UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Computación

Análisis exploratorio y estadístico sobre el alumnado de Ciencias de la Computación en FCEN-UBA

Tesis de Licenciatura en Ciencias de la Computación

Daniel Paz 407/08
Nicolas Varaschin 187/08

Directores de tesis: Dr. Carlos 'Greg' Diuk y Dr. Diego Fernández Slezak.

Resumen

Actualmente en la Argentina existen diversas carreras enfocadas en la tecnología, informática y computación. La cantidad de profesionales informáticos que forman dichas carreras no alcanza a satisfacer la demanda del mercado laboral. Dada la gran demanda actual en temas computacionales, conocer nuevos datos acerca de los alumnos actuales y graduados de la carrera de Ciencias de la Computación de FCEN puede aportar información valiosa sobre la problemática general. Otro problema de importancia es la baja proporción de mujeres en el alumnado. En la actualidad, saber las causas de por qué las mujeres no eligen una carrera de computación es de interés para muchos centros de estudios del mundo entero.

En el presente trabajo recolectamos los datos necesarios, los analizamos estadísticamente y obtenemos conclusiones para tratar el problema de descubrir los factores que afectan al rendimiento académico y a la deserción de los alumnos. Dentro del análisis pretendemos poder predecir si, dado un alumno, éste presenta tendencia a tener un rendimiento académico bajo o alto. Asimismo, poder analizar los factores que indican la regularidad de un alumno o futura falta de ella tanto si abandona la carrera como si se gradúa. Dentro de la investigación esperamos poder generar un perfil de los alumnos actuales del departamento y sus graduados. Dimos especial énfasis a determinar los factores de interés de las mujeres sobre nuestra carrera.

Los resultados obtenidos proveen información útil sobre preguntas clásicas y creencias típicas del alumnado. Se encontró evidencia que indica que no son necesarios conocimientos previos de programación antes de ingresar a la carrera para tener éxito en la misma. Los datos reafirman que el hecho de realizar los ejercicios de las guías prácticas influye positivamente en los alumnos logrando que un porcentaje menor de éstos postergue la carrera. En nuestra carrera se suele pensar que el examen final de la materia Análisis II se posterga hacia el final de la misma. Sin embargo, en otro análisis realizado en este trabajo se observó qué, en promedio, dicho examen se rinde en la fecha acorde al plan de estudios. A pesar de esto, los alumnos suelen postergar los examenes finales de materias como Métodos Numéricos, Organización del Computador II e Ingeniería del Software II. Los datos ponen en descubierto que las fechas acordadas por el plan de estudios son difíciles de cumplir por la mayoría del alumnado. En este sentido, se analizaron cuáles materias son las más postergadas.

El contexto de la tesis se enmarca en la obtención y procesamiento de datos del alumnado. Dichos datos se obtuvieron mediante encuestas diseñadas por nosotros y a través de datos provistos por el Departamento de Alumnos de la Facultad. Los grupos a analizar comprenden tanto a los alumnos regulares como a los graduados. Utilizando los datos, se construyó un método predictivo para poder inferir si un alumno es propenso a postergar o no la carrera. Se consideraron modelos de Regresión Generalizada y modelos de clasificacion propios del área de Machine Learning (Random Forests y Gradient Boosting). Dentro de los predictores más influyentes del clasificador encontramos el hecho de utilizar Internet para el estudio, las horas que trabaja el alumno, si consideró dejar la carrera alguna vez, si mantuvo un mismo grupo de trabajos prácticos durante varias materias y si tuvo o no alguna beca. Analizando a los alumnos que quedaron libres, se descubrió que más de la mitad abandona la carrera sin aprobar ningún final y, en segunda instancia, se abandona luego de rendir Análisis II o

Álgebra (primeras dos materias del plan de estudios).

Complementando el trabajo, se realizó un estudio sobre el grupo de alumnas de nuestra carrera. El objetivo fue sumar información a la tendencia en estudios acerca de mujeres en el sector informático. Se recolectaron algunas características que son elegidas por la mayoría de las alumnas para elegir nuestra carrera. Se destacan el interés desde el secundario por las matemáticas y la computación. En cuanto al rendimiento académico, el mayor factor que afecta a éste grupo en particular es el hecho de trabajar y estudiar a la vez.

Palabras clave: Análisis exploratorio de datos, Inferencia estadística, Machine learning, Random Forests, Gradient Boosting

Abstract

Currently in Argentina there are multiple bachelor and master degrees focused in technology, informatics and computer science. The amount of professionals graduated from those schools are not enough to satisfy the market's demand. Given the current big demand in computational orientations, it is relevant to gather and study new data about the undergraduate and graduate students from the Computer Science Department at FCEN-UBA. Another issue of importance relies on the low proportion of female students within the student body. Knowing the causes that drive the female students not to choose a degree in Computer Science, is of interest to many study centers worldwide.

The following work gathers the necessary data, performs statistical analyses and acquires new insights to deal with the problem of learning which factors impact the academic performance and desertion. With our analysis we pretend to be able to predict if given a student, he or she shows an inclination of having a good or bad academic performance. Furthermore, we will analyze which features suggest the regularity of a student, or a future lack of regularity given that the student abandoned or finished the degree. With our research we hope to produce a profile for the current and graduated students. We will be giving special attention to determine the factors of interest for women in our degree.

Results obtained provide feasible information about common questions and common beliefs that students have. Evidence was found that supports that it's not necessary to be successful in this particular career no previous programming knowledge is necessary . Data reaffirms that completing the optional exercises and handouts provided by each course (as it's always suggested) is very influential over students, achieving a lower percentage of them to delay the obtention of their degree. It's a common belief of students from our career that the final exam of Calculus II it's delayed until near the end of the career. However, another analysis performed we observed that this final exam is , in average, taken on the planned date stupilated by the career syllabus. Nevertheless, final exams from courses like Numerical Methods, Computer Organization II and Software Engineering II are usually delayed by students. Data obtained shed light that dates proposed by the carrer syllabus are hard to accomplish for the majority of the student body. Following this line of thought, we analysed which courses are the most delayed by students.

The context of our thesis focuses on obtaining and processing of students data. This information was obtained from surveys and data provided by the students department. The groups to be analyzed were composed by the general student body and graduated students. Using this data, we built a predictive model to be able to infer if a student is likely to delay the career or not. We considered Generalized Regression models and other models belonging to the area of machine learning (Random Forests and Gradient Boosting). Within the most influential predictors we found the fact of student using internet to complement his/her study, if he/she kept a common study group throughout several courses and if he/she has a scholarship. Analyzing the dropout students, we discovered that more than half of them give up the career without passing any final exam and, secondly, the students drop out after passing Caluculus II or Algebra (first two subjects of the career syllabus).

To complement our work, we performed a study over the female students of our career. The objective was to add information to the already existing studies about women in computer science. Some features that are chosen by

the majority of the female students at the moment to choose our career were gathered. We highlight the interest for mathematics and computer subjects in high school. Regarding the academic performance, the biggest factor that affects this group in particular, is the fact of working and studying at the same time.

Keywords: Exploratory Data Analysis, Statistical Inference, Machine Learning, Random Forests, Gradient Boosting

Agradecimientos

A mi familia por sobre todo, mis viejos, mi hermano y mi abuela. A los amigos y compañeros que conocí en la facultad y de los cuales aprendí mucho. A mis amigos de la vida, por estar en todo momento. A ambos GARG, dos grandes grupos que me acompañaron estos años. A nuestros directores y jurados por su total apoyo para poder realizar esta tesis. Al Departamento de Computación y la Universidad por formarme profesionalmente y facilitar los datos y herramientas que hicieron posible este trabajo.

