/) fue desarrollada el año 2014 por el estudiante Roberto Riquelme como parte de su memoria de Ingeniería Civil en Computación [41] dirigida por el profesor guía Sergio Ochoa. Esta herramienta tenía por objetivo el diseño y la implementación de un sistema de evaluación del desempeño de los miembros de un equipo de trabajo. La aplicación utiliza el framework Django[2] tanto para el frontend como el backend y PostgreSQL[3] como motor de base de datos.

Los cursos que tienen acceso a la plataforma son aquellos que imparten cursos específicos, como Ingeniería de Software, Ingeniería de Software II, Proyecto de Software y Formulación, Evaluación y Gestión de Proyectos. En ellos las evaluaciones se dividen en iteraciones o sprints. En cada iteración, los estudiantes presentan los avances en sus proyectos y, como parte de la evaluación, deben responder a una coevaluación que evalúa el desempeño de sus compañeros de grupo. Estas coevaluaciones son obligatorias y contribuyen a un porcentaje específico de la nota de la iteración, lo que agrega un elemento adicional de relevancia a esta plataforma en el ámbito académico.

Las coevaluaciones se componen de preguntas, que se mantienen consistentes a lo largo de los semestres y cursos. Estas preguntas se dividen en dos tipos: abiertas y evaluativas. En las preguntas abiertas, los alumnos deben proporcionar respuestas en forma de texto libre en relación a las fortalezas y debilidades del alumno evaluado. En contraste, las preguntas evaluativas presentan una escala Likert[4] del 1 al 5, y los alumnos deben seleccionar una opción que mejor refleje su evaluación. La escala se define con los siguientes términos:

[2] Framework de desarrollo web de código abierto, escrito en Python.
[3] Sistema de gestión de bases de datos relacional de código abierto y de alto rendimiento.
[4] Método de medición utilizado por los investigadores con el objetivo de evaluar la opinión y actitudes de las personas.

Tabla 4.1: Escala de notas de preguntas evaluativas

1	Nunca
2	Con dificultad
3	A veces
4	Regularmente
5	Siempre

También, las preguntas evaluativas tienen asociados aspectos cualitativos que están vinculados con las habilidades de trabajo en equipo que estas representan. A continuación se muestra una tabla con cada pregunta, su tipo y aspectos cualitativos asociados.

Tabla 4.2: Preguntas de coevaluación junto con su tipo y aspectos cualitativos

	Pregunta	**Tipo**	**Aspectos Cualitativos**
P1	Demuestra compromiso con el proyecto	evaluativa	compromiso
P2	Cumple de manera adecuada con las tareas que le son asignadas	evaluativa	compromiso, contribución, colaboración
P3	Demuestra iniciativa para lograr el éxito del proyecto	evaluativa	motivación, actitud
P4	Mantiene buena comunicación con el resto del equipo	evaluativa	comunicación
P5	Mantiene buena coordinación entre sus tareas y las de sus pares	evaluativa	coordinación
P6	La calidad de su trabajo es la apropiada para lograr el éxito del proyecto	evaluativa	contribución, colaboración
P7	Ofrece apoyo en las tareas que van más allá del rol asignado	evaluativa	motivación, colaboración, actitud
P8	Es capaz de admitir sus equivocaciones y recibir críticas	evaluativa	actitud
P9	Fortalezas	abierta	-
P10	Debilidades	abierta	-

Los aspectos cualitativos asociados también pueden ser cuantificados con un puntaje, el cual se calcula como una ponderación de las preguntas. Esta ponderación se utilizaba en una versión anterior de la plataforma y se ha integrado en este estudio para evaluar objetivamente los aspectos de trabajo en equipo en cuestión de una forma compacta y simple de interpretar. Esta se presenta en la tabla 4.3.

Tabla 4.3: Relación entre aspectos cualitativos y preguntas de coevaluación

Aspectos Cualitativos	**Ponderación**
Actitud	$\frac{1}{3} \cdot P3 + \frac{1}{3} \cdot P7 + \frac{1}{3} \cdot P8$
Colaboración	$\frac{1}{4} \cdot P2 + \frac{1}{4} \cdot P6 + \frac{1}{2} \cdot P7$
Contribución	$\frac{1}{2} \cdot P2 + \frac{1}{2} \cdot P6$
Motivación	$\frac{1}{2} \cdot P3 + \frac{1}{2} \cdot P7$
Coordinación	$P5$
Comunicación	$P4$
Compromiso	$\frac{1}{2} \cdot P1 + \frac{1}{2} \cdot P2$

La plataforma está diseñada para admitir tres tipos de usuarios: profesores, ayudantes y alumnos. Cada uno de estos usuarios debe autenticarse en el sistema utilizando un nombre de usuario y una contraseña. En el contexto de esta tesis, se centrará en los roles de profesor y alumno, ya que son los más relevantes para su estudio.

Las funciones de los profesores incluyen la capacidad de crear cursos, formar grupos dentro de estos cursos, diseñar y publicar coevaluaciones, especificar fechas de inicio y finalización para estas evaluaciones, notificar a los alumnos mediante correos electrónicos sobre las coevaluaciones publicadas y, al finalizar el plazo de cada coevaluación, revisar y descargar los resultados en formato Excel, lo que proporciona una visión detallada del rendimiento de cada alumno, grupo y curso.

Por otro lado, los alumnos tienen dos funciones principales en la plataforma. En primer lugar, pueden responder a las coevaluaciones asignadas a su curso, lo que les permite evaluar el desempeño de sus compañeros de grupo. Una vez finalizado el plazo para responder, los alumnos pueden acceder a sus notas de coevaluación asignadas anónimamente por sus compañeros junto con comentarios positivos y/o negativos. Se incluye el promedio de cada pregunta y su nota final, lo que proporciona una visión completa de su rendimiento en el grupo desde la perspectiva de sus compañeros.

En la figura 4.1 se muestra la vista de un alumno del menú inicial de la plataforma en la que se ve un resumen de todas las coevaluaciones que ha tenido por curso y año. La vista de los profesores es prácticamente igual con la diferencia de que tienen un botón para crear coevaluaciones nuevas, y además en cada coevaluación tienen un botón para editarla (cambiar el nombre y extender fecha de término). En la figura 4.2 se muestra un ejemplo de una coevaluación, donde el alumno evalúa a sus compañeros respondiendo las 8 preguntas de la

tabla 4.2 seleccionando la cantidad de estrellas que considere pertinente, donde la cantidad de estrellas se rige por la tabla 4.1.

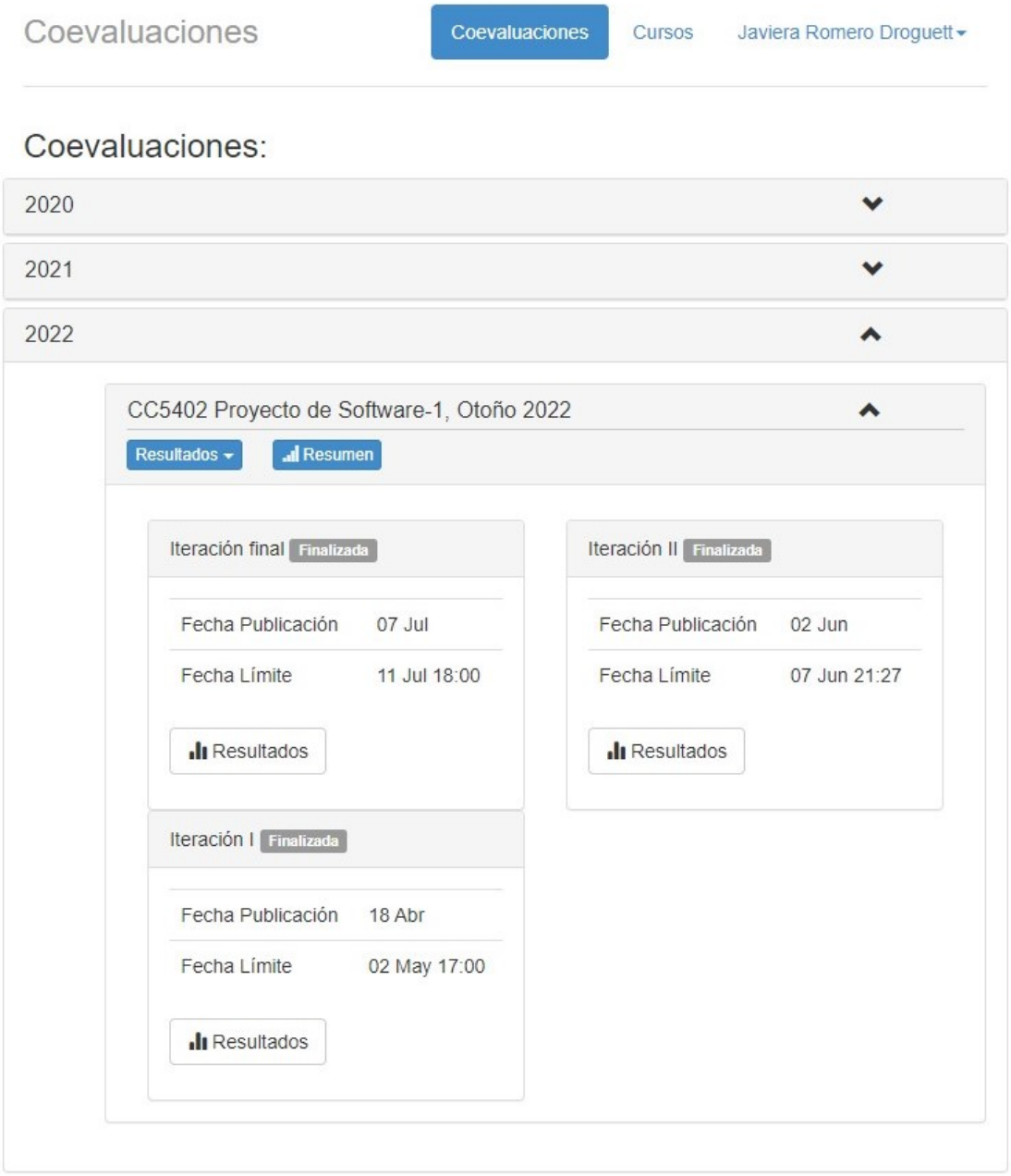


Figura 4.1: Página inicial de coevaluaciones, con el resumen de las coevaluaciones por año

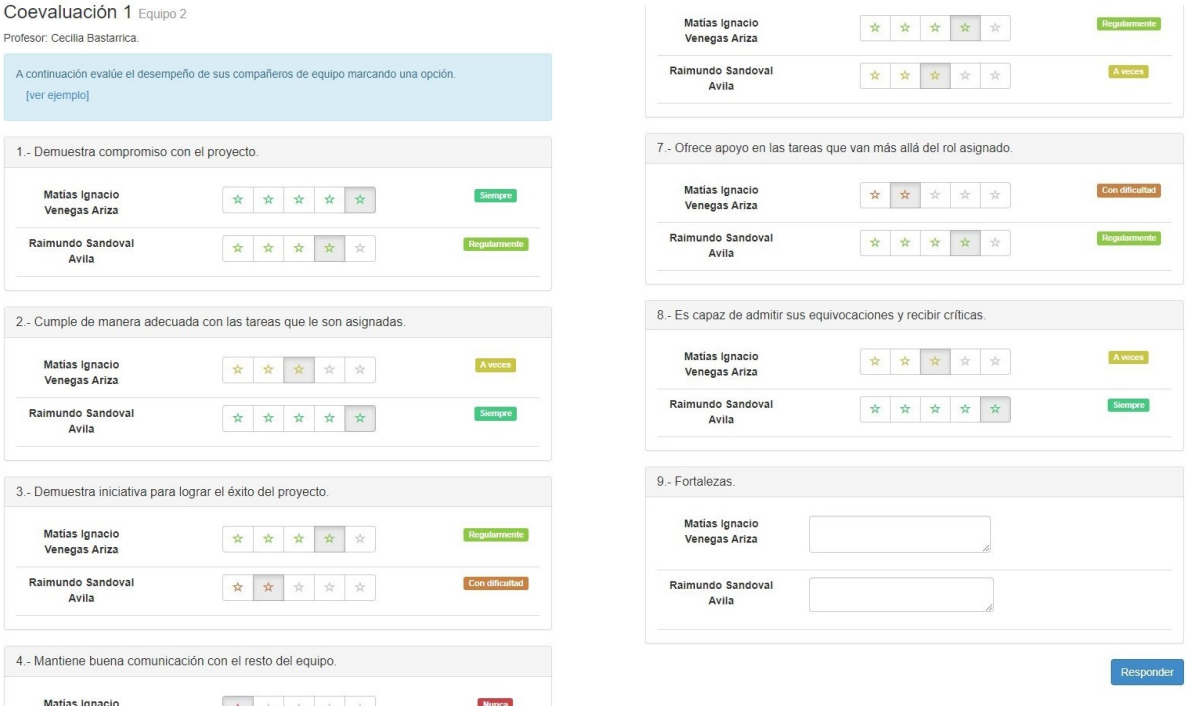


Figura 4.2: Ejemplo de coevaluación

4.3. Metodología

4.3.1. Desarrollo e implementación de plataforma de autoevaluaciones

La plataforma de coevaluaciones existente explicada en la sección 4.2 fue modificada para incluir la nueva herramienta de autoevaluación. Para la implementación de esta nueva herramienta se decidió mantener el diseño y la estructura de la interfaz actual, agregando una nueva pestaña a la barra de menú con un botón para dirigirse a Autoevaluaciones. Las autoevaluaciones (análogas a las coevaluaciones descritas anteriormente) se crean con el objetivo de que los alumnos puedan además de evaluar a sus compañeros, evaluarse a sí mismos respecto al desempeño que tuvieron dentro del equipo. Las preguntas de autoevaluación son las mismas usadas para la coevaluación (Tabla 4.2) cumpliendo tambien la transformación a aspectos cualitativos (Tabla 4.3).

Para llevar a cabo estos cambios, se trabajó en el repositorio de Github[5] del proyecto a través de la herramienta Gitpod[6]. En primer lugar se trabajó con el backend de la plataforma agregando tres nuevas tablas a la base de datos existente, *Autoevaluación*, *RespuestaEvalAuto* y *RespuestaTxtAuto*, las cuales se crean a través de modelos de Django. El modelo de datos de las tablas que se agregaron se puede ver en la figura 4.3.

[5] Plataforma de alojamiento y colaboración en línea para proyectos de desarrollo de software que utilizan el sistema de control de versiones Git.

[6] Entorno de desarrollo en la nube que promueve la colaboración efectiva entre los miembros del equipo en proyectos de software alojados en GitHub.

La tabla *Autoevaluación* se relaciona con la tabla *Curso*. Esto significa que todos los alumnos inscritos en un curso específico están automáticamente asociados a la autoevaluación creada. También, guarda la información del nombre de la autoevaluación, la fecha límite para responderla y el estado actual de la autoevaluación que puede ser "creada", "publicada" o "finalizada".

La tabla *RespuestaEvalAuto* se relaciona con una pregunta específica, un estudiante, un grupo y una autoevaluación. Además, se guarda la respuesta de tipo numérica con un valor de 1 a 5, lo que permite una evaluación cuantitativa.

La tabla *RespuestaTxtAuto* es análoga a *RespuestaEvalAuto*, con la diferencia de que la respuesta es de tipo texto, lo que permite una evaluación cualitativa y más detallada.

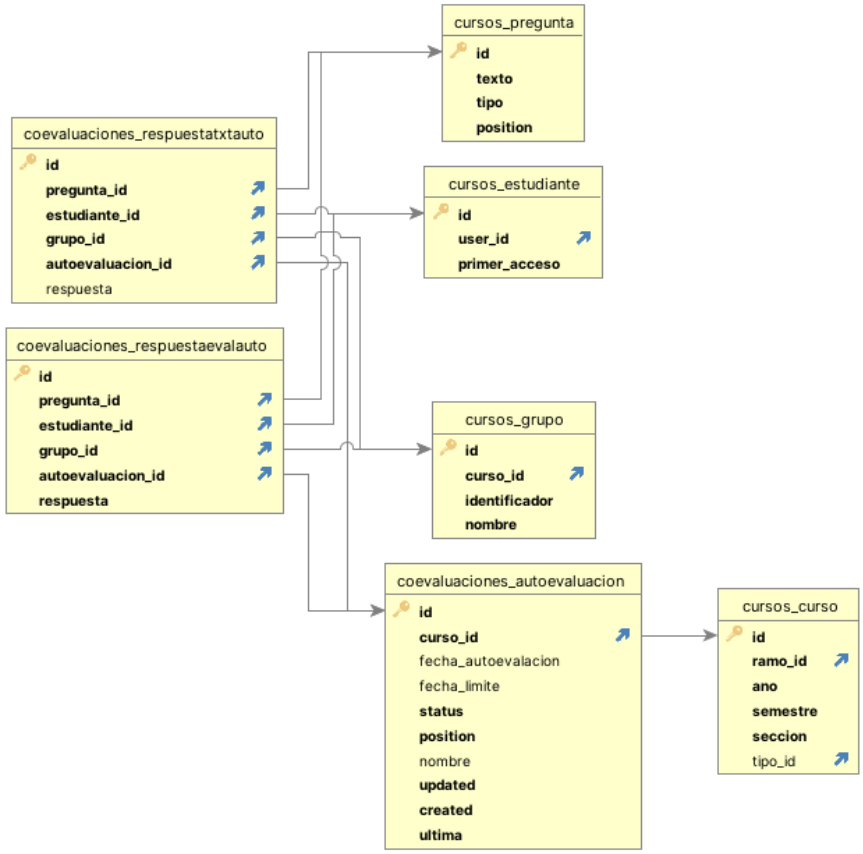


Figura 4.3: Modelo de datos de las tablas *Autoevaluación*, *RespuestasEvalAuto* y *RespuestasTxtAuto*

Luego, en el frontend se realizó un cambio en el botón de la barra de menú. Se reemplazó por una lista desplegable que permite a los usuarios seleccionar entre "Coevaluaciones" o "Autoevaluaciones" para dirigirse a las diferentes secciones. Esto se puede observar en la figura 4.4.

Figura 4.4: Barra de menú actualizada

La creación de autoevaluaciones sigue siendo una función del profesor, quien selecciona la fecha de publicación y la fecha límite para completarlas. La vista para los estudiantes se mantiene prácticamente idéntica a la de las coevaluaciones, lo que facilita su uso. La única diferencia radica en que, al final del cuestionario, los estudiantes deben otorgar su autorización para que su información sea utilizada para la realización de esta investigación. La vista del cuestionario de las 10 preguntas y la autorización que deben dar los alumnos se muestra en la figura 4.5.

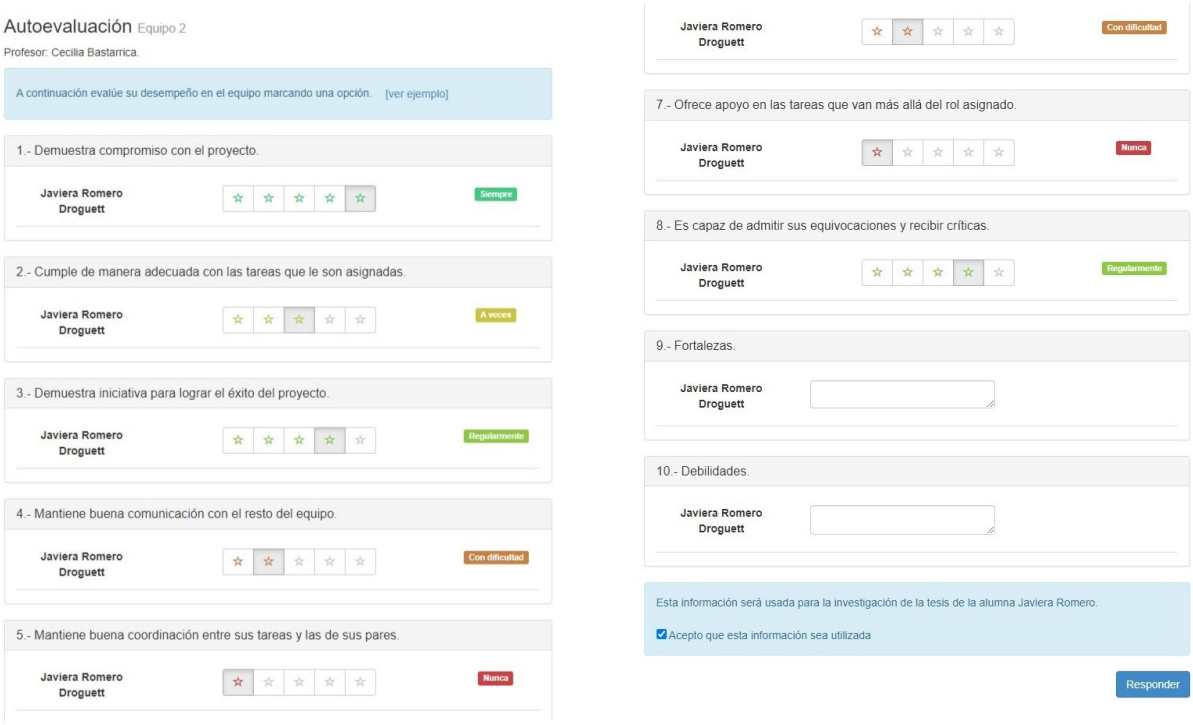


Figura 4.5: Ejemplo de autoevaluación

Esta actualización de la herramienta fue testeada y enviada a producción por el equipo de Software del DCC en junio de 2023, pero no fue desplegada en el servidor de Coevaluaciones

(url mencionada en la sección 4.2). En su lugar, se encuentra en un ambiente productivo del DCC (apps.dcc.uchile.cl/coevaluaciones). Esta decisión se tomó con el propósito de evitar cualquier interferencia con las coevaluaciones publicadas en los cursos del semestre. El sistema fue desplegado en el ambiente con una copia de la base de datos del sistema en producción, por lo que todos los alumnos pueden acceder al servidor con sus credenciales habituales.

4.3.2. Fuente y recolección de datos

Como se mencionó en la sección 4.3.1, la implementación de la herramienta de autoevaluaciones tuvo lugar en junio de 2023. En ese momento, se decidió lanzar una autoevaluación de carácter voluntario al término del primer semestre en los cursos CC4401 y CC5402. Esta autoevaluación se llevó a cabo de manera simultánea con las coevaluaciones finales de ambos cursos.

Para proporcionar una visión más detallada de los datos recopilados, se presenta a continuación la tabla 4.4, que muestra la cantidad de alumnos y las respuestas obtenidas en cada uno de los cursos durante el semestre Otoño 2023:

Tabla 4.4: Cantidad de alumnos que respondieron las coevaluaciones y autoevaluaciones en los cursos del semestre Otoño 2023

	Alumnos en el curso			**Respuestas coevaluación**			**Respuestas autoevaluación**		
	Total	Hombres	Mujeres	Total	Hombres	Mujeres	Total	Hombres	Mujeres
CC4401	136	109	27	134	107	27	36	32	4
CC5402	43	37	6	41	35	6	21	20	1

Las cifras mencionadas anteriormente consideran la cantidad de alumnos que respondieron las preguntas evaluativas, ya que estas son obligatorias al enviar los formularios de la coevaluación y autoevaluación. En cambio, las preguntas abiertas son opcionales por lo que se tiene un número diferente de respuestas. Para el análisis de estas preguntas se considerará la cantidad de respuestas obtenidas por curso, evaluación y género, los cuales se muestran en la tabla 4.5.

Tabla 4.5: Cantidad de respuestas obtenidas en preguntas abiertas de coevaluaciones y autoevaluaciones

	Respuestas abiertas coevaluación						**Respuestas abiertas autoevaluación**					
	Fortalezas			Debilidades			Fortalezas			Debilidades		
	Total	Hombres	Mujeres	Total	Hombres	Mujeres	Total	Hombres	Mujeres	Total	Hombres	Mujeres
CC4401	336	267	69	213	178	35	20	19	1	21	20	1
CC5402	28	21	7	8	8	0	3	3	0	3	3	0

La obtención de datos de cada curso se realizó desde la herramienta, exportando los archivos con las respuestas de coevaluación y autoevaluación de cada curso en formato Excel.

4.3.3. Estructura de los datos

Para el análisis y la manipulación de los datos, se utilizó la librería Pandas[7], que permite importar los datos desde archivos Excel y transformarlos en DataFrames en Python, lo que facilita su procesamiento y análisis.

La tabla 4.6 presenta una descripción de las columnas de los DataFrames y especifica qué columnas están presentes tanto en las coevaluaciones como en las autoevaluaciones.

Tabla 4.6: Descripción de las columnas del DataFrame original al cargar el Excel con los datos de coevaluación y autoevaluación

Columna del DataFrame	**Descripción**	**Coevaluación**	**Autoevaluación**
Equipo	Nombre del equipo	✓	✓
Evaluador	Usuario del alumno que evalúa	✓	✓
Género evaluador	Género del alumno que evalúa	✓	✓
Evaluado	Usuario del alumno evaluado	✓	
Género evaluado	Género del alumno evaluado	✓	
P1	Nota pregunta 1	✓	✓
P2	Nota pregunta 2	✓	✓
P3	Nota pregunta 3	✓	✓
P4	Nota pregunta 4	✓	✓
P5	Nota pregunta 5	✓	✓
P6	Nota pregunta 6	✓	✓
P7	Nota pregunta 7	✓	✓
P8	Nota pregunta 8	✓	✓
Nota parcial	Promedio de las 8 preguntas de las columnas anteriores, es decir la nota asignada por la evaluación de un compañero	✓	✓
Fortalezas	Pregunta abierta	✓	✓
Debilidades	Pregunta abierta	✓	✓

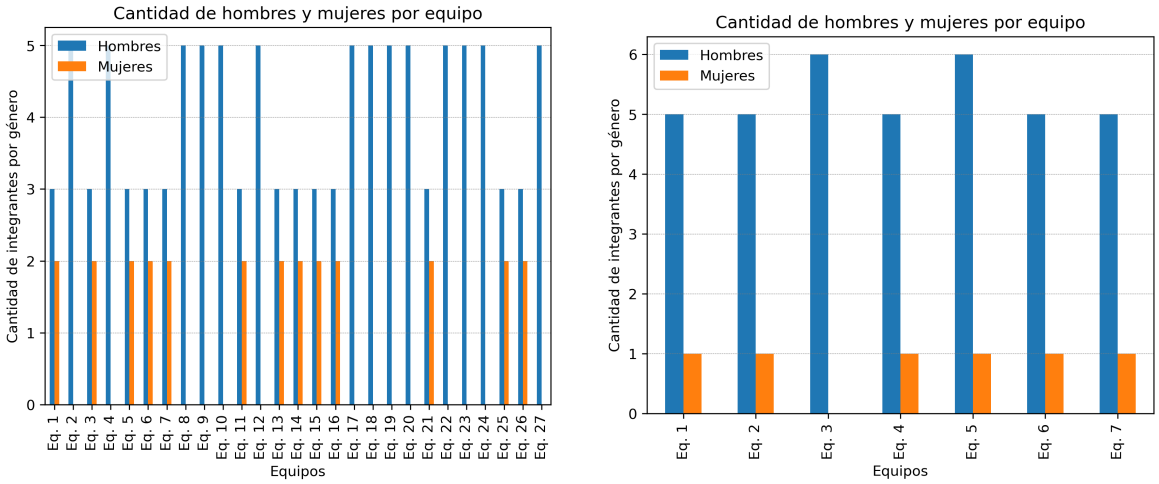
Es importante señalar que en el DataFrame de coevaluación, cada fila o instancia corresponde a la coevaluación realizada por un miembro del equipo (Evaluador) a otro compañero (Evaluado). Como un equipo consta de N integrantes, un estudiante evaluado tendrá $N-1$ filas asociadas correspondiente a las coevaluaciones de sus compañeros, y de la misma forma, un estudiante evaluador tendrá $N-1$ filas asociadas correspondiente a las coevaluaciones que respondió de sus compañeros.