- Nico

A mi familia principalmente, que siempre fue un apoyo en casa en momentos de alegría, miedos y tensiones. A mis amigos de toda la vida por estar siempre, prestar una oreja y palabras de aliento siempre. A mis compañeros de facultad que me acompañaron a lo largo de la carrera, con los que pude avanzar a la par, aprender juntos y que fácilmente se convirtieron en un grupo de amigos que hoy sigue más unido que nunca. A nuestros directores y jurados por sugerirnos y acompañarnos a lo largo de la tesis. Al Departamento de Computacion por formarme y con quien muchos quedaremos en deuda siempre. Gracias...totales!

- Daniel

Índice

1. Introducción

El área de Data Science se define como un enfoque estructurado para la extracción de conocimiento y hechos sobre un cuerpo de datos determinado. Abarca el estudio de los datos de un punto de vista estadístico y computacional.

Hoy en día, es un área muy solicitada, con varias aplicaciones reales en los campos de las ciencias biológicas, informática médica, ciencias sociales y humanidades como también economía, finanzas y negocios. La aparición de grandes conjuntos de datos, y el desarrollo del poder de cómputo para poder procesarlos, hacen que los métodos de Data Science sean herramientas muy útiles para la industría y la investigación.

El objetivo de esta tesis es lograr construir un perfil de los alumnos del Departamento de Computación de la Facultad de Ciencias Exactas y Naturales que nos permita entender el rendimiento académico de cada uno, para luego poder detectar y tomar acciones de forma más eficiente en las variables que puedan afectarlo. ¿Será acaso posible determinar (cuantitativamente) cuánto afectan las largas jornadas laborales (más de siete horas) en el rendimiento académico del alumnado?. Para responder esta y otras preguntas que puedan surgir, se pretende analizar a los actuales alumnos de nuestra facultad, mediante la realización de un estudio estadístico. A través de este estudio, observaremos qué rasgos y características distinguen a nuestros alumnos. Separaremos a los alumnos en diferentes grupos teniendo en cuenta los valores de ciertas variables, analizando cómo se comporta el resto de las mismas al realizar estas particiones. En este marco, nos proponemos realizar un análisis exploratorio para cada grupo, así como también utilizar diversas técnicas estadísticas y otras herramientas predictivas dentro del marco de Data Science.

Se analizaron factores que afectan de manera significativa al rendimiento o éxito académico, a fines de mejorar la calidad de estudio en nuestra facultad. Diseñamos y utilizamos una definición propia que determina cuando un alumno está postergando o no la carrera para elaborar un modelo que sirva de base de nuestro análisis. Utilizando este modelo construímos predictores que indican si dado un alumno, éste está postergando la carrera o no. Con esta información, pudimos analizar qué factores en concreto afectan de forma directa en que un alumno postergue la carrera.

Por otro lado, a lo largo de la historia, la presencia femenina en las carreras relacionadas a la computación se ha ido disminuyendo. A tal punto de que antiguamente la presencia de hombres y mujeres estaba casi dividida en un cincuenta porciento para cada género, algo impensado hoy en día. La brecha se ha ido incrementando con el paso de los años a tal punto que hoy en día, muchas mujeres escogen no estudiar estas carreras porque piensan que no cuentan con la habilidad que se requiere para tener éxito en las mismas. Sin embargo, este es sólo uno de los factores. ¿Será posible detectar otros factores significativos que influyan en éste fenómeno? ¿Cómo podemos mitigar esta tendencia? Nos propusimos realizar un análisis específico para alumnas de nuestra carrera para conocer y entender mejor los factores que tienen como consecuencia la escasez de mujeres en computación. Si bien nuestro análisis se limitó a las alumnas de nuestra carrera, como mencionamos anteriormente, este problema es global en el área informática.

Para realizar nuestro estudio, recolectamos y utilizamos datos de los alumnos de diferentes fuentes. Entre ellas, se encuentran dos encuestas que realizamos al

alumnado: La primera enfocada en alumnos egresados y no egresados de la carrera sin distinción de género. La segunda, enfocada exclusivamente en el alumnado de género femenino. También contamos con la colaboración del Departamento de Alumnos de la Facultad, el cual nos proveyó de datos anónimos del alumnado. Estos datos comprenden la condición de regularidad de los alumnos y su desempeño en algunas materias importantes de la carrera.

La cantidad de datos suministrados por el Departamento de Alumnos fueron suficientes para generar análisis más interesantes, incluyendo clasificadores predictivos sobre alguna de las variables. Estos datos contemplan el comportamiento de los alumnos frente a un subconjunto elegido de materias y también la condición de regularidad, indicando cuales características son importantes para determinar si un alumno está más cercano a recibirse o a abandonar la carrera.

1.1. Análisis exploratorio de datos

El análisis exploratorio de datos, es un enfoque estadístico que engloba un conjunto de herramientas para el análisis de conjuntos de datos. Principalmente, se utiliza para tener un panorama general y descriptivo del comportamiento y estructura de los datos. La exploración sirve de base para crear hipótesis sobre los datos y obtener premisas sobre las cuales construir la inferencia estadística.

Las técnicas del análisis exploratorio son mayormente observacionales. Teniendo en cuenta fuentes como gráficos, tablas y correlacciones entre variables, se pueden descubrir hechos y plantear conjeturas sobre los datos.

Dentro de las herramientas de estadística descriptiva utilizadas se encuentran estilos de gráficos como boxplot o histogramas, análisis de correlacción o distribución, o análisis de componentes principales. Estos métodos buscan obtener una representación de las variables que resuman los datos y que aceleren el proceso de estudio de los mismos. Conocer los resultados de estos análisis y representaciones, revelan la información más básica sobre los datos en cuestión.

En un paso posterior al proceso exploratorio, se puede realizar inferencia estadística. La inferencia se compone de técnicas para obtener conclusiones acerca del modelo de los datos, y cómo deberían comportarse posibles nuevos datos de la misma población.

Resulta interesante comparar variables entre sí desde diferentes puntos de vista, así como también tener una noción de las diferencias entre dos grupos que se conforman al fijar una de éstas variables. Para comprobar si dicha comparación es significativa, se utilizan tests estadísticos que se construyen sobre los resultados. Los tests estadísticos trabajan sobre ciertas hipótesis sobre los datos, las cuales son puestas a prueba comparando *p-values* obtenidos de los tests.

1.2. Tests estadísticos utilizados

A lo largo de esta tesis utilizaremos diferentes tipos de tests estadísticos para poner a prueba algunas hipótesis que surgirán a lo largo de la misma. Dos de los tests estadísticos clásicos del área son el test de Student y el test de Kolmogorov-Smirnov.

1.2.1. Test de Student

La hipótesis nula del test de Student es: 'Las medias de ambos grupos son iguales'. El test de Student (llamado *t-test*) analiza si la media de dos grupos son estadísticamente diferentes entre sí. Las diferencias entre las medias nos interesan sólo si son significativas. Existe una diferencia significativa cuando la probabilidad de que las medias de cada grupo sean diferentes por pura casualidad es baja.

El hecho de que la diferencia sea significativa también es afectada por la varianza de cada grupo. Si los grupos tienen baja varianza, aumenta la probabilidad de que la diferencia encontrada sea real y no causada por el azar. En cambio, si existe gran varianza dentro de cada grupo, es más probable que el resultado obtenido haya sido por casualidad.

En la figura 1 se puede ver un ejemplo.