Otro punto a destacar es el hecho que para el caso de las autoevaluaciones el DataFrame no tiene las columnas de “Evaluado” y “Género Evaluado” ya que el estudiante se evalúa a sí mismo transformándolo en evaluador y evaluado. Por esta razón solo se poseen las columnas de “Evaluador” y “Género Evaluador” en este caso.

[7] Librería de código abierto en Python diseñada para el análisis y manipulación de datos de manera eficiente y fácil.

4.3.4. Información preliminar de los datos

De la tabla 4.4 se puede observar que en general la cantidad de mujeres en los cursos es bastante menor comparada con los hombres. En CC4401 las mujeres representan el 19.9 % y en CC5402 el 16.2 %. Dado esto, la cantidad de mujeres en los grupos fue repartido de manera equitativa, lo que significa que hubo intervención por parte del cuerpo docente en la formación de los equipos. Este punto es importante ya que la distribución de género en los equipos de trabajo no es aleatorio. En particular, en la figura 4.6 se puede observar la composición de género por equipo de cada curso. En CC4401 hay 27 equipos con 5 integrantes, algunos contienen 5 hombres, mientras otros 3 hombres y exactamente 2 mujeres. En el caso de CC5402, hay 7 equipos, donde 5 de los equipos se componen de 5 hombres y 1 mujer, un equipo se compone de 6 hombres y 1 mujer, mientras que el restante se conforma únicamente de hombres.



(a) CC4401 (b) CC5402

Figura 4.6: Gráficos de barras que muestran la cantidad de hombres y mujeres por equipo

Se puede ver que en ambos cursos existen grupos con y sin mujeres. En el caso de CC4401 los equipos con mujeres tienen exactamente 2 mujeres, mientras que en CC5402 tienen sólo una. En la tabla 4.7 se resume la información según la presencia de mujeres en los equipos y de ella se extrae que en CC4401 la cantidad de equipos con y sin mujeres está más equilibrado; 13 equipos con mujeres y 14 sin mujeres. En contraste, en CC5402, existen 6 equipos con mujeres y sólo uno sin mujeres. El hecho que existan grupos con y sin mujeres ayuda al estudio de esta tesis, ya que permite hacer comparaciones según la presencia de mujeres en los equipos.

Tabla 4.7: Caracterización del tipo de equipo para CC4401 y CC5402

	Tipo de equipo	**Cantidad de equipos**	**Cantidad de mujeres**
CC4401	**Con mujeres**	13	2
	Sin mujeres	14	0
CC5402	**Con mujeres**	6	1
	Sin mujeres	1	0

4.3.5. Preprocesamiento de los datos

Antes de poder calcular estadísticas, los datos deben ser preprocesados. De la tabla 4.6 se puede observar que existen 8 preguntas que se responden en una escala numérica mientras que las dos últimas de fortalezas y debilidades son preguntas abiertas de texto. Ambos tipos de preguntas son preprocesadas de distintas formas como se explicará a continuación.

En los DataFrames de ambos tipos de evaluación se transformaron las columnas de las preguntas (P1 a P8) a columnas correspondientes a los aspectos cualitativos de acuerdo a las ponderaciones de la tabla 4.3. De esta manera se trabajó directamente con los aspectos cualitativos que representan el trabajo en equipo. En la tabla 4.8 se describen estas nuevas columnas.

Por otro lado, las preguntas de fortalezas y debilidades de los DataFrames se les aplica análisis de sentimiento utilizando el Cloud Natural Language API de Google ® mencionado en la sección 3.4. Dada la naturaleza de estas preguntas solo se utilizará la medida de score del análisis de sentimiento. Existe una claridad en la polaridad del sentimiento; las respuestas en fortalezas debiesen estar inclinadas a sentimientos positivos y las debilidades a sentimientos negativos, por lo que el valor absoluto de score ya entrega una medida relativa de la intensidad del sentimiento. La magnitud en este caso se vuelve redundante ya que este valor es más útil cuando se encuentra una mezcla de sentimientos. Por ejemplo, un texto que presenta sentimientos positivos fuertes y sentimientos negativos fuertes podría mostrar un score cercano a 0, pero la magnitud puede tener un valor alto. Este tipo de casos no deberían ocurrir dado el contexto unipolar de las preguntas. Por último, dada estas razones, el score que se presentará en los resultados será el valor absoluto, omitiendo el signo de este ya que se infiere la dirección del sentimiento por el contexto. Así con este análisis de sentimiento y la transformaciones a aspectos cualitativo se crea el DataFrame base para el resto del procesamiento de aquí en adelante y se muestra en la tabla 4.8.

Tabla 4.8: Descripción de las columnas del DataFrame de la coevaluación y autoevaluación con los aspectos cualitativos

Columna del DataFrame	**Descripción**	**Coevaluación**	**Autoevaluación**
Equipo	Nombre del equipo	✓	✓
Evaluador	Usuario del alumno que evalúa	✓	✓
Género evaluador	Género del alumno que evalúa	✓	✓
Evaluado	Usuario del alumno evaluado	✓	
Género evaluado	Género del alumno evaluado	✓	
Actitud	$\frac{1}{3} \cdot P3 + \frac{1}{3} \cdot P7 + \frac{1}{3} \cdot P8$	✓	✓
Colaboración	$\frac{1}{4} \cdot P2 + \frac{1}{4} \cdot P6 + \frac{1}{2} \cdot P7$	✓	✓
Contribución	$\frac{1}{2} \cdot P2 + \frac{1}{2} \cdot P6$	✓	✓
Motivación	$\frac{1}{2} \cdot P3 + \frac{1}{2} \cdot P7$	✓	✓
Coordinación	$P5$	✓	✓
Comunicación	$P4$	✓	✓
Compromiso	$\frac{1}{2} \cdot P1 + \frac{1}{2} \cdot P2$	✓	✓
Score de Fortalezas	Score obtenido del análisis de sentimiento del texto de las fortalezas	✓	✓
Score de Debilidades	Score obtenido del análisis de sentimiento del texto de las debilidades	✓	✓

4.3.6. Definición de Métricas

En el contexto de la evaluación del rendimiento de las mujeres en los equipos de trabajo de los cursos CC4401 y CC5402, es esencial establecer una serie de métricas específicas que permitan un análisis. Son fundamentales para obtener una visión clara y detallada de la dinámica de género en los equipos y su impacto en las evaluaciones, y estas se calcularán de forma separada para cada uno de los cursos.

Es importante destacar que es posible calcular estas métricas tanto para los puntajes de aspectos cualitativos como para los scores del análisis de sentimiento de las preguntas de fortalezas y debilidades al ser ambos valores numéricos. Para los puntajes de aspectos cualitativos se calcularán las 3 métricas, mientras que para el caso de las fortalezas y debilidades solo se calculará la segunda métrica. A continuación se exponen y explican las 3 métricas a utilizar y cómo se calculan a partir de los datos.

- **Métrica 1: Promedio de los puntajes/scores por género**

Esta métrica se enfoca en el rendimiento de los estudiantes en función de su género. Se

calculará el promedio de los puntajes asociados a cada aspecto cualitativo y los scores de las fortalezas y debilidades de acuerdo a las respuestas de los alumnos en las coevaluaciones y autoevaluaciones, desglosado por género. Esto permitirá una comparación directa para determinar si existen diferencias significativas en el rendimiento entre hombres y mujeres en estos cursos. Para obtener estas métricas se agrupan los datos según el género evaluado creando dos grupos: hombres evaluados y mujeres evaluadas. Luego se calcula el promedio de cada grupo por columna usando métodos de Pandas.

También se calcula el promedio general de los 7 aspectos cualitativos junto con la desviación estándar (a partir de los promedios de cada aspecto cualitativo) de ambos grupos como una métrica más general para la comparación.

- **Métrica 2: Promedio de los puntajes/scores por equipo según presencia femenina**

  El análisis de esta métrica es de particular importancia, ya que se centra en cómo se desempeñan los grupos en función de si tienen integrantes mujeres o no. En ambos cursos existe dos tipos de equipos: con mujeres y sin mujeres. Esta métrica busca medir cómo son coevaluados y autoevaluados los grupos según la presencia femenina.

  Para calcular esta métrica primero se calcula el rendimiento individual de cada alumno. En el caso de la autoevaluación no es necesario al ser el mismo alumno el que se evalúa así mismo, pero en el caso de la coevaluaciones hay que calcular el promedio de los puntajes (tanto de aspectos cualitativos como de análisis de sentimiento) asignados al alumno evaluado por sus compañeros. Luego se obtiene el puntaje promedio por equipo calculando el promedio de los puntajes de cada alumno. Y por último se agrupan los equipos según la presencia femenina, calculando el promedio de los puntajes grupales de los equipos con mujeres y sin mujeres.

  De igual forma que en la métrica anterior se calcula un promedio general y desviación estándar de los 7 aspectos cualitativos de cada tipo de equipo.

- **Métrica 3: Promedio de los puntajes de aspectos cualitativos según coevaluación entre géneros**

  Esta métrica ofrece una visión más detallada de cómo se llevan a cabo las evaluaciones entre los estudiantes de diferentes géneros. Se analizarán las coevaluaciones realizadas entre hombres a mujeres, hombres a hombres, mujeres a hombres y mujeres a mujeres. Esto permitirá descubrir posibles patrones o diferencias significativas en la forma en que los estudiantes se evalúan mutuamente según el género. Para obtener esta métrica se agrupan los datos según el género evaluador y género evaluado obteniendo las cuatro combinaciones mencionadas anteriormente. Después se calcula el promedio de cada uno de estos grupos.

Al igual que en las métricas anteriores se calcula el promedio general y desviación estándar de los 7 aspectos cualitativos de cada tipo de combinación de géneros de evaluador/evaluado.

4.3.7. Análisis estadístico de las diferencias

En la sección anterior se explicó cómo se calculan las métricas y se puede notar que al hacer esto los datos se agrupan según el objetivo de la métrica. En la métrica 1 se agrupan los estudiantes según el género, en la métrica 2 por grupo según la presencia femenina y en la métrica 3 según el género evaluado y género evaluador. Las comparaciones entre los grupos se hacen usando el p-value del análisis de varianza multivariado permutacional PERMANOVA [37]. Este es un test de hipótesis multivariado de permutación estadístico no-paramétrico que compara estadísticamente la media y la varianza de distintas poblaciones (o grupos) para probar su parecido (o no), como se explicó en la sección 3.3.

En particular se usó este tipo de test por tres razones:

1. Se pueden comparar más de dos grupos
2. Es multivariado: cada muestra de las evaluaciones consiste en un vector de 7 dimensiones correspondientes a los 7 aspectos cualitativos.
3. Es no-paramétrico: esto se relaciona con ser permutacional, lo cual asegura robustez frente a muestras pequeñas (pocos alumnos), que por la misma razón, no siguen una distribución conocida.

Para cuantificar la significancia estadística de las diferencias se usó el **p-value** del PERMANOVA. Mientras menor es el p-value, más significativa es la diferencia. El p-value se acota por nivel de significancia α, por lo que un p-value $< \alpha$ se traduce en que la diferencia es signficativa al $(100 \cdot \alpha)\,\%$ (ver sección 3.1).

En general, en los estudios se usan niveles de significancia α de 0.01 (1 %) y 0.05 (5 %) ya que logran un buen equilibrio entre sensibilidad y especificidad. Para este estudio se considerará un nivel del 5 % por la poca cantidad de datos y la naturaleza subjetiva de las respuestas de coevaluación y autoevaluación.

Al calcular cada métrica, se hacen múltiples PERMANOVAs para comparar más de 2 grupos y también entre ambos cursos, por lo que amerita realizar correcciones de Bonferroni. Aún así, en este estudio particular se decidió no realizarlas. Esto porque, como se discutió en la sección 3.2, al disminuir el nivel de significancia α al hacer la corrección, también disminuye la potencia del test. Este efecto es crítico por el hecho de que el tamaño de muestra es pequeño. Además, como se verá en los resultados, las respuestas de los alumnos contienen un efecto techo, lo que significa que los alumnos tienden a calificar con el mayor puntaje posible, por lo que los efectos que se pueden esperar en los resultados son sutiles, perdiendo la capacidad de detectar un diferencia significativa al disminuir α.