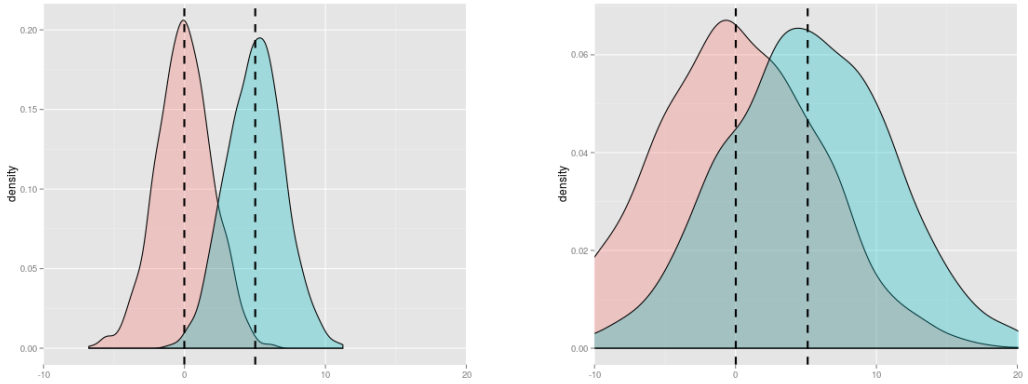


Figura 1: Dos distribuciones normales A y B con $\mu_A = 0$ y $\mu_B = 5$. A la izquierda las varianzas son $S_A^2 = S_B^2 = 2$ y a la derecha $S_A^2 = S_B^2 = 6$.

El *t-test* entre dos muestras X_A y X_B se define como:

$$t = \frac{\mu_A - \mu_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

Donde S^2 refiere a la varianza de cada muestra y n al tamaño. El test devuelve un valor t que puede ser comparado contra la distribución de Student y así, obtener un *p-value*.

1.2.2. Test de Kolmogorov-Smirnov

El test de Kolmogorov-Smirnov es un test no paramétrico, lo que significa que no necesita prerrequisitos en cuanto al tipo de distribución, medias o varianza de los grupos. La hipótesis nula del test de Kolmogorov-Smirnov es: 'Las distribuciones de ambos grupos son iguales'

Similarmente al test de Student, Kolmogorov-Smirnov compara dos muestras. Pero en vez de compararlas respecto a sus medias, las compara respecto a sus distribuciones probabilísticas subyacentes.

Las diferencias buscadas serán entre las distribuciones de los grupos. De existir una diferencia significativa para éste test, puede deberse a una diferencia no sólo entre las distribuciones, sino también entre las mencionadas medias o varianzas. El test trabaja sobre las funciones de distribución acumulada de cada grupo. Para cada posible valor, se aplica la función de distribución acumulada perteneciente a cada grupo, y luego, se compara la distancia entre los dos nuevos valores obtenidos. La mayor distancia encontrada será el resultado estadístico del test (D).

$$D = \sup_x |F_A(x) - F_B(x)|$$

Al igual que el test de Student, el valor D que devuelve Kolmogorov-Smirnov puede ser utilizado para obtener un *p-value* y así poder evaluar la existencia de una diferencia significativa entre distribuciones.

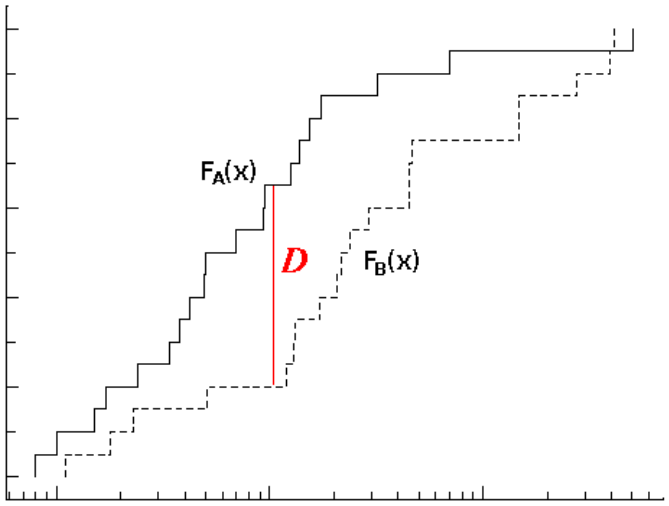


Figura 2: Ejemplo de las probabilidades acumuladas de dos distribuciones, mostradas como $F_A(x)$ y $F_B(x)$. La mayor distancia entre las distribuciones se muestra como D y ese valor será el resultado del test.

2. Analisis sobre el alumnado de nuestra carrera

En esta sección realizaremos un análisis sobre los alumnos de la carrera de Ciencias de la Computación de la FCEN. Nuestro enfoque no sólo será en aspectos puramente académicos sino también buscaremos observar si algunas características propias de cada alumno, como el tener un trabajo o realizar algún deporte, influyen en su desempeño académico.

A partir de los diferentes conjuntos de datos, agruparemos a los alumnos según ciertos criterios para poder analizar diferentes aspectos de los mismos y ver en qué se parecen y diferencian.

2.1. Obtención de datos

Existen varias fuentes de datos disponibles acerca de los alumnos de la carrera de Ciencias de la Computación en la Facultad de Ciencias Exactas y Naturales qué a modo anecdótico vale la pena mencionar. Algunas de ellas son:

- Censo de la UBA: La Universidad de Buenos Aires pone a disposición un análisis de los datos obtenidos en el censo que realiza a estudiantes todas las carreras (incluida Computación).
- Encuestas Docentes: El departamento de Computación de la Facultad de Ciencias Exactas y Naturales de la UBA pone a disposición los resultados de las encuestas docentes de los diferentes cuatrimestres. En las encuestas se pueden encontrar datos como puntuaciones a los diferentes docentes otorgadas por los alumnos y comentarios anónimos (de los alumnos) acerca de la cursada o los docentes que forman parte de la misma.
- Sistema de inscripciones: El sistema de inscripciones cuenta con información académica sobre los alumnos, como ser cursadas y finales aprobados, notas y condición de regularidad. Esta información no está disponible públicamente, para obtener los datos se debe realizar el pedido justificado al Departamento de Alumnos de la FCEN.

Sin embargo, estos no fueron suficientes para realizar un análisis cualitativo de interés, o bien, no contaban con la información necesaria para realizar el tipo de análisis buscado. De las fuentes mencionadas, el presente trabajo utiliza los datos del Sistema de Inscripciones.

Nuestro análisis se enfocaba en conocer no sólo aspectos académicos (como notas de exámenes o cursadas aprobadas) sino también antecendentes previos a ingresar a la facultad, propios de cada alumno (como saber programar antes de ingresar a la carrera, desde que año trabaja etc.), algunas actitudes propias de los alumnos en cuanto a la facultad (como si suelen hacer preguntas durante la clase, si alguna vez pensaron en dejar la carrera etc.) y otras preguntas que no pueden responderse con los datos disponibles mencionados anteriormente.

En lo que respecta al Sistema de Inscripciones, pudimos solicitar datos puramente académicos de los alumnos. De todos modos, cabe destacar que los datos solicitados al Departamento de Alumnos fueron proporcionados posteriormente al análisis de esta sección, por lo que realizaremos un análisis utilizando exclusivamente estos datos más adelante.

Para solucionar el problema de falta de datos de interés para nuestro análisis, decidimos obtener datos por nuestra cuenta. Como primera fuente de datos,

realizamos una encuesta para alumnos de ámbos géneros, graduados y no graduados de la facultad. La encuesta se realizó en base a factores de los cuáles nos interesaba analizar y estudiar su influencia en el éxito académico de los encuestados.

2.2. Elaboración de la encuesta

La característica más importante a la que apunta la encuesta, y la que guió la formulación de las preguntas, fue el rendimiento académico. Fue elaborado un conjunto de preguntas posibles, del cual fueron seleccionadas treinta y cinco para la versión final. Las respuestas dependen de percepción personal de cada individuo y son sensibles a la fiabilidad del encuestado. Decidimos no analizar estos puntos en nuestro trabajo, quedando como una limitación del mismo y como posible trabajo futuro.

Las preguntas principales hacen foco en cuestiones académicas. Estos datos no sólo ayudarán a analizar el comportamiento del alumnado, sino que también a formular una definición de rendimiento académico. Con dicha definición, se podrán comparar los datos obtenidos en las otras preguntas y medir el efecto sobre el rendimiento. Este tipo de preguntas incluyen: cuantas materias le restan cursar y cuántas veces recursó el encuestado, cuántos finales adeuda, qué tan bien le fue durante sus cursadas, entre otras.

También se incluyeron preguntas sobre antecedentes personales relacionados con conocimientos de inglés y de programación, previos al ingreso a la carrera. Sobre este último punto, una pregunta que suelen realizar futuros ingresantes y nuevos estudiantes es, si hay que tener nociones de programación como requisito previo para entender los temas de la carrera.

Nuestra hipótesis sobre este tema, es que los conocimientos previos son irrelevantes para el éxito academico y usaremos los datos de la encuesta para ver si es cierta o no.

Se decidió incluir preguntas referidas a la vida extracurricular de los alumnos, por el interés que generaría encontrar una correlación entre estos datos y el rendimiento académico. Estas preguntas incluyen pasatiempos, que pueden estar relacionados a la computación o no.

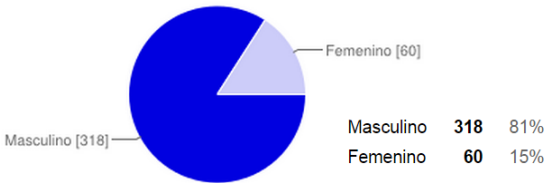
Como la muestra incluye datos de estudiantes actuales y graduados, se pueden conocer las propiedades y diferencias de cada uno de los subgrupos. Esto ayudará a reforzar nuestra definición de rendimiento académico.

Una vez conformada la encuesta, incentivamos al alumnado a llenarla, difundiendo la encuesta vía e-mail mediante la lista de distribución de la carrera y diferentes grupos de alumnos creados en las redes sociales que los alumnos de la carrera frecuentan. Se obtuvieron un total de 391 respuestas en el periodo de Junio y Julio de 2014.

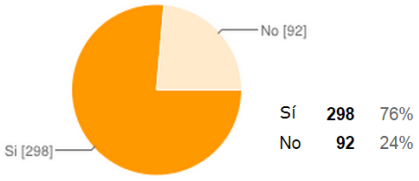
La encuesta completa con todas sus preguntas se puede ver en el Anexo 1.

2.3. Análisis exploratorio

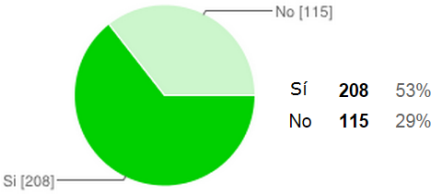
A continuación se exhiben los resultados más relevantes obtenidos a partir de las respuestas de los alumnos que completaron la encuesta (la lista completa de resultados se puede ver desde aquí).



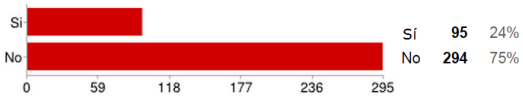
(a) Género de los encuestados



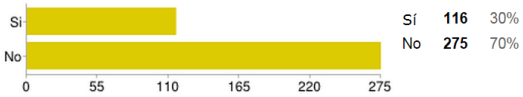
(b) ¿Trabaja?



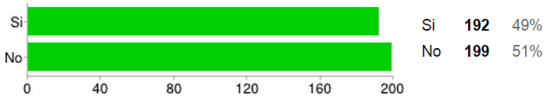
(c) ¿Cursaste menos por trabajar?



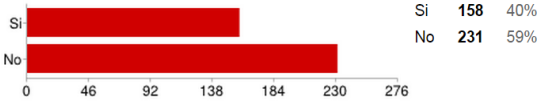
(d) ¿Sos Graduado?



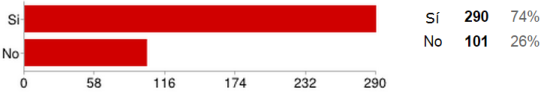
(e) ¿Sos/fuiste docente?



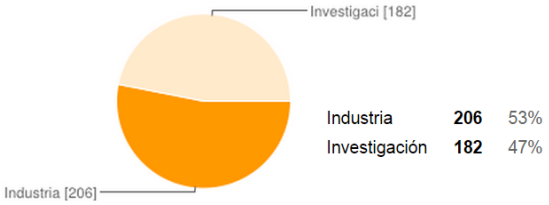
(f) ¿Sabias programar antes de entrar a la carrera?



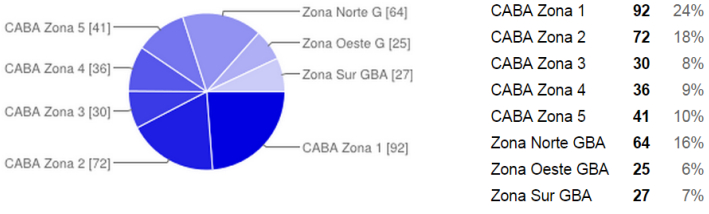
(g) ¿Alguna vez pensaste en dejar la carrera?



(h) ¿Hacés cosas relacionadas a la computación en tu tiempo libre?



(i) ¿Preferís el perfil de industria o investigación?



(j) ¿En que zona vivís actualmente?(ver mapa para CABA)

Figura 3: Algunos resultados de la encuesta realizada a alumnos de ambos géneros de la carrera Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales de la UBA.

El mapa correspondiente al último gráfico se puede consultar en el Anexo 1 donde se encuentran todas las preguntas que fueron incluidas en la encuesta. Como suele ocurrir en general en nuestra carrera [5] [18] [1] [8] no contamos con una presencia femenina significativa (apenas el 15 %), como se puede ver en la Figura 3a. En la sección 4 describiremos un estudio más exhaustivo sobre los

alumnos de género femenino [1].

También contamos con varios graduados, aún ligados a la facultad (al menos por medio de la lista de difusión de noticias de la carrera), un 24 % de los encuestados eran graduados (Figura 3d).

De la Figura 3b se desprende que la gran mayoría de los alumnos encuestados trabaja (el 76 %), como era de esperarse en una carrera como la nuestra, con una amplia salida laboral.

Uno de los resultados que llama la atención es la cantidad de alumnos que cursaron menos materias por trabajar (un 53 % de los encuestados), como se aprecia en la Figura 3c. La encuesta no especificaba en que período los alumnos cursaron menos materias. Es decir, pudieron haber cursado menos en años anteriores, o bien, durante el 2014. Es posible, que haya sido un problema pasado y, como consecuencia, en la actualidad hayan dejado de trabajar para poder cursar con normalidad. Otra posiblidad es el hecho de que actualmente se encuentren cursando una menor cantidad de materias debido a la demanda de tiempo que conlleva tener un trabajo. Lo cierto es que el problema afectó a una gran porción de los encuestados (más de la mitad).