Capítulo 5

Resultados y Análisis

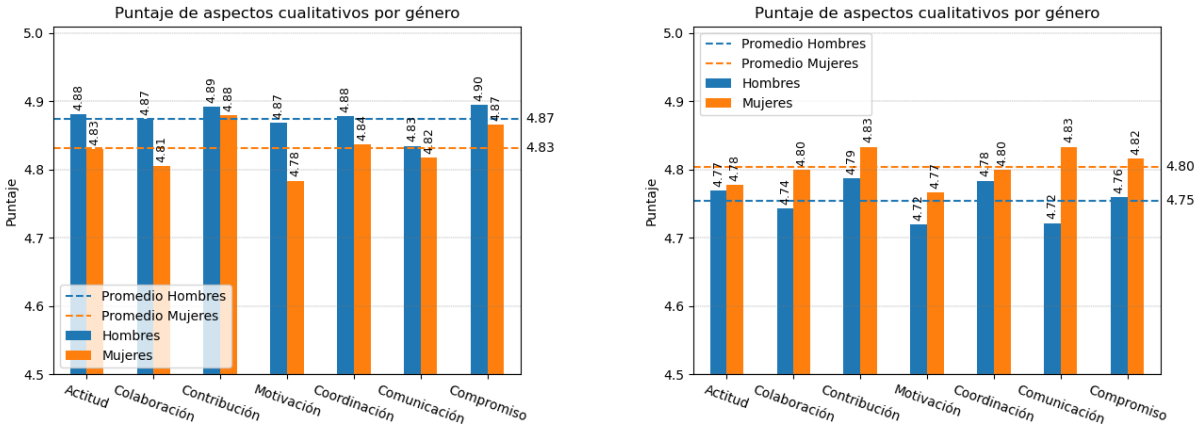
A continuación se muestran los gráficos y tablas con los resultados del estudio hecho según las métricas descritas en la sección 4.3.6, junto con su análisis numérico y de significancia estadística de las diferencias. La discusión de estos resultados serán abordados en el capítulo 6.

5.1. Promedio de los puntajes de aspectos cualitativos por género

En esta sección se presentan los resultados relacionados a la primera métrica tanto para las coevaluaciones y autoevaluaciones. Como se detalló en la sección 4.3.6, se calcula el promedio de los puntajes de los aspectos cualitativos separado por género.

5.1.1. Coevaluación

En la figura 5.1 se muestran dos gráficos de barras que presentan el promedio del puntaje de cada aspectos cualitativos de los alumnos evaluados según su género.



(a) CC4401 (b) CC5402

Figura 5.1: Gráficos de barras del promedio del puntaje de cada aspecto cualitativo separada por género del estudiante evaluado para la coevaluación

En la figura 5.1.a (CC4401) el género masculino alcanzó mayor puntaje en todos los aspectos cualitativos, obteniendo en promedio 0.04 puntos más que las mujeres. En cambio, como se ve en la figura 5.1.b en CC5402 ocurre lo opuesto, es decir el género femenino obtuvo mayor puntaje en todos los aspectos cualitativos, obteniendo un promedio 0.05 puntos mas alto que los hombres.

Los datos de la figura 5.1 se resumen en la tabla 5.1. En esta tabla se muestra además la desviación estándar de las preguntas según el género por curso y el coeficiente de variación, que indica la magnitud de la desviación estándar con respecto al promedio ($\frac{desv.estándar}{promedio} \cdot 100$). La mayor variación fue para los hombres de CC5402, con un valor de 0.63 %. Esto muestra que los puntajes de los aspectos cualitativos están estrechamente agrupados alrededor del promedio por lo que los promedios de cada género por curso son representativos.

Tabla 5.1: Promedio de los puntajes de cada aspecto cualitativo junto con el promedio, desviación estándar y coeficiente de variación de ellas, separado por género del estudiante evaluado en las coevaluaciones

		Actitud	**Colaboración**	**Contribución**	**Motivación**	**Coordinación**	**Comunicación**	**Compromiso**	**Promedio**	**Desviación Estándar**	**Coeficiente de Variación**
CC4401	**Hombres**	4.88	4.87	4.89	4.87	4.88	4.83	4.90	**4.87**	**0.02**	**0.41 %**
	Mujeres	4.83	4.81	4.88	4.78	4.84	4.82	4.87	**4.83**	**0.03**	**0.62 %**
CC5402	**Hombres**	4.77	4.74	4.79	4.72	4.78	4.72	4.76	**4.75**	**0.03**	**0.63 %**
	Mujeres	4.78	4.80	4.83	4.77	4.80	4.83	4.82	**4.80**	**0.02**	**0.42 %**

Dado el análisis de la tabla 5.1 y que los promedios por curso y género son representativos, se puede analizar la diferencia entre los promedios de los géneros por cursos con mayor significancia. Como se mencionó antes, en la figura 5.1.a (CC4401) los hombres obtienen en promedio 0.04 puntos más que las mujeres y en 5.1.b (CC5402) las mujeres obtiene en promedio 0.05 puntos más que los hombres. Estas diferencia entre géneros son relativamente pequeñas y podrían considerarse no significantes.

Para medir la significancia estadística de estas diferencias, se realizó un PERMANOVA usando los 7 aspectos cualitativos. Este test de hipótesis se realizó entre ambos géneros del mismo curso para dicha diferencia como también entre el mismo género de distintos cursos. Los p-values de estas combinaciones se aprecia en la tabla 5.2.

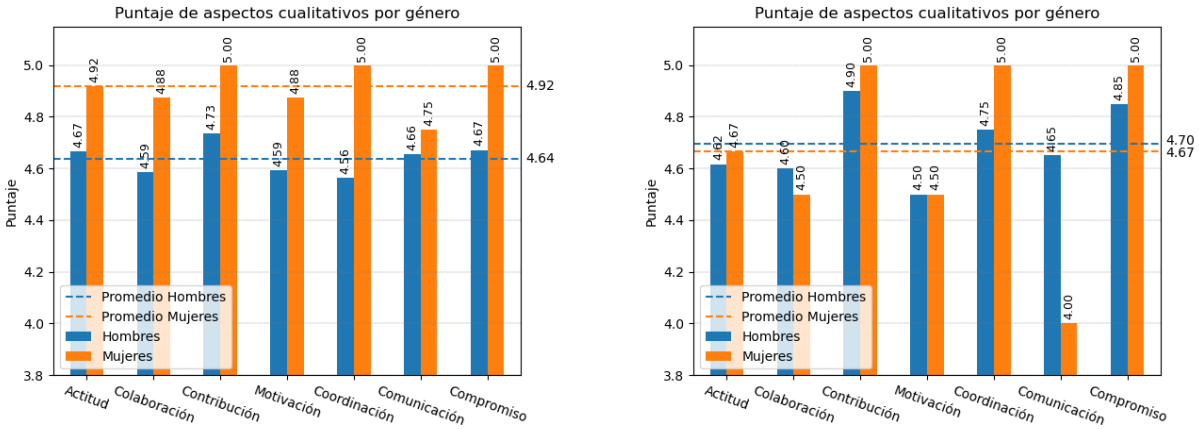
Tabla 5.2: P-values del PERMANOVA realizado entre géneros de los cursos CC4401 y CC5402 para los aspectos cualitativos de las coevaluaciones

		CC4401	**CC5402**
		Mujeres	**Hombres**
CC4401	**Hombres**	0.815	0.212
CC5402	**Mujeres**	1	0.901

Se puede confirmar que la diferencia entre distintos géneros del mismo curso no es significativa al obtener p-values mucho mayores al nivel de significancia $\alpha = 0.05$ (0.815 y 0.901). Lo mismo ocurre comparando el mismo género entre cursos donde entre mujeres el p-value es de 1 y entre hombres 0.212.

5.1.2. Autoevaluación

En la figura 5.2 se muestran gráficos de barras en los que se compara el promedio de los puntajes de los aspectos cualitativos de autoevaluación por género del estudiante.



(a) CC4401 (b) CC5402

Figura 5.2: Gráficos de barras del promedio de los puntajes de cada aspecto cualitativo separada por género del estudiante para la autoevaluación

Analizando los datos de la figura 5.2.a se puede ver que en el curso CC4401, las mujeres tienen mejor puntaje que los hombres en todos los aspectos cualitativos, logrando una diferencia de promedio de 0.28 puntos. Por otro lado, en 5.2.b se observa que los hombres se autoevaluaron mejor en "Colaboración" y "Comunicación", y a pesar de que en todos los otros aspectos las mujeres tienen mejores puntajes, en promedio los hombres las superan por 0.03 puntos.

En la tabla 5.3 se resume la información de las figuras, donde se puede ver que los coeficientes de variación son más altos que en el caso de las coevaluaciones. Esto tiene sentido, ya que la cantidad de respuestas de autoevaluaciones es mucho menor a la de coevaluaciones. La mayor variación fue para las mujeres en CC5402 con 7.92 %, ya que en este curso solo una mujer respondió la encuesta.

Tabla 5.3: Promedio de los puntajes de cada aspecto cualitativo junto con el promedio, desviación estándar y coeficiente de variación de ellas, separado por género del estudiante evaluado en las autoevaluaciones

		Actitud	**Colaboración**	**Contribución**	**Motivación**	**Coordinación**	**Comunicación**	**Compromiso**	**Promedio**	**Desviación Estándar**	**Coeficiente de Variación**
CC4401	**Hombres**	4.67	4.59	4.73	4.59	4.56	4.66	4.67	**4.64**	**0.06**	**1.29 %**
	Mujeres	4.92	4.88	5.00	4.88	5.00	4.75	5.00	**4.92**	**0.09**	**1.83 %**
CC5402	**Hombres**	4.62	4.60	4.90	4.50	4.75	4.65	4.85	**4.70**	**0.14**	**2.98 %**
	Mujeres	4.67	4.50	5.00	4.50	5.00	4.00	5.00	**4.67**	**0.37**	**7.92 %**

Después, se realizó un PERMANOVA para medir la significancia estadística de las diferencias usando los aspectos cualitativos entre ambos géneros del mismo curso como también entre el mismo género de distintos cursos. Los p-values resultantes se pueden ver en la tabla 5.4.

Tabla 5.4: P-values del PERMANOVA realizado entre géneros de los cursos CC4401 y CC5402 para los aspectos cualitativos de las autoevaluaciones

		CC4401	**CC5402**
		Mujeres	**Hombres**
CC4401	**Hombres**	0.291	0.587
CC5402	**Mujeres**	0.401	0.905

Con estos resultados se confirma que las diferencias no son significativas al obtener p-values mucho mayores al nivel de significancia $\alpha = 0.05$.

5.2. Promedio de los scores de fortalezas y debilidades por género

En esta sección se presentan los resultados del análisis de sentimiento relacionado a la primera métrica tanto para las coevaluaciones como para las autoevaluaciones. Se calculan de forma análoga al caso de los puntajes de aspectos cualitativos.

5.2.1. Coevaluación

En la tabla 5.5 se presentan los promedios de los scores de fortalezas y debilidades agrupados por género. De esta tabla se puede notar en primer lugar, que para las fortalezas del curso CC4401 los scores están más equiparados donde los hombres superan a las mujeres por solo 0.06 puntos. En CC5402 esta diferencia está más acentuada donde los hombres reciben opiniones notablemente más fuertes con un diferencia de score de 0.34 puntos. De esta manera se puede decir que en general los hombres reciben coevaluaciones más positivas que las mujeres.

Para las debilidades se puede observar que para el curso CC4401 las opiniones son bastante más débiles en ambos géneros comparado con las fortalezas. Aún así se sigue manteniendo una diferencia pequeña entre los géneros, donde en este caso las mujeres superan a los hombres por 0.08 puntos. Para el curso CC5402 se puede observar que en ambos géneros las opiniones son más fuertes que en el caso de las fortalezas ya que los scores son más altos. En este caso las mujeres reciben opiniones más fuertes superando a los hombres por 0.18 puntos. De esta manera se puede concluir que en cuanto a debilidades, las mujeres reciben en promedio comentarios más negativos que los hombres.

Tabla 5.5: Promedio de los scores de fortalezas y debilidades por género del alumno evaluado en coevaluaciones

		Fortalezas	**Debilidades**
CC4401	**Hombres**	0.79	0.13
	Mujeres	0.73	0.21
CC5402	**Hombres**	0.57	0.63
	Mujeres	0.23	0.81

5.2.2. Autoevaluación

En el caso de las autoevaluaciones, en la tabla 5.6 se muestran los promedios de los scores de los comentarios de fortalezas y debilidades de los alumnos agrupados según el género. Como se observa en la tabla 4.5, ninguna mujer de CC5402 respondió las preguntas abiertas de autoevaluación, por lo que no se pueden observar estos resultados en la tabla 5.6. En la tabla se puede ver que en el curso CC4401 los hombres reciben respuestas de fortalezas con menor score que las mujeres, con un diferencia de 0.14. Observando los resultados de las debilidades se nota el mismo patrón pero con una diferencia más acentuada, donde el score de las mujeres superan al de los hombres por 0.47 puntos. De esta manera es posible observar que para CC4401, las mujeres se autoevalúan de forma más positiva en cuanto a las fortalezas y a la vez se autoevalúan de manera más negativas que los hombres en cuanto a debilidades.

En el caso del curso CC5402, no existen autoevaluaciones de las mujeres. Aún así es posible analizar la intensidad de la opinión de los hombres de su trabajo personal en el curso. En particular en este curso los hombres autoevalúan sus fortalezas con menor intensidad que ambos géneros del otro curso. En cuanto a las debilidades se pueden notar que muestran opiniones más fuertes que en el caso de las fortalezas mostrando que se autoevalúan más negativamente en las debilidades que positivamente en las fortalezas.