En la Figura 3e observamos que el 30 % de los encuestados decidió dedicar su tiempo a la docencia en la facultad, lo cual es un porcentaje considerable de alumnos.

Otro resultado que se desprende de las respuestas, es que gran parte de los alumnos pensaron alguna vez en dejar la carrera (Figura 3g). Esto será motivo de análisis para saber si realmente influye o no en el rendimiento académico. Abordaremos el problema en secciones posteriores.

A su vez, se observa que la mayoría de los encuestados realiza tareas relacionadas con la computación frecuentemente en su tiempo libre. Esto último habla de la predisposición de los encuestados a practicar la computación como pasatiempo y no únicamente con fines académicos o laborales.

En cuanto a la preferencia entre el perfil de industria o investigación, en la Figura 3i nos encontramos con una distribución de respuestas equilibrada. Esto demuestra que nuestra carrera se adecúa a cualquiera de los dos perfiles a los que aspiren nuestros estudiantes. El 53 % de los encuestados escogió el sector de industria, mientras que el 47 % se inclinó por el de investigación. De todos modos, recibimos algunos comentarios sobre esta pregunta. Algunos alumnos mencionaban no estar del todo seguros de la respuesta escogida, lo cual hace aún más notoria la poca diferencia que hay para escoger uno de los dos perfiles.

En el gráfico de de las zonas donde viven los estudiantes (Figura 3j), se puede apreciar que las predominantes son la Zona 1 y 2 (las más cercanas al predio de Ciudad Universitaria) con un 24 % y 18 % de los encuestados, respectivamente. Sin embargo, no hay una gran diferencia con respecto a las demás zonas, en general los alumnos pertenecen a diversos lugares. Fuera de la Ciudad de Buenos Aires, llama la atención el porcentaje de alumnos provenientes de Zona Norte (un 16 %). Este hecho puede ser indicador de que quizás convenga ampliar y concentrar tareas de divulgación de la carrera en dicha parte de la provincia de Buenos Aires.

Otro resultado interesante de este primer análisis exploratorio, es la cantidad de alumnos que poseía conocimientos de programacion antes de empezar la carrera. Casi un 50 % de los encuestados tenían esta habilidad (Figura 3f).

[1] Género autopercibido por cada individuo

En vistas de que la carrera de Ciencias de la Computación tiene una amplia salida laboral (aún cuando los alumnos cuentan con poca experiencia) se generó un gráfico para analizar desde que año comienzan a trabajar los alumnos encuestados (ver Figura 4).

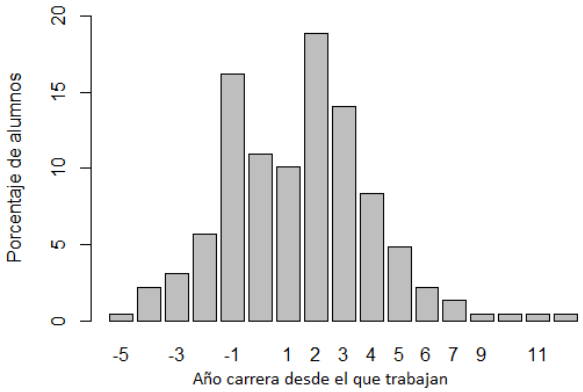


Figura 4: ¿En que año de la carrera comienzan a trabajar los alumnos? Se puede ver que gran parte comienza a trabajar en el segundo año (donde los alumnos ya cuentan con las herramientas para hacerlo) o bien un año antes de ingresar. A partir del tercer año también vemos un pequeño pico, representando a los alumnos que comienzan a trabajar cuando cuentan con un poco más de herramientas provistas por otras materias de la carrera.

El eje X del grafico de barras muestra la diferencia entre el año de ingreso a la carrera y el año de comienzo de actividades laborales. Entonces, un valor positivo en el eje X muestra a partir de que año de la carrera comenzó a trabajar un alumno; mientras que un valor negativo siginfica que el alumno comenzó a trabajar esa cantidad de años antes de entrar a la carrera.

Muchos alumnos comienzan a trabajar en el segundo año de la carrera. También existe otro pico en el año anterior al de ingreso, correspondiente al CBC. Esto refleja la facilidad que tienen los alumnos de nuestra carrera para conseguir trabajo, aún con poca experiencia en el área. Este fenómeno probablemente se deba a la gran demanda de puestos laborales en nuestro campo de estudio[7].

Otra hipótesis posible es que los alumnos comiencen la carrera porque se lo hayan exigido en su trabajo actual, para obtener más conocimientos en el área en el que se desarrollan laboralmente.

Estos resultados muestran una tendencia, por parte de los alumnos, a querer empezar la carrera y formarse profesionalmente al mismo tiempo. El porcentaje de alumnos encuestados que comienza a trabajar luego de recibirse es pequeña si se considera que la duración de la carrera es de por lo menos cinco años.

Más allá de estos resultados parciales, resulta útil separar el alumnado en ciertos grupos de interés para observar sus diferencias y similitudes. En la siguiente sección analizaremos características de diferentes grupos en los que se dividió al alumnado.

2.4. Análisis exploratorio por grupos

Como dijimos anteriormente, se realizaron dos tipos de divisiones dentro del alumnado:

- Los que trabajan y los que no trabajan
- Graduados y no graduados

Elegimos realizar una distinción entre los alumnos que trabajan y los que no, dado que el hecho de trabajar demanda tiempo y dedicación y puede afectar el desempeño académico de los alumnos. Un ejemplo de esto son (como vimos en el análisis exploratorio de la sección 2.3) los alumnos que cursan menos materias para poder trabajar. Otros factores incluyen falta de tiempo a la hora de estudiar, llegar tarde a clase por los horarios laborales que se deben cumplir y otros derivados de las consecuencias de tener menos tiempo para dedicar al estudio.

Por el lado de los graduados y los no graduados, se busca observar que características presentan los alumnos graduados y compararlas con los alumnos que aún continúan estudiando.

Para esta última división, es importante aclarar que todos los alumnos graduados fueron incluidos dentro del mismo grupo sin importar el momento en que se hayan graduado. Debido a ello, hay que considerar que los factores sociales que pudieran haber afectado a los diferentes alumnos graduados de este grupo son dinámicos, y puede que hayan variado según la época de graduación de cada uno de ellos.

Al realizar las comparaciones entre las propiedades de los diferentes grupos, se aplicará el test estadístico de Student (mencionado en la sección 1.2) que nos indicará si los resultados obtenidos son estadísticamente significativos respecto a alguna medición en particular.

2.4.1. Trabajan vs No Trabajan

En esta sección analizaremos las propiedades de los alumnos que trabajan comparándolos con los que no lo hacen. Veremos que los alumnos que trabajan tienen una mayor tendencia a dejar la carrera, y menor tendencia a respetar el plan de estudios que los alumnos que no trabajan durante la cursada. Por otro lado, veremos que el trabajo parece no influir en la realización de los ejercicios de las guías prácticas de las materias, y tampoco en que los alumnos se inclinen por elegir el perfil de industria sobre el de investigación [2], como se podría suponer intuitivamente.

En un primer análisis, se decidió investigar cómo se distribuye la proporción de alumnos que pensaron dejar la carrera en estos grupos.

[2]Se entiende por perfil industrial al ejercicio de la profesión en entes privados o públicos y al perfil de investigación como una carrera científica en una organición del sistema nacional de ciencia y tecnologia

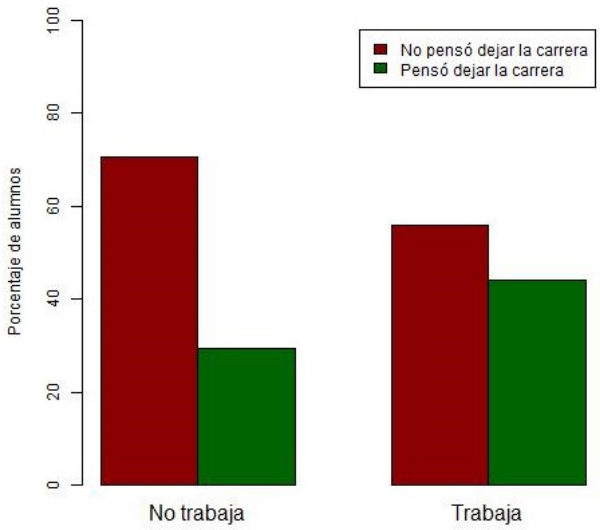


Figura 5: Porcentaje de alumnos que pensaron dejar la carrera en cada grupo, los que trabajan y los que no trabajan. El porcentaje de alumnos que no trabaja y tampoco pensó en dejar la carrera es mucho mayor que el de alumnos que trabaja y alguna vez pensó dejarla.

Como se observa en la Figura 5, hay una mayor tendencia a pensar en dejar la carrera entre los alumnos que trabajan. Cerca del 30 % de los alumnos que no trabajan pensaron en dejar la carrera, mientras que poco más del 70 % no pensó jamás en abandonarla. Por el lado de los alumnos que trabajan, la cantidad de alumnos que pensó en dejar la carrera y la que no son mucho más parejas con porcentajes cercanos al 45 % y 55 % respectivamente.

La carga horaria laboral quita tiempo de estudio. Por este motivo, pensamos que es uno de los hechos que dificultan las cursadas y, consecuentemente, pueden llevar al alumno a pensar en abandonar la carrera.

Al realizar el test de Student para este caso, obtuvimos un resultado significativo ($p < 0.0154$).

Otra característica interesante para analizar sobre esta división de grupos fue la cantidad de alumnos que respetan los tiempos del plan de estudios en cada uno de ellos.

El gráfico de la Figura 6 muestra cómo los estudiantes que no trabajan tienden a respetar el plan de estudios un poco más que los que lo sí lo hacen. Decidimos realizar el test para comprar si existían diferencias interesantes entre la cantidad de alumnos que respetan el plan de estudios y los que no, en cada grupo. El test de Student arrojó un resultado significativo ($p < 0.0410$). Por lo tanto, el test confirma que existe una diferencia entre los alumnos que trabajan y los que no en lo que respecta a cumplir con los tiempos del plan de estudios de la carrera. Esto era esperable, dado que los tiempos con los que cuentan los alumnos que trabajan son menores a los de los que no lo hacen.

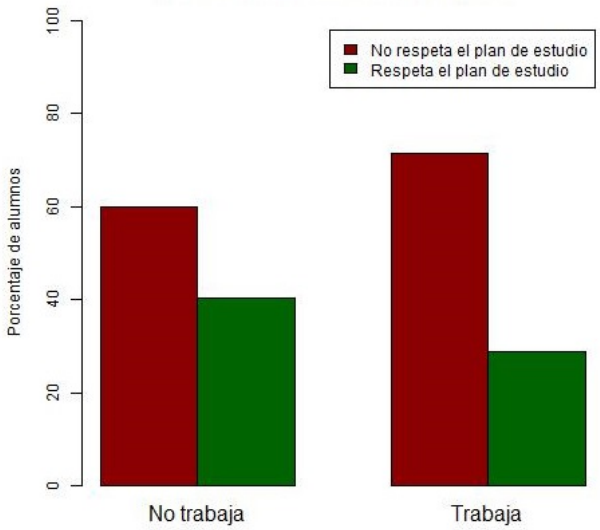


Figura 6: Porcentaje de alumnos que respeta el plan de estudios de cada grupo, los que trabajan y los que no trabajan. A priori parecería no haber mucha diferencia entre los porcentajes de alumnos que respetan el plan en uno y otro grupo, sin embargo los tests estadísticos confirman que la diferencia existe. Los alumnos que trabajan respetan con menos frecuencia el plan de estudios que los que no lo hacen.

Es un hecho que en nuestra facultad los docentes aconsejan fuertemente realizar las guías prácticas de las materias. Las mismas cuentan con ejercicios que sirven como entrenamiento y fuente de consultas para despejar dudas antes de un examen. En este marco, decidimos ver cuánto influye la situación laboral en la realización de los ejercicios prácticos.

Esperábamos observar que los alumnos que trabajan, al contar con menos tiempo libre, realicen menor cantidad de prácticas que los alumnos que no trabajan. Sin embargo, los resultados fueron otros.

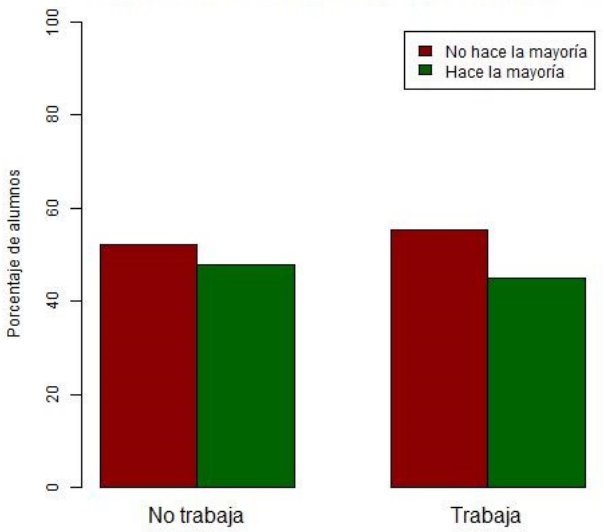


Figura 7: El porcentaje de alumnos que trabaja y hace la mayoría de los ejercicios no es muy diferente del los que no trabajan y también lo hacen. El trabajar no influye en realizar las guías prácticas de la materia.

En la Figura 7 se puede observar que no hay grandes diferencias entre ambos grupos con respecto a esta característica. Según el test de Student, la diferencia no es significativa ($p < 0.6114$).

En este caso podemos concluir, según nuestros datos, que el trabajar no influye directamente en que los alumnos realicen o no los ejercicios de las prácticas. La diferencia entre los alumnos que no realizan las prácticas de uno y otro grupo favorece por muy poco a los encuestados que no trabajan.

Otro planteo a analizar fue observar si el hecho de trabajar influye al elegir el perfil de industria sobre el de investigación.

En la Figura 8 se pueden observar los resultados de este análisis.

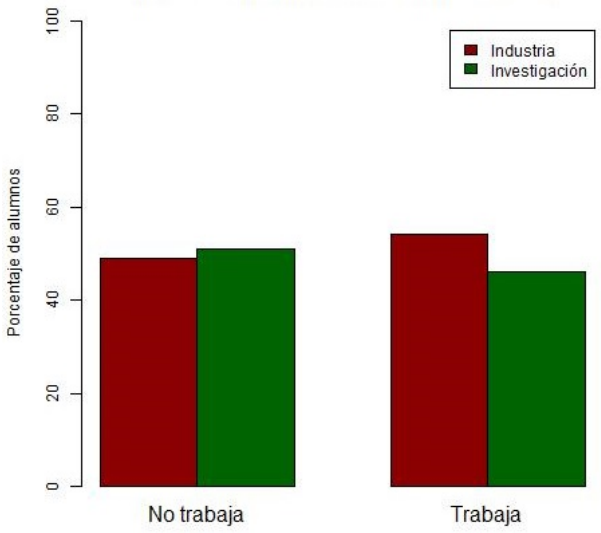


Figura 8: No hay muchas diferencias entre las preferencias de los perfiles de industria e investigación en ambos grupos (alumnos que trabajan y no trabajan). Apenas se diferencia una pequeña inclinación por la industria de los alumnos que trabajan. Los tests estadísticos confirman que la diferencia no es significativa.