Tabla 5.6: Promedio de los scores de fortalezas y debilidades por género en autoevaluaciones

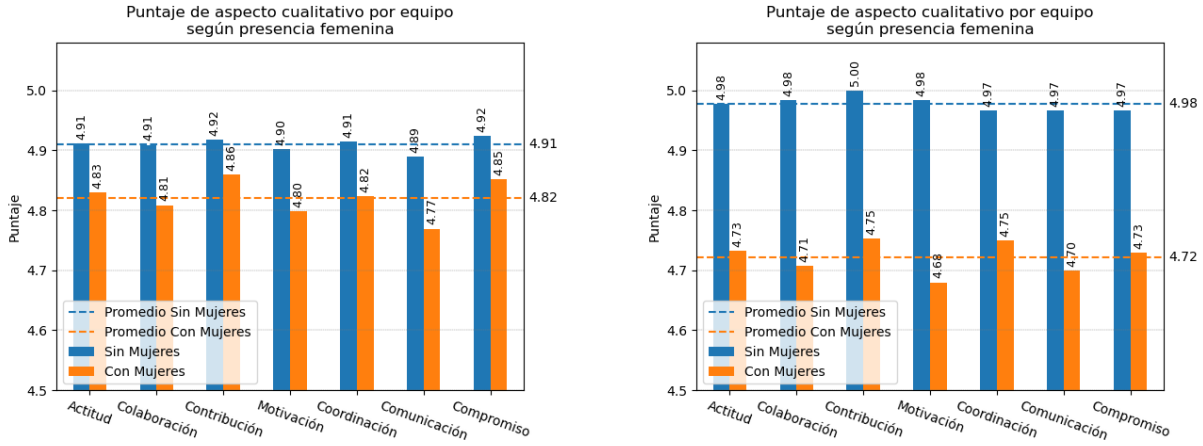
		Fortalezas	Debilidades
CC4401	**Hombres**	0.63	0.36
	Mujeres	0.77	0.83
CC5402	**Hombres**	0.52	0.62
	Mujeres	-	-

5.3. Promedio de los puntajes de aspectos cualitativos por equipo según presencia femenina

En esta sección se presentan los resultados relacionados a la segunda métrica tanto para las coevaluaciones y autoevaluaciones. Como se detalló en la sección 4.3.6, se calcula el promedio de los puntajes de los aspectos cualitativos separado por equipos según la presencia de mujeres en estos.

5.3.1. Coevaluación

En la figura 5.3 se muestran dos gráficos de barras que presentan el promedio de los puntajes de cada aspecto cualitativo de coevaluación de los equipos que tienen mujeres y los que no. Se debe recordar de la sección 4.3.6 que el puntaje por equipo en el caso de coevaluación es el promedio simple de los puntajes obtenidos por cada estudiante.



(a) CC4401 (b) CC5402

Figura 5.3: Gráficos de barras del promedio de los puntajes de aspectos cualitativos por equipo para la coevaluación, separados en los que tienen integrantes mujeres y los que no

Se puede notar que en ambos cursos en todos los puntajes de cada aspecto cualitativo de la coevaluación de los equipos sin mujeres se obtiene en promedio mejor puntaje que los equipos con mujeres. También se observa en línea punteada el promedio general de los 7 aspectos cualitativos. En CC4401 el promedio general de los equipos sin mujeres es mayor por 0.09 puntos y en CC5402 por 0.26 puntos. Una observación interesante es que a pesar

de que los equipos sin mujeres tienen mejor promedio en ambos cursos que los equipos con mujeres, la diferencia entre los promedios generales del curso CC4401 es notoriamente menor que en el curso CC5402.

Los datos de los gráficos se resumen en la tabla 5.7, en la que además se puede ver que las notas de las preguntas están estrechamente agrupadas alrededor del promedio, con coeficientes de variación bastante bajos. La mayor variación fue 0.64 % en los grupos con mujeres del curso CC5402, lo que indica que los promedios de cada tipo de grupo por curso son representativos.

Tabla 5.7: Promedio de los puntajes de cada aspecto cualitativo junto con el promedio, desviación estándar y coeficiente de variación de coevaluación por grupo, separados en los que tienen integrantes mujeres y los que no

		Actitud	**Colaboración**	**Contribución**	**Motivación**	**Coordinación**	**Comunicación**	**Compromiso**	**Promedio**	**Desviación Estándar**	**Coeficiente de Variación**
CC4401	**Sin Mujeres**	4.91	4.91	4.92	4.90	4.91	4.89	4.92	**4.91**	**0.01**	**0.20 %**
	Con Mujeres	4.83	4.81	4.86	4.80	4.82	4.77	4.85	**4.82**	**0.03**	**0.62 %**
CC5402	**Sin Mujeres**	4.98	4.98	5.00	4.98	4.97	4.97	4.97	**4.98**	**0.01**	**0.20 %**
	Con Mujeres	4.73	4.71	4.75	4.68	4.75	4.70	4.73	**4.72**	**0.03**	**0.64 %**

Dado el análisis anterior, se realizó un PERMANOVA para medir la significancia estadística de las diferencias. Los p-values resultantes se muestran a continuación en la tabla 5.8.

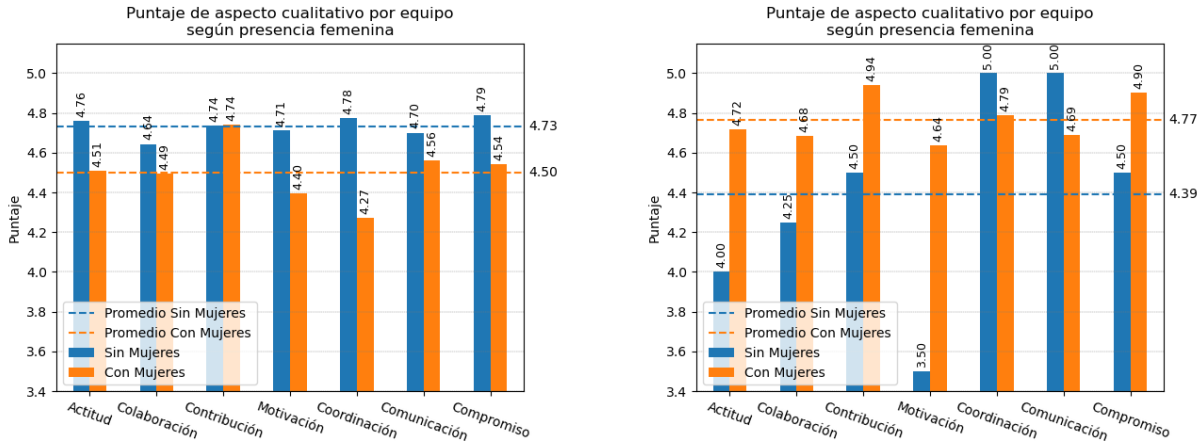
Tabla 5.8: P-values del PERMANOVA realizado entre los equipos con y sin mujeres de los cursos CC4401 y CC5402 para los aspectos cualitativos de las coevaluaciones

		CC4401	**CC5402**
		Con Mujeres	**Sin Mujeres**
CC4401	**Sin Mujeres**	0.316	0.733
CC5402	**Con Mujeres**	0.582	0.570

Con los resultados obtenidos se ve que la diferencia de los tipos de grupos de los cursos CC4401 y CC5402 no es significativa ya que todos los p-values son mucho mayores que el nivel de significancia $\alpha = 0.05$.

5.3.2. Autoevaluación

En la figura 5.4 se muestran gráficos de barras que presentan el promedio de los puntajes de cada aspecto cualitativo de autoevaluación de los grupos que tienen mujeres y los que no.



(a) CC4401 (b) CC5402

Figura 5.4: Gráficos de barras del promedio de puntajes por aspecto cualitativo por grupo para la autoevaluación, separados en los que tienen integrantes mujeres y los que no

Se observa que en el caso de CC4401 los equipos sin mujeres tienen mayor o igual puntaje en cada aspecto cualitativo, obteniendo un promedio general superior a los equipos con mujeres por 0.23 puntos. Por otro lado, en CC5402 los equipos con mujeres obtienen mejor promedio por aspecto cualitativo en todos expecto "coordinación" y "comunicación", teniendo en general un comportamiento invertido al otro curso. Ahora el promedio general de los equipos con mujeres supera al de los equipos sin mujeres por 0.38 puntos.

Los datos de los gráficos se muestran en la tabla 5.9. En ella se ve que los coeficientes de variación son más altos que en las coevaluaciones, teniendo como mayor variación 12.30 % en los grupos sin mujeres del curso CC5402.

Tabla 5.9: Promedio de los puntajes de cada aspecto cualitativo junto con el promedio, desviación estándar y coeficiente de variación de autoevaluación por grupo, separados en los que tienen integrantes mujeres y los que no

		Actitud	**Colaboración**	**Contribución**	**Motivación**	**Coordinación**	**Comunicación**	**Compromiso**	**Promedio**	**Desviación Estándar**	**Coeficiente de Variación**
CC4401	**Sin Mujeres**	4.76	4.64	4.74	4.71	4.78	4.70	4.79	**4.73**	**0.05**	**1.06 %**
	Con Mujeres	4.51	4.49	4.74	4.40	4.27	4.56	4.54	**4.50**	**0.15**	**3.33 %**
CC5402	**Sin Mujeres**	4.00	4.25	4.50	3.50	5.00	5.00	4.50	**4.39**	**0.54**	**12.30 %**
	Con Mujeres	4.72	4.68	4.94	4.64	4.79	4.69	4.90	**4.39**	**0.12**	**2.73 %**

Luego, se realizó un PERMANOVA para medir la significancia estadística de las diferencias usando los 7 aspectos cualitativos entre ambos tipos de grupo del mismo curso como también entre los distintos cursos. Los p-values resultantes se pueden ver en la tabla 5.10.

Tabla 5.10: P-values del PERMANOVA realizado entre los equipos con y sin mujeres de los cursos CC4401 y CC5402 para los aspectos cualitativos de las autoevaluaciones

		CC4401	**CC5402**
		Con Mujeres	**Sin Mujeres**
CC4401	**Sin Mujeres**	0.286	0.200
CC5402	**Con Mujeres**	0.481	0.142

Con los resultados obtenidos se ve que la diferencia no es significativa ya que todos los p-values son mucho mayores que el nivel de significancia $\alpha = 0.05$.

5.4. Promedio de los scores de fortalezas y debilidades por equipo según presencia femenina

En esta sección se presentan los resultados del análisis de sentimiento relacionado a la segunda métrica tanto para las coevaluaciones como para las autoevaluaciones. Como se detalló en la sección 4.3.6, se calculan de forma análoga al caso de los puntajes de aspectos cualitativos.

5.4.1. Coevaluación

En la tabla 5.11 se presentan los promedios de los scores de fortalezas y debilidades agrupados por equipos con mujeres y sin mujeres. De esta tabla se puede notar en primer lugar, que para las fortalezas en ambos cursos, los equipos con mujeres muestran opiniones más fuertes que los equipos sin mujeres. Esta diferencia se ve más acentuada en el curso CC5402 donde los equipos con mujeres en promedio tienen un score 0.27 más alto que los equipos sin mujeres. En CC4401 la diferencia entre ambos tipos de equipos es harto menor, pero con la diferencia que los scores de ambos tipos de equipos superan los de CC5402. De esta manera en CC5402 existe un brecha relativa más acentuada entre los grupos con mujeres y sin mujeres pero con sentimiento más débiles comparado con CC4401.

En cuanto a las debilidades se ve un patrón bastante diferente. En el caso del curso CC4401, se puede notar que el score de ambos tipos de equipos es notablemente menor, incluso llegando a un valor casi nulo en el caso de los equipos sin mujeres, lo que significa que en promedio las opiniones fueron más neutras. Por otro lado, los scores de CC5402 muestran de nuevo un brecha acentuada entre los equipos con mujeres y sin mujeres, pero de manera invertida al caso de las fortalezas; ahora los equipos sin mujeres muestran en promedio un score 0.42 puntos más alto que los equipos con mujeres.

Tabla 5.11: Promedio de los scores de fortalezas y debilidades por equipo agrupados según la presencia femenina a partir de los datos de coevaluaciones

		Fortalezas	Debilidades
CC4401	**Sin Mujeres**	0.78	0.01
	Con Mujeres	0.82	0.22
CC5402	**Sin Mujeres**	0.38	0.87
	Con Mujeres	0.65	0.45

5.4.2. Autoevaluación

En el caso de las autoevaluaciones, en la tabla 5.12 se muestran los promedios de los scores de los comentarios de fortalezas y debilidades de los alumnos según la presencia femenina en los equipos. En la tabla se puede ver que en el curso CC4401 los equipos con mujeres presentan mayor promedio de scores en las fortalezas con una diferencia de promedio de 0.11 puntos. En cambio, en el curso CC5402 es al revés, donde los grupos con mujeres tienen un menor promedio de scores en las fortalezas con una diferencia de 0.47 puntos, lo que indica que los grupos sin mujeres tienen opiniones más fuertes respecto a las fortalezas.

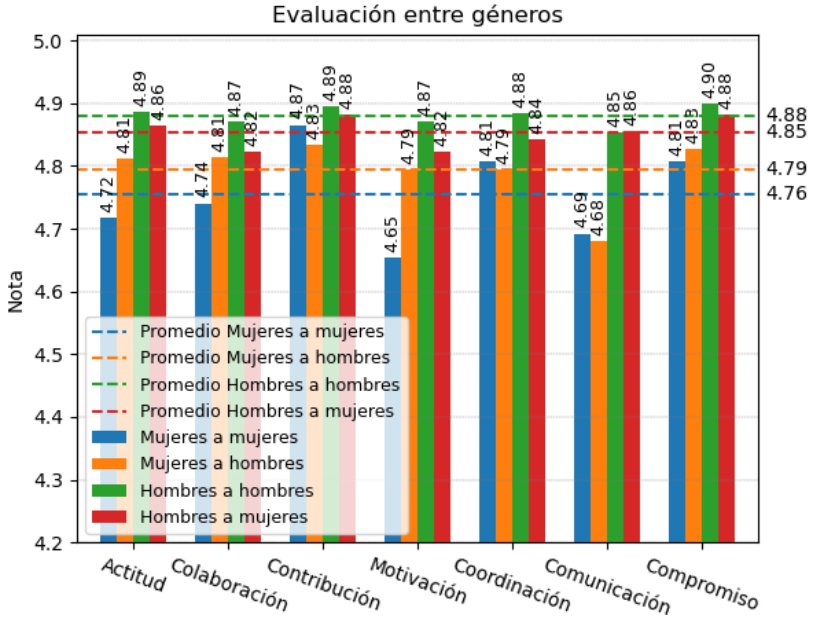
Para las debilidades, en ambos cursos los equipos con mujeres presentan promedios de scores más altos que los equipos sin mujeres, en el curso CC4401 con una diferencia de 0.05 puntos y en CC5402 con una diferencia de 0.26 puntos, lo que significa que los grupos con mujeres tienen opiniones más fuertes que los equipos sin mujeres respecto a las debilidades.