El test de Student indica que la diferencia entre las preferencias de ambos perfiles para cada uno de los grupos, no es significativa ($p < 0.2731$). Es decir, el test confirma lo que se observa en las figuras: Trabajar no influye a los alumnos para elegir el perfil de industria sobre el de investigación.

2.4.2. Graduados vs No Graduados

En esta sección analizaremos a los alumnos separándolos según si son graduados o no. Separando los datos de esta manera, veremos que se mantiene el hecho de que no saber programar antes de comenzar la carrera no influye en el rendimiento académico. También observamos que los datos reflejan la gran salida laboral que tiene la carrera, viendo como casi la totalidad de los graduados trabaja y gran parte de los no graduados también.

Se decidió observar si la característica de saber programar antes de comenzar la carrera se mantenía si se separa al alumnado entre graduados y no graduados. Recordemos que los encuestados se dividían casi en un 50 y 50 entre los que sabían programar antes de ingresar a la carrera y los que no. Con este análisis esperábamos analizar si este factor se mantiene aún al realizar una distinción entre estos dos grupos.

Los resultados nos muestran que, en efecto, la característica se mantiene, como se observa en la Figura 9. El test de Student indica que no hay gran diferencia entre las medias de ambos grupos ($p < 0.1425$), por lo que no hay evidencia de que haga falta saber programar para graduarse.

Sumando este último resultado, podemos decir que no hace falta saber programar con anterioridad para estudiar en nuestra facultad, ni tampoco para lograr graduarse en la misma.

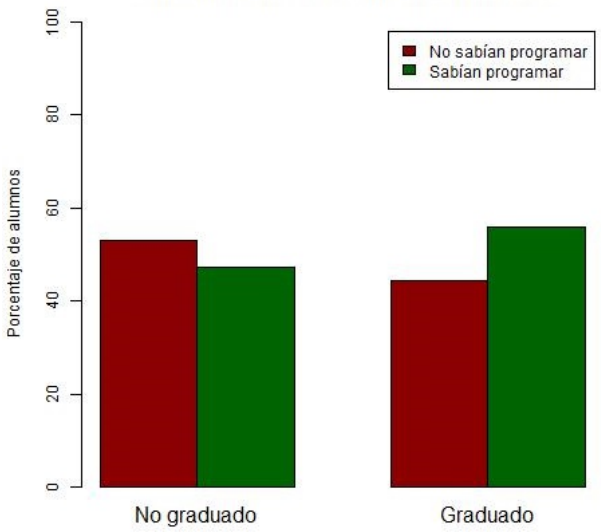


Figura 9: Porcentaje de alumnos que sabían programar antes de ingresar a la carrera y no entre graduados y no graduados. No existen grandes diferencias entre los graduados y los no graduados que sabían programar, los tests estadísticos confirman que no hay diferencias significativas en este aspecto.

Por último, decidimos ver la situación laboral de estos dos grupos. Dada la demanda laboral de nuestro campo de estudio, y lo concluido en secciones pasadas acerca del año en que empiezan a trabajar los alumnos, esperamos encontrar una amplio porcentaje de alumnos con trabajo.

Observando la Figura 10, se aprecia que los graduados trabajan casi en su totalidad (cerca del 100 %), mientras que los no graduados también tienen una gran tendencia a trabajar (cerca del 70 %), confirmando así nuestra hipótesis.

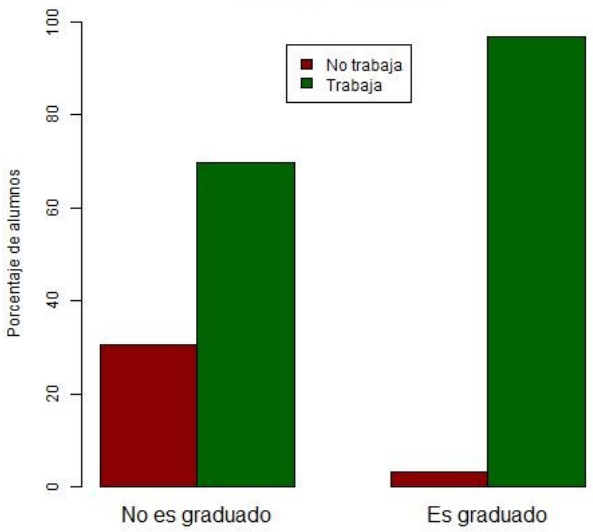


Figura 10: Porcentaje de alumnos que sabían trabajan y no, entre graduados y no graduados. Casi la totalidad de los graduados cuenta con un trabajo. Los alumnos no graduados trabajan en menor medida, sin embargo, el porcentaje es elevado con respecto al total de su grupo (casi un 70 %)

El test de Student indica que hay una gran diferencia entre las medias de estos dos grupos ($p < 5.0167e^{-16}$). Esto confirma que si bien gran parte de los alumnos no graduados trabaja, no se compara con los graduados donde el porcentaje de alumnos que no trabaja es ínfimo.

2.5. Análisis del rendimiento académico

Una vez realizado un análisis exploratorio de los datos obtenidos a través de la encuesta general al alumnado, se decidió construir un método predictivo sobre ellos. El objetivo radica en poder predecir si un alumno está rindiendo acorde a lo que se espera según el plan de estudios de la carrera.

El primer paso fue excluir del análisis a los alumnos encuestados que fueran graduados (sólo nos interesan los alumnos que se encuentran todavía cursando).

Tomamos cada pregunta de la encuesta como un *feature* a utilizar en nuestro modelo (por ejemplo género (M/F): indica el genero del alumno encuestado; Cómo sentís que te esta yendo: determina la percepción del alumno sobre cuán bien le está yendo en la carrera, etc.).

Luego, identificamos y extrajimos las *features* más representativas para evaluar el rendimiento de los alumnos. Utilizando estas últimas, se pudo construir una nueva *feature* que indica si a un alumno le va según lo esperado por el plan de estudios.

2.5.1. Extracción de features

Tradicionalmente, los métodos de extración de *features* crean nuevas *features* de un set de datos, a partir de la aplicación de funciones sobre las *features* originales. Mayormente se aplican a problemas que involucran texto, imagen o video ya que pueden generar *features* útiles automáticamente. Estos problemas

se caracterizan por poseer vectores con un gran número de dimensiones y, por consiguiente, son difíciles de analizar. Utilizando *Feature* Extraction se obtienen ciertas características de los vectores que cumplen ser importantes y con significado.

En nuestro caso, el enfoque será el de una construcción no automática de las *features*. Nos basaremos en el análisis de los datos obtenidos en la encuesta para poder determinar, a nuestro criterio, un indicador de rendimiento académico o de postergación de la carrera.

2.5.2. Escogiendo features

Se observaron todas las *features* con las que se contaba en la encuesta a fin de decidir cuales eran las más significativas. Con esto se busca determinar si un alumno está o no rindiendo acorde al plan de estudios.