Tabla 5.12: Promedio de los scores de fortalezas y debilidades por equipo agrupados según la presencia femenina a partir de los datos de autoevaluaciones

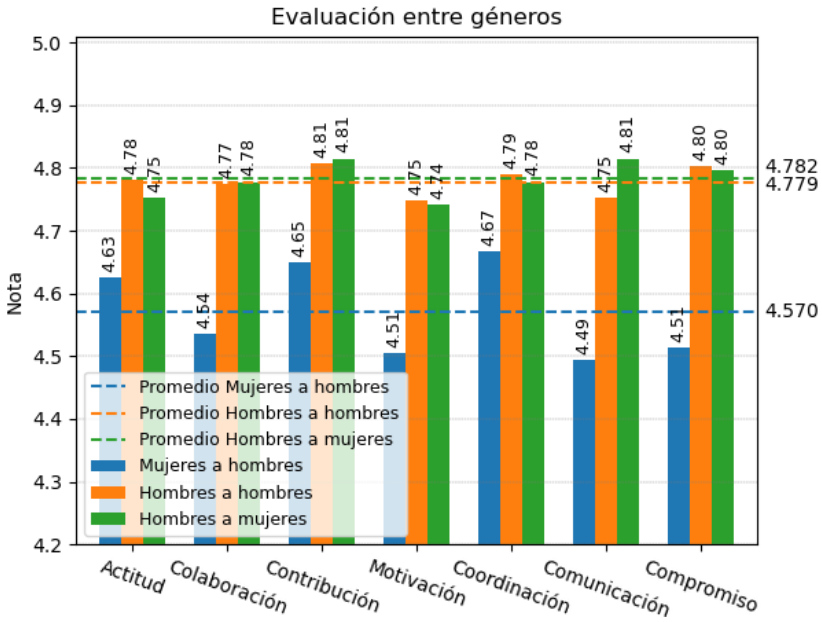
		Fortalezas	Debilidades
CC4401	**Sin Mujeres**	0.58	0.40
	Con Mujeres	0.69	0.45
CC5402	**Sin Mujeres**	0.83	0.44
	Con Mujeres	0.36	0.70

5.5. Promedio de los puntajes de aspectos cualitativos según coevaluación entre géneros

La figura 5.5 expone dos gráficos de barras que presentan el promedio de las coevaluaciones por pregunta entre género, es decir, cómo evalúa un género a su propio género y al opuesto. En el caso del curso CC5402 no hay evaluaciones entre mujeres, ya que no hay ningún grupo con más de una mujer (figura 4.6.b y tabla 4.7).



(a) CC4401



(b) CC5402

Figura 5.5: Gráficos de barras del promedio de los puntajes por aspecto cualitativo por grupo para la coevaluación, separados en los que tienen integrantes mujeres y los que no

Observando los promedio generales, en ambos cursos los hombres coevalúan en promedio mejor que las mujeres independiente del género al que coevalúen. Específicamente en CC4401, les asignan mejores notas a otros hombres que a mujeres con una diferencia de 0.03 puntos, mientras en CC5402 la diferencia es de 0.003. En CC4401, las mujeres coevalúan mejor a hombres que a su propio género con una diferencia de 0.03 puntos. También se destaca el hecho de que en CC4401 los hombres coevalúan mejor a otros hombres que mujeres a otras

mujeres, con una diferencia de 0.12, donde se presenta la mayor diferencia. Y por último se muestra en ambos cursos la tendencia de que las mujeres evalúan peor en general independiente del género.

En la tabla 5.13 se resumen los datos de los gráficos, mostrando además la desviación estándar y coeficiente de variación general entre los 7 aspectos cualitativos. Se puede observar que en general los coeficientes son relativamente bajos, mostrando la mayor variación para mujeres evaluando a hombres en CC5402 con un 1.53 %. De esta forma el promedio general de los puntajes de los aspectos cualitativos son representativos de cada tipo de grupo.

Tabla 5.13: Promedio de notas de cada aspecto cualitativo de cómo se coevalúan entre los distintos géneros

		Actitud	**Colaboración**	**Contribución**	**Motivación**	**Coordinación**	**Comunicación**	**Compromiso**	**Promedio**	**Desviación Estándar**	**Coeficiente de Variación**
CC4401	**Mujeres a mujeres**	4.72	4.74	4.87	4.65	4.81	4.69	4.81	**4.76**	**0.07**	**1.47 %**
	Mujeres a hombres	4.81	4.81	4.83	4.79	4.79	4.68	4.83	**4.79**	**0.05**	**1.04 %**
	Hombres a hombres	4.89	4.87	4.89	4.87	4.88	4.85	4.90	**4.88**	**0.02**	**0.41 %**
	Hombres a mujeres	4.86	4.82	4.88	4.82	4.84	4.86	4.88	**4.85**	**0.03**	**0.62 %**
CC5402	**Mujeres a hombres**	4.63	4.54	4.65	4.51	4.67	4.49	4.51	**4.57**	**0.07**	**1.53 %**
	Hombres a hombres	4.78	4.77	4.81	4.75	4.79	4.75	4.80	**4.779**	**0.02**	**0.42 %**
	Hombres a mujeres	4.75	4.78	4.81	4.74	4.78	4.81	4.80	**4.782**	**0.03**	**0.63 %**

En la tabla 5.14 se muestran los p-values de los PERMANOVAs realizados entre cada par de posibles grupos de géneros coevaluando a géneros. Se recuerda que CC5402 no posee grupos con más de una mujer por lo que no existe la relación de "mujeres a mujeres". En particular se nota que ninguno alcanza un nivel de signficancia del 5 %, pero se encuentra un valor relativamente bajo del 14.1 % al comparar la coevaluación de "mujeres a mujeres" con "hombres a hombres" de CC4401. Esto en realidad se condice con el gráfico 5.5.a ya que los promedios de estos grupos (que son representativos) son los más alejados.

Tabla 5.14: P-values del PERMANOVA realizado entre grupos de cierto género evaluando a un género para los aspectos cualitativos de coevaluación en cada curso

		CC4401	**CC5402**
Mujeres a mujeres	**Mujeres a hombres**	0.852	-
Mujeres a mujeres	**Hombres a hombres**	0.141	-
Mujeres a mujeres	**Hombres a mujeres**	0.508	-
Mujeres a hombres	**Hombres a hombres**	0.232	0.624
Mujeres a hombres	**Hombres a mujeres**	0.532	0.703
Hombres a hombres	**Hombres a mujeres**	0.714	0.963

Por otro lado, en la tabla 5.15 se observan los p-values al comparar los mismos tipos de grupos de un género evaluando a otro entre ambos cursos. De nuevo en ningún caso se cumple el nivel de significancia del 5 %, aunque para el caso de "mujeres a hombres" se ve un valor de 0.228, el más bajo. Este valor se puede considerar bajo mostrando mayor diferencia entre estos grupos.

Volviendo a la figura 5.5 y tabla 5.13 se puede observar que comparando el promedio de los puntajes del caso "mujeres a hombres" de ambos cursos el promedio varía 0.22 aproximadamente (desde 4.779 a 4.88), mientras que las desviaciones estándar son parecidas. Esta diferencia en los promedios se podría considerar relevante, por lo que el valor del p-value se podría deber más a la diferencia entre los promedios que en la variabilidad de la población.

Tabla 5.15: P-values del PERMANOVA realizados entre grupos de ciertos género evaluando a un género entre CC4401 y CC5402 para los aspectos cualitativos

	p-value
Mujeres a hombres	0.228
Hombres a hombres	0.285
Hombres a mujeres	0.709

Capítulo 6

Discusión

En esta sección se busca responder las tres preguntas de investigación basándose en el análisis de los resultados obtenidos, los patrones encontrados tanto en los puntajes de aspectos cualitativo como en los scores del análisis de sentimiento, y el análisis de significancia de las diferencias. Se debe tener en cuenta que estos resultados representan el semestre estudiado (otoño 2023), por lo que no son generalizables.

6.1. Autoevaluación de las mujeres

A partir del análisis de la figura 5.2 se destaca que en términos de promedios de aspectos cualitativos, las mujeres en general obtuvieron mejor nota mostrando que se autoevalúan mejor que los hombres. Sin embargo, no se puede concluir nada concreto, ya que las diferencias entre los géneros no son estadísticamente significativas. Esto puede atribuirse a dos factores principales. En primer lugar, se tienen pocos datos. Aunque el test PERMANOVA, al ser permutacional y no paramétrico, está diseñado para ser robusto frente a muestras limitadas, la escasez de datos de todas formas afecta la calidad de los resultados, debilitando la validez del análisis. Esta poca cantidad de datos con respecto a las autoevaluaciones se debe primordialmente al hecho de que es una actividad opcional debido a ser el primer semestre en implementar este tipo de evaluación. Y en segundo lugar, se cree que las autoevaluaciones no son respondidas a conciencia; la falta de significancia en las diferencias se podría atribuir, en parte, a la clara tendencia de los alumnos a asignarse la puntuación máxima en todas las preguntas, revelando una falta de variabilidad en las respuestas.

En cuanto a las respuestas de fortalezas y debilidades, en general las muejres tienen opiniones más fuertes que los hombres en ambos campos. Las mujeres se autoevalúan con opiniones más positivas en las fortalezas, pero al mismo tiempo se autoevalúan con opiniones más negativas en las debilidades. Esto podría ser un resultados interesante, pero no confiable por el hecho que este resultado es en base a la única autoevaluación recibida por el género femenino y sólo en el curso CC4401 (como se puede verificar en la tabla 4.5).

Concluyendo esta sección, la respuesta a la primera pregunta de investigación (**RQ1**) sobre cómo las mujeres se autoevalúan en comparación con los hombres no puede ser respondida con certeza. A pesar de que las cifras indican que, en promedio, las mujeres se autoevalúan mejor que los hombres en cuanto a los aspectos cualitativos, la falta de significancia estadística se-

gún el PERMANOVA sugiere que estas diferencias en los promedios no deben interpretarse como una realidad. Por otro lado, la poca cantidad de respuestas en las preguntas abiertas de autoevaluación por parte de las mujeres también deja ambigüedad en la confianza de los resultados. Por lo anterior, debido a la baja tasa de respuestas en las autoevaluaciones, las afirmaciones presentadas en esta sección no pueden ser consideradas concluyentes sin un estudio futuro que recopile datos adicionales.

6.2. Evaluación de equipos según la presencia femenina

En cuanto a los aspectos cualitativos se obtuvo que los integrantes de los equipos sin mujeres son mejor coevaluados en promedio en ambos cursos. La diferencia entre equipos con y sin mujeres varió bastante entre los cursos, pero siempre los equipos sin mujeres por encima de los equipos con mujeres. En el caso de CC4401 la diferencia de los promedio generales fue mucho menor que en CC5402. Por otro lado, a partir del análisis de las preguntas abiertas, se observó que en las coevaluaciones los equipos con mujeres resaltan las fortalezas en ambos cursos, destacando que en CC5402 la diferencia entre ambos tipos de equipos es bastante mayor comparado con CC4401. Se observa entonces que en cuanto a aspectos cualitativos los equipos sin mujeres se mantienen con puntajes mayores en ambos cursos, mientras que en cuanto a fortalezas, son los equipos con mujeres. Aun así se observa el mismo efecto de que la diferencia entre ambos tipos de equipos es menor en CC4401 comparado con CC5402.

Esto podría deberse a un carácter más crítico en el equipo ante la presencia de mujeres, ya que a pesar de mostrar opiniones más positivas en los comentarios de las fortalezas en los equipos con presencia feminina, estos no se sobrecalifican tanto en los aspectos cualitativos comparado con los equipos sin mujeres. En CC5402 los equipos sin mujeres llegan a un promedio general de 4.98 mostrando una clara sobrecalificación.

De todas formas, no se puede concluir con certeza este resultado ya que se observó que dentro de cada curso las diferencias no son estadísticamente significativas para ambos tipos de evaluaciones. De nuevo, esto puede deberse a las mismas razones anteriores. Probablemente los alumnos en general no responden a conciencia las preguntas de ambas evaluaciones, tienden a evaluar a sus compañeros y a sí mismos con el mayor puntaje posible.

Por otro lado, en el caso de las autoevaluaciones ocurre un efecto diferente. En el curso CC4401 los equipo sin mujeres superan en puntaje promedio a los equipos con mujeres, mientras que en el curso CC5402 ocurre lo contrario. En el análisis de las preguntas abiertas, se observa el mismo efecto de las coevaluaciones, los promedios de scores de fortalezas se comportan de manera contraria a los promedios de aspectos cualitativos, es decir que en CC4401 los equipos con mujeres tienen mejor score en fortalezas y en CC5402 los equipos sin mujeres tienen el mejor score.

Dado lo anterior, la respuesta a la segunda pregunta de investigación (**RQ2**) sobre las diferencias de los equipos con y sin mujeres en ambas evaluaciones tampoco se puede responder con certeza debido a que los resultados no son significativos, sin embargo en este

caso específico podemos decir que en las coevaluaciones los equipos sin mujeres son mejor evaluados que los con mujeres. En cuanto a las autoevaluaciones los patrones no son claros y podrían deberse a otro factor aparte de la presencia feminina en los equipos de trabajo.