Previamente a escoger las *features*, es importante mencionar y tener en cuenta que todos los valores que pueden tomar están relacionados con el año de ingreso del alumno a la carrera. Por ejemplo, es esperable que un alumno que ingresó hace cuatro años tenga pendiente menos cursadas que uno que ingresó hace uno.

Entonces, dado que el año de ingreso permite saber hace cuánto tiempo se encuentra en la facultad el alumno y cuánto tuvo para realizar sus estudios, se pueden estimar los valores que se espera que tomen los *features* a escoger. La lista completa de *features* y su descripción se puede consultar en el Anexo 3.

Finalmente se decidieron utilizar las siguientes *features*:

- CantVecesRecurso
- CantFinalesAdeuda
- CursadasSinFinalFaltan
- DebeAnalisis

A partir de las *features* listadas se puede tener una noción de qué tan avanzado se encuentra un alumno en la carrera según el plan de estudios y su año de ingreso.

En base a la lista se creó una nueva *feature* para decidir si un alumno se encuentra dentro de los valores esperados según el plan de estudios. De no ser así, diremos que ese alumno está postergando la carrera.

Cabe mencionar que dentro de la lista completa de *features* se encuentra 'CómoSentísQueTeEstáYendo'. Esta *feature* representa la opinión personal de cada alumno acerca de su propio rendimiento. Como se observa en la lista de *features* elegidas, se decidió dejar fuera esta *feature*. Si bien se podría haber agregado a la lista argumentando que la opinión del alumno es importante para conocer el rendimiento real, decidimos centrarnos solamente en los hechos objetivos. En la sección 2.6 hablaremos más sobre este *feature* que presenta particularidades interesantes.

2.5.3. Creando una nueva *feature*

Se creó entonces una nueva *feature* binaria a la que llamamos 'PostergaCarrera'. Esta *feature* pretende describir si un alumno tarda más de lo esperado

en cursar las materias. Nos centraremos en esta faceta del éxito académico. La *feature* 'PostergaCarrera' es binaria, por lo tanto: Diremos que un alumno posterga la carrera si la *feature* tiene un valor de 1 (uno). Por contrario, si el alumno cumple con los tiempos esperados, el *feature* 'PostergaCarrera' tendrá un valor de 0 (cero).

El criterio utilizado se basa en las *features* anteriormente listadas. Como mencionamos, hay que tener en cuenta el año de ingreso de los alumnos, es por ello que el valor asignado para 'PostergaCarrera' de cada alumno dependerá del año de ingreso del mismo, siguiendo un criterio. El criterio se muestra en el Cuadro 1.

Año de ingreso	Cursadas adeuda [3]	Finales adeuda	Cant veces recursó	Adeuda análisis II(C)
2013	≥ 15	≥ 3	>4	No se aplica
2012	≥ 14	≥ 4	>4	No se aplica
2011	≥ 10	≥ 6	>4	No se aplica
2010	≥ 7	≥ 8	No se aplica	1
2009	≥ 7	≥ 7	No se aplica	1
2008	≥ 5	≥ 5	No se aplica	1
2005-2007	≥ 5	≥ 4	No se aplica	1
≤ 2004	True	True	True	True

Cuadro 1: Criterio utilizado para definir el nuevo *feature* 'PostergaCarrera' el cual pretende describir si un alumno tarda más de lo esperado en cursar las materias o si su rendimiento académico esperado no es adecuado. La columna 2 en adelante contienen condiciones sobre el *feature* correspondiente según el año de ingreso de la columna 1. El listado completo de *Features* y su descripción se encuentra en el Anexo 3

Para cada año de ingreso de la tabla, si se cumple alguna de las condiciones para los *features* de su misma fila, se clasifica como positivo para 'PostergaCarrera', por lo tanto, se le asigna el valor 1 (uno).

En cuanto a la columna 'Adeuda análisis II(C)', al tratarse de una *feature* binaria, nos interesa saber si se adeuda dicho final a partir de cierto año de ingreso. Los años que nos interesa evaluar son aquellos para los que un alumno se encuentra avanzado en la carrera y (a nuestro entender) ya debería haberlo aprobado.

En cuanto al *feature* que indica la cantidad de veces que recursó un alumno, fue tomado en cuenta únicamente para los casos en que el año de ingreso del alumno fuera 2011 o mayor (es decir, los alumnos con menos años en la carrera).

Consideramos que un alumno puede haber recursado varias veces luego de varios años de cursada y sin embargo estar al día con la carrera (si cumple con los otros requisitos del criterio). Por otro lado, el hecho de recursar varias veces siendo un estudiante con pocos años en la facultad, puede llevarlo a postergar la carrera a futuro.

Teniendo esto en cuenta, al no considerar el *feature* de cantidad de veces que recursó para los alumnos con año de ingreso anteriores a 2011 (fila 4 del Cuadro 1, en adelante), enfatizamos la importancia de las otras características para los alumnos que hace más de cuatro años transitan la carrera. Esto último se debe, como dijimos, a que nos interesan más las cursadas y finales que adeudan esos alumnos que la cantidad de veces que recursaron materias.

[3] En base a la cantidad de materias obligatorias

Por último, los alumnos cuyos años de ingreso son menores o iguales a 2004 (fila 7 del Cuadro 1), se etiquetan como positivos para 'PostergaCarrera' (PostergaCarrera = 1) sin importar los valores que tomen el resto de los *features* de la condición. Esto se debe a que consideramos que si un alumno cursa la carrera desde hace diez años o más, alcanza para afirmar que no está cumpliendo con los tiempos pactados por el programa de la carrera.

A modo de ejemplo de lo discutido, según el criterio definido, un alumno de primer año [4] (año ingreso 2013) debería al menos haber aprobado la cursada correspondiente a tres materias, adeudar a lo sumo dos finales y no haber recursado más de 3 veces.

Una vez acordado el criterio, se etiquetó al conjunto de datos para que cuenten con este nuevo *feature* definido.

2.6. Decidiendo qué feature predecir

Antes de crear un modelo debemos escoger que *feature* se va a intentar predecir. Nuestro objetivo es poder caracterizar el rendimiento de un alumno. Entonces, parecería que 'PostergaCarrera' es un buen candidato. Sin embargo, al realizar una correlación entre las *features* veremos que surge otro buen candidato. Dada esta situación deberemos optar por predecir uno u otro, mediante algún tipo de análisis.

Sobre el *feature* 'PostergaCarrera' se puede realizar un análisis para determinar qué otras características de los alumnos están correlacionadas.

Dentro de la encuesta, uno de los ítems preguntaba al alumno cómo sentía que le estaba yendo en la carrera en una escala de 1 a 5 (donde 1 significa mal, y 5 bien). Consideramos que esta *feature* es determinada en parte por el rendimiento real del alumno. Si un alumno tiene un rendimiento no acorde al plan de estudios, puede que sea conciente de este hecho y haya contestado a acorde a este hecho.

En base a los *features* que tenemos, armamos una matriz de correlación para poder corroborrar si hay realmente relación entre 'CómoSentísQueTeEstáYendo' y 'PostergaCarrera', entre otras. Decidimos quedarnos con aquellos pares de *features* con correlación moderada. Para lograr esto, decidimos tomar las *features* que tienen valor de correlación absoluta mayor a 0.4.

Feature 1	Feature 2	Correlación
PostergaCarrera	CómoSentísQueTeEstáYendo	-0.5022
PostergaCarrera	PensoDejarCarrera	0.4138
Trabaja	HorasTrabaja	0.6117
Trabaja	AnioDesdeTrabaja