6.3. Evaluación según la cantidad de mujeres en cada equipo

En cuánto al número de mujeres en los equipos, a pesar de que en ambos cursos los equipos sin mujeres obtuvieron mejores puntajes de aspectos cualitativos que los grupos con mujeres, se ve que en el curso CC4401 donde los equipos tienen dos mujeres hay una menor diferencia de promedios que en el curso CC5401 , donde los equipos tienen solamente una mujer (Figura 5.3). En el caso de la autoevaluación, los equipos sin mujeres se autoevalúan mejor que los equipos con mujeres en CC4401 mientras que en CC5402 es al revés.

De los análisis de significancia (tabla 5.8 y 5.10) fue posible observar que esta diferencias no son significativas. Entonces dado este resultado, se puede inferir que la cantidad de mujeres en los grupos con mujeres no cambia significativamente sus resultados, pero sí influye en general en el curso para aumentar o disminuir la diferencia de los promedios, como también invertir los resultados en el caso de las autoevaluaciones.

Este comportamiento de las diferencias de los promedios en las coevaluaciones puede indicar que al haber dos mujeres en el equipo hay mayor igualdad de trabajo. Como se puede verificar en la tabla 4.7, los equipos con mujeres en CC4401 tienen 2 mujeres, mientras que en CC5402 solo una. Con este dato se puede explicar que los promedios generales de los aspectos cualitativos en CC4401 son más cercanos entre sí, mostrando una menor diferencia, ya que al haber dos mujeres en los equipos con presencia femenina, pareciera haber mayor igualdad de trabajo. Caso contrario al curso CC5402 donde los equipos con mujeres contienen sola una mujer, la diferencia se acentúa notoriamente.

Esto también puede sugerir que las mujeres al no tener compañeras de su mismo género podrían sentir menos motivación y confianza para trabajar en equipo, por lo que tener dos mujeres o más aumentaría su comodidad y trabajarían de mejor manera. Es importante destacar que no existe un estudio de causalidad para poder afirmar esto con certeza, pero se cree que esta igualdad de trabajo es uno de los factores que explican los resultados.

Además, de los resultados de la figura 5.1 se ve que en el curso con una mujer por equipo (CC5402) se muestra un promedio general de los aspectos cualitativos de coevaluación de los hombres mucho menor que en el curso con dos mujeres por equipo (CC4401). Esto se condice con el p-value de la tabla 5.2, que aunque no es significativo al 5 % muestra el menor valor de la tabla. En cambio el promedio general de los aspectos cualitativos de coevaluación de las mujeres entre CC4401 y CC5402 es más cercano, 4.83 y 4.80 respectivamente.

Por otro lado, de los scores de fortalezas y debilidades tanto por género como por equipos agrupados según la presencia femenina,se puede observar que en promedio los hombres y las mujeres del CC4401 (2 mujeres por equipo) obtienen comentarios más positivos con

respecto a CC5402 (una mujer por equipo). Por otro lado, en CC4401 ambos géneros reciben debilidades con scores más cercano a 0, mostrando opiniones más neutras, mientras que en CC5402 las opiniones de debilidades son más negativas al tener valores cercanos 1. Esto se repite para el análisis según presencia femenina

Una posible teoría para explicar los resultados obtenidos tanto en aspectos cualitativos como en las respuestas a preguntas abiertas puede separarse en dos partes. Primero, en el curso CC4401, que incluye dos mujeres en cada equipo, se observa una mejor coevaluación en aspectos cualitativos en comparación con el curso CC5402, que solo incluye una mujer por equipo. Esto sugiere que la presencia de dos mujeres puede mejorar el ambiente de trabajo, reflejando una percepción general más positiva de todos los estudiantes. Esta mejora en la percepción también se refleja en los scores de las fortalezas, ya que para ambos géneros de CC4401 son más altos que los deCC5402, en otras palabras, los comentarios son más positivos. Lo mismo ocurre para los scores de fortalezas según la presencia feminina en los equipos. De la misma forma, en los scores de las debilidades de ambos géneros, en CC4401 son menores comparados con CC5402, por lo que los comentarios son menos negativos, incluso casi neutros.

En segundo lugar, la diferencia notable en los promedios de aspectos cualitativos entre los hombres de ambos cursos podría indicar que tener dos mujeres en el equipo no solo mejora la percepción del grupo hacia los hombres, sino que también influye positivamente en cómo las mujeres perciben a sus compañeros masculinos, por razones similares a las ya mencionadas.

Lo anterior se condice con lo que aparece en la figura 5.5. En ella se nota que los hombres de ambos cursos coevalúan sin mucha diferencia a otros hombres. Pero observando cómo evalúan las mujeres a los hombres, en CC4401 (grupos con 2 mujeres) las mujeres coevalúan en promedio a los hombres de mucha mejor manera con respecto a CC5402. Este resultado respalda con mayor certeza la teoría anterior; un equipo con dos mujeres mejora el ambiente de trabajo, pero desde la perspectiva de la mujer, mostrando esta alza en cómo son coevaluados los hombres por parte de las mujeres.

Finalmente, la respuesta a la tercera pregunta de investigación (**RQ3**) sobre si influye la cantidad de mujeres que hay en un equipo de trabajo con la forma en que se evalúan los estudiantes tampoco se puede generalizar con seguridad, ya que las diferencias no son significativas. Sin embargo, en este caso específico se puede ver que sí influye, comparando los cursos CC4401 y CC5402, en los que hay equipos sin mujeres, con una y con dos. Por esto se plantea la idea de que la presencia de dos o más mujeres en un equipo de trabajo ayuda a mejorar el ambiente de trabajo y la comodidad de los alumnos, especialmente de las alumnas, que lo componen.

Capítulo 7

Conclusiones

Esta tesis tiene como objetivo analizar el impacto de la presencia de mujeres en los equipos de trabajo del área de desarrollo de software. Esto debido a que dado el bajo número de mujeres que estudian y trabajan en campos de ingeniería y computación, existen diversos estereotipos que sugieren que las mujeres tienen menor rendimiento en estas áreas y que no son un aporte en los equipos de trabajo. Este aporte puede evaluarse considerando diversas dimensiones del trabajo en equipo, tales como actitud, compromiso, motivación, entre otros, los cuales pueden medirse con coevaluaciones y autoevaluaciones de cada miembro del equipo. Estas son evaluaciones formativas que estimulan el pensamiento crítico, fomentando la responsabilidad y autonomía de los alumnos en su aprendizaje.

Los objetivos planteados se lograron cumplir exitosamente, implementando la sección de autoevaluaciones correctamente en la plataforma del DCC y utilizándola en dos cursos del semestre Otoño 2023. Se obtuvieron los datos correctamente para su manipulación y análisis exploratorio, respaldado por gráficos y tablas, además de un análisis de significancia estadística de las diferencias usando el test de hipótesis PERMANOVA y un análisis de sentimiento para comentarios respecto a las fortalezas y debilidades de los alumnos.

Los resultados obtenidos indican que la presencia de mujeres en los equipos no genera diferencias estadísticamente significativas en la calidad del trabajo en equipo, según las percepciones evaluadas. Esto desafía los estereotipos iniciales planteados en la motivación del problema que sugiere que las diferencias de género podrían influir de manera significativa en la dinámica y rendimiento de los equipos de desarrollo de software. Sin embargo, es importante destacar que la presencia de más de una mujer en los equipos parece mejorar el ambiente de trabajo y la comodidad de los integrantes, sugiriendo que una mayor inclusión femenina podría contribuir positivamente al clima y la eficacia colaborativa. A pesar de no mostrar una diferencia significativa a favor de la presencia de las mujeres, se demuestra que esta no empeora la percepción del trabajo del equipo.

Este estudio enfrenta limitaciones asociadas al contexto y al método de investigación. Primero, la investigación es realizada en un ambiente académico, que puede no reflejar completamente las dinámicas de los entornos de trabajo en la industria. Además, aunque las herramientas de coevaluación y autoevaluación proporcionan datos valiosos sobre las percepciones del desempeño, estas pueden estar sujetas a sesgos personales y de grupo.

Se recomienda para futuras investigaciones ampliar el alcance de los resultados incluyendo datos de otros semestres y cursos. Aumentar el dataset podría revelar diferencias significativas y ofrecer una comprensión más completa de la influencia de las mujeres en los equipos de desarrollo de software. Además, se puede expandir al sector industrial para comparar las dinámicas de equipo y el impacto de la presencia femenina en entornos académicos versus corporativos. De igual forma, sería interesante investigar cómo las interacciones dentro de los equipos varían con diferentes proporciones de género y qué medidas pueden fomentar que se respondan a consciencia las coevaluaciones y autoevaluaciones en equipos de trabajo.

Así mismo, se propone buscar métricas adicionales de acuerdo a los datos disponibles, lo que podría proporcionar una comprensión más profunda de las dinámicas del trabajo en equipo. La inclusión de datos como las calificaciones de los alumnos en los proyectos puede ser clave para encontrar correlaciones y relaciones causales entre la percepción de los estudiantes y su desempeño académico.

Este estudio aporta nuevos conocimiento sobre el funcionamiento de los grupos de trabajo y el efecto de la presencia de mujeres en ellos, específicamente en la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, contribuyendo a la diversificación y transversalidad de la investigación en STEM. La singularidad de este enfoque es considerar las autoevaluaciones y coevaluaciones, proporcionando una noción de la percepción entre los estudiantes y la de ellos mismos.

Bibliografía

[1] D. Baker, "Where is gender and equity in science education?," Journal of Research in Science Teaching, vol. 39, no. 8, pp. 659–663, 2002.

[2] R. Noonan, "Women in STEM: 2017 Update. ESA Issue Brief# 06-17.," US Department of Commerce, 2017.

[3] F. Alessandri Cuevas and T. Villarroel Oyarzun, "Participación de Mujeres en STEM, la situación chilena Y comparación internacional," Acción Educar, Oct 2021.

[4] Facultad de Ciencias Físicas y Matemáticas Universidad de Chile, "Programa de Ingreso Prioritario de Equidad de Género (PEG)." https://uchile.cl/i94355, Dec 2022.

[5] J. Simmonds, M. C. Bastarrica, and N. Hitschfeld-Kahler, "Impact of Affirmative Action on Female Computer Science/Software Engineering Undergraduate Enrollment," IEEE Software, vol. 38, pp. 32–37, March 2021.

[6] Subsecretaría de Educación Superior, Ministerio de Educación, "Estadísticas por carrera." https://www.mifuturo.cl/buscador-de-estadisticas-por-carrera/, Oct 2023.

[7] A. Menezes and R. Prikladnicki, "Diversity in Software Engineering," in 2018 IEEE/ACM 11th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE), pp. 45–48, 2018.

[8] H. Blackburn, "The status of women in STEM in higher education: A review of the literature 2007–2017," Science & Technology Libraries, vol. 36, no. 3, pp. 235–273, 2017.

[9] N. Jimenez-Tenorio, L. Aragón, S. Sanchez-Rodriguez, C. Aragon, P. Azcárate, J. Cardeñoso, and F. Moreno, "La coevaluación/autoevaluación como instrumentos para valorar la competencia en el trabajo de equipo," in IV Jornadas de Innovación Docente. Abriendo caminos para la mejora educativa Sevilla, España: Universidad de Sevilla. Facultad de Ciencias de la Educación., May 2014.

[10] A. Rivera Picon, "La coevaluación como instrumento de la evaluación en el proceso de enseñanza-aprendizaje basado en el modelo por competencias en la educación superior," Diplomado en Docencia Universitaria, Mención en Ciencias Jurídicas y Políticas, Versión 15TA. Universidad Mayor de San Simón, Dec 2018.

[11] S. Fernández, "La autoevaluación como estrategia de aprendizaje," marcoELE. Revista de didáctica español lengua extranjera, no. 13, pp. 1–15, 2011.

[12] UNESCO, Cracking the code: girls' and women's education in science, technology, engineering and mathematics (STEM). Jan 2017.

[13] J. Kim and S. Celis, Women in STEM in Chilean Higher Education: Social movements and institutional transformations, pp. 105–120. Taylor and Francis Inc., Aug. 2021.

[14] A. Fausto-Sterling, "Women and science," Women's Studies International Quarterly, vol. 4, no. 1, pp. 41–50, 1981. Women in futures research.

[15] D. K. Pérez, ""¿No soy buena en esto o no soy buena en lo absoluto?"Brechas de género en la educación superior STEM," Tesis de maestría, Universidad de los Andes, Colombia, 2021.

[16] C. Rojas and M. C. Bastarrica, "Woman Performance in a Computer Science Program," in Proceedings of the 42 Conference of the Chilean Computer Science Society, (Concepción), Nov 2023.

[17] Facultad de Ciencias Físicas y Matemáticas, "Comisión local de Autoevaluación. Informe anual 2021," 2021.

[18] S. Mora-Rivera, M. Coto-Chotto, and J. Villalobos-Murillo, "Participación de las mujeres en la carrera de Ingeniería Informática de la Universidad Nacional y su desempeño en los cursos de programación," Revista Electrónica Educare, Costa Rica, 2017.

[19] S. J. Ceci, W. M. Williams, and S. M. Barnett, "Women's underrepresentation in science: sociocultural and biological considerations," Psychological bulletin, vol. 135, no. 2, 2009.

[20] J. S. Hyde, "The future of sex and gender in psychology: Five challenges to the gender binary," The American psychologist, vol. 74, no. 2, pp. 171–193, 2019.

[21] N. Bedregal-Alpaca, D. Tupacyupanqui-Jaén, and V. Cornejo-Aparicio, "Análisis del rendimiento académico de los estudiantes de Ingeniería de Sistemas, posibilidades de deserción y propuestas para su retención," Ingeniare. Revista chilena de ingeniería, vol. 28, pp. 668 – 683, Dec 2020.

[22] A. Murphy, B. Kelly, K. Bergmann, K. Khaletskyy, R. V. O'Connor, and P. M. Clarke, "Examining Unequal Gender Distribution in Software Engineering," in Systems, Software and Services Process Improvement (A. Walker, R. V. O'Connor, and R. Messnarz, eds.), (Cham), pp. 659–671, Springer International Publishing, 2019.

[23] M. Zukerfeld, G. Yansen, and N. Mura, "¿Por qué las mujeres no programan? Acerca de los vínculos entre Género, Tecnología y Software," IX Jornadas de Sociología. Facultad de Ciencias Sociales, Universidad de Buenos Aires, 2011.

[24] A. C. de Empresas de Tecnologías de Información AG, "Mujeres y tecnologías," Mar 2019.

[25] K. Davies, "Understanding the Gender Imbalance in STEM," STEM Women Whitepaper, 2021.

[26] M. R. J. Qureshi, S. A. Alshamat, and F. Sabir, "Significance of the teamwork in agile software engineering," Science International-Lahore, vol. 26, no. 1, pp. 117–120, 2014.

[27] J. Hayes, T. Lethbridge, and D. Port, "Evaluating individual contribution toward group software engineering projects," in 25th International Conference on Software Engineering, 2003. Proceedings., pp. 622– 627, June 2003.

[28] M. Davcheva and V. González-Romá, "Proportion of women in work teams and team performance: a moderated mediation model," Current Psychology, Aug 2022.

[29] M. Marques, "Software engineering education — Does gender matter in project results? — A Chilean case study," in 2015 IEEE Frontiers in Education Conference (FIE), pp. 1–8, 2015.

[30] J. J. Cruzado Saldana, "La evaluación formativa en la educación," Comuni@cción, vol. 13, pp. 149 – 160, April 2022.

[31] J. Delgado, N. Medina, and J. Fernández, "La evaluación por pares. Una alternativa de evaluación entre estudiantes universitarios," ReHuSo: Revista de Ciencias Humanísticas y Sociales. e-ISSN 2550-6587. URL: www.revistas.utm.edu.ec/index.php/Rehuso, vol. 5, pp. 14–26, May 2020.

[32] F. Cruz Núñez and A. Quiñones Urquijo, "Importancia de la evaluación y autoevaluación en el rendimiento académico," Zona Próxima, 2012.

[33] S. Luttenberger, M. Paechter, and B. Ertl, "Self-Concept and Support Experienced in School as Key Variables for the Motivation of Women Enrolled in STEM Subjects With a Low and Moderate Proportion of Females," Frontiers in Psychology, vol. 10, 2019.

[34] C. Schindler and M. Müller, "Gender Gap? A Snapshot of a Bachelor Computer Science Course at Graz University of Technology," in ECSA '19 - Proceedings of the 13th European Conference on Software Architecture, vol. 2, (United States), pp. 100–104, Association of Computing Machinery, Sept. 2019.

[35] D. R. D. L. Giannina Costa, Juan Felipe Calderón and L. P. S. M. Medina, "Relationship between guided interactive activities and self-concept in engineering students," in 2021 ASEE Virtual Annual Conference Content Access, (Virtual Conference), ASEE Conferences, July 2021. https://peer.asee.org/37656.

[36] D. Wackerly, W. Mendenhall, W. Mendenhall, and R. Scheaffer, "Estadística Matemática con Aplicaciones, Séptima Edición," ch. 10, 2009.

[37] M. Anderson, "A new method for non-parametric multivariate analysis of variance," Austral Ecology, vol. 26, pp. 32 – 46, 02 2001.

[38] IBM, "¿Qué es el análisis de sentimiento?," 2024.

[39] A. K. Dwivedi and M. Dwivedi, "A study on the role of machine learning in natural language processing," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp. 192–198, 07 2022.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 10 2018.

[41] R. Riquelme Muñoz, "Sistema de evaluación del desempeño de los miembros de un equipo de desarrollo de software," Trabajo de título, Universidad de Chile, 2014.

[42] F. Tobar, "Estadística - Notas de Clase MA3402," pp. 75–77, 2022.

[43] R. A. Johnson and D. W. Wichern, Applied Multivariate Statistical Analysis. USA: Prentice-Hall, Inc., 1988.

Anexos

Anexo A. Ejemplo de Test de Hipótesis

Para mayor entendimiento, se usará el ejemplo presentado en [42]. En este escenario hipotético se supone que los recién nacidos (RN) en Santiago, Chile, siguen una distribución normal con una media de 3000 gramos y una desviación estándar de 500 gramos. Se cree que los RNs en Osorno pesan, en promedio, más que los RNs en Santiago. En este contexto, se formulan las siguientes hipótesis:

- Hipótesis alternativa H_1: Los RNs en Osorno pesan en promedio más de 3000 gramos.
- Hipótesis nula H_0: Los RNs en Osorno pesan en promedio 3000 gramos.

En [42] se expone una situación interesante. Se asume que se tomó una muestra de 50 RNs en Osorno, y se ha obtenido un peso promedio de 3200 gramos. ¿Se puede concluir con certeza que los RNs en Osorno pesan más que los de Santiago? No necesariamente, ya que es posible que esta muestra de 50 RNs con un peso promedio de 3200 gramos provenga de una población de RNs distribuidos según $\mathcal{N}(\mu = 3000, \sigma^2 = 500^2)$.

Para respaldar la validez del resultado anterior, es necesario determinar cuál de las dos hipótesis es verdadera. En particular se quiere comprobar H_1, y esto se logra demostrando que H_0 es falsa. Para ello, es necesario calcular el **p-value**, el cual corresponde a la probablidad de obtener el resultado observado (es decir, un promedio de 50 RNs de 3200 gramos) condicionado a que la hipótesis nula es cierta (es decir, que la muestra de los 50 RNs proviene de una distribución normal $\mathcal{N}(3000, 500^2)$).

Matemáticamente, si se asume que H_0 es cierta, la muestra de 50 RNs se expresa como:

$$X_1, \ldots, X_{50} \sim \mathcal{N}(3000, 500^2)$$

Usando propiedades de variables aleatorias gaussianas, la media de esta muestra se define como $\bar{X} = \frac{1}{50}\sum_{i=1}^{50} X_i \sim \mathcal{N}(3000, 500^2/50)$. Con esto, el p-value se define como:

$$\text{p-value} = \mathbb{P}(\underbrace{\frac{1}{50}\sum_{i=1}^{50} X_i = \bar{X} \geq 3200}_{\substack{\text{Resultado observado bajo la} \\ \text{hipótesis alternativa}}} \mid \overbrace{X_i \sim \mathcal{N}(3000, 500^2)\, i = 1, \ldots, 50}^{\text{Hipótesis nula es cierta}})$$

Usando técnicas de probabilidades como el *z-test* se puede calcular esta probabilidad y corresponde a p-value = 0.00233. Si se elige un un nivel de significancia $\alpha = 0.01$, se cumpliría que se rechaza H_0 al 1 %.

Anexo B. *Analysis of Variance* (ANOVA)

Basado en [43], el análisis de varianza (ANOVA) es un test de hipótesis paramétrico univariado que busca comparar g poblaciones o grupos para investigar si las medias poblacionales son iguales, y si no, encontrar cuales medias difieren significativamente.

Se asumen dos cosas:

1. $X_{l1}, \ldots, X_{ln_l}$ es una muestra aleatoria de la población l de tamaño n_l de una población con media $\mu_l \in \mathbb{R}$. Las muestras de diferentes poblaciones son independientes.
2. Todas las poblaciones tienen la misma varianza σ^2, en otras palabras, $X_{lj} \sim \mathcal{N}(\mu_l, \sigma^2)\ \forall\, l, j$

Ej: Un curso de alumnos separado en dos poblaciones: hombres y mujeres.

De esta manera, aunque la hipótesis nula de igualdad de medias poblacionales se puede formular como $\mu_1 = \mu_2 = \cdots = \mu_g$, se acostumbra considerar μ_l como la suma de un componente de media general designado como μ y un componente debido a la población específica τ_l conocido como *efecto de tratamiento*. En el ejemplo anterior el tratamiento sería el género de los alumnos. Se puede reescribir $\mu_l = \mu + (\mu_l - \mu) = \mu + \tau_l$ donde $\tau_l = \mu_l - \mu$. Asi, la reparametrización es:

$$\underbrace{\mu_l}_{\substack{l\text{-ésima media}\\ \text{poblacional}}} = \underbrace{\mu}_{\substack{\text{media}\\ \text{general}}} + \underbrace{\tau_l}_{\substack{l\text{-ésimo efecto}\\ \text{de tratamiento}}}$$

Esta reparametrización lleva a la reformulación de la hipótesis nula H_0:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = 0$$

Con esto es posible reparametrizar la muestra $X_{lj} \sim \mathcal{N}(\mu + \tau_l, \sigma^2)$ como:

$$X_{lj} = \underbrace{\mu}_{\substack{\text{media}\\ \text{general}}} + \underbrace{\tau_l}_{\substack{l\text{-ésimo efecto}\\ \text{de tratamiento}}} + \underbrace{\epsilon_{lj}}_{\substack{\text{error}\\ \text{aleatorio}}} \tag{B.1}$$

donde $\epsilon_{lj} \sim \mathcal{N}(0, \sigma^2)$ son las variables aleatorias independientes. Para definir de forma única los parámetros del modelo y sus estimaciones de mínimos cuadrados, se impone la restricción $\sum_{l=1}^{g} n_l \tau_l = 0$.

De la reparametrización (B.1), el análisis de varianza está basado en un descomposición análoga de las observaciones:

$$\underbrace{x_{lj}}_{\text{observación}} = \underbrace{\bar{x}}_{\substack{\text{media muestral}\\ \text{general}}} + \underbrace{(\bar{x}_l - \bar{x})}_{\substack{\text{efecto de tratamiento}\\ \text{estimado}}} + \underbrace{(x_{lj} - \bar{x}_l)}_{\text{residuo}} \tag{B.2}$$

donde $\bar{x}$ es una estimación de μ, $\hat{\tau} = (\bar{x}_l - \bar{x})$ una estimación de τ_l y $(x_{lj} - \bar{x}_l)$ una estimación del error ϵ_{lj}. De esto, se puede notar que para las estimación $\hat{\tau}_l$ siempre se satisface que $\sum_{l=1}^{g} n_l \hat{\tau}_l = 0$ y bajo H_0, cada $\hat{\tau}_l$ es una estimación de cero.

Finalmente, el ANOVA consiste en responder la pregunta de la igualdad entre las medias **al evaluar si las contribuciones de los efectos de tratamiento es grande relativo a los residuos**. Esto se logra reformulando las ecuaciones B.1 y B.2.

Para lograr lo anterior, de la ecuación (B.2) se puede llegar a la siguiente expresión de sumas de cuadrados (SS):

$$\underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj}-\bar{x})^2}_{SS_{cor}} = \underbrace{\sum_{l=1}^{g} n_l(\bar{x}_l-\bar{x})^2}_{SS_{tr}} + \underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj}-\bar{x}_l)^2}_{SS_{res}} \tag{B.3}$$

donde SS_{cor} es la suma de cuadrados total corregido, SS_{tr} es la suma de cuadrados **entre muestras de distintos grupo o tratamiento** y SS_{res} sería la suma de cuadrados **entre muestras del mismo grupo o tratamiento**. En otras palabras SS_{tr} y SS_{res} se relacionan con la varianza, donde el primero mide **la variación debido al tratamiento**, y el segundo **la variación dentro de cada tratamiento**.

Finalmente se usa un F-test para rechazar H_0 a un nivel de significancia α si:

$$F = \frac{SS_{tr}/(g-1)}{SS_{res}/\left(\sum_{l=1}^{g} n_l - g\right)} > F_{g-1,\sum n_l - g}(\alpha)$$

donde $F_{g-1,\sum nl-g}(\alpha)$ es el 100α-ésimo percentil de la distribución F con $(g-1)$ y $(\sum n_l - g)$ grados de libertad.

Anexo C. *Multivariate Analysis of Variance* (MANOVA)

El análisis de varianza multivariado (MANOVA) es un extensión multivariada del ANOVA (Anexo B). En este caso,

$$X_{lj} = \mu + \tau_l + \epsilon_{lj},\ j = 1, 2, \ldots, n_l \ y\ l = 1, 2, \ldots, g \tag{C.1}$$

donde ahora $X_{lj} \sim \mathcal{N}(\mu, \Sigma)$ con $\mu \in \mathbb{R}^n$ y $\Sigma \in \mathbb{R}^{n\times n}$ (el resto de los vectores tienen las dimensiones adecuadas).

Se puede notar que para el modelo en (C.1) cada componente del vector de observación X_{lj} satisface el modelo univariado (B.1).

El desarrollo es completamente análogo al caso del ANOVA y se llega a:

$$\underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj}-\bar{x})(x_{lj}-\bar{x})^t}_{SS_{cor}} = \underbrace{\sum_{l=1}^{g} n_l(\bar{x}_l-\bar{x})(\bar{x}_l-\bar{x})^t}_{SS_{tr}} + \underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj}-\bar{x}_l)(x_{lj}-\bar{x}_l)^t}_{SS_{res}} \tag{C.2}$$

donde SS_{cor} es la suma de cuadrados total corregido, SS_{tr} es la suma de cuadrados **entre muestras de distintos grupo o tratamiento** y SS_{res} sería la suma de cuadrados **entre**

muestras del mismo grupo o tratamiento. La diferencia con ANOVA es que en este caso las sumas de cuadrados son matrices y consideran los productos cruzados entre las componentes.

En este caso el el test de hipótesis que se realiza es el de *Wilks' lambda* test (tabla 6.3 de [43]), el cual usa el ratio de varianzas generalizado:

$$\Lambda^* = \frac{|SS_{res}|}{|SS_{tr} + SS_{res}|} \tag{C.3}$$

donde $|*|$ corresponde al determinante de las matrices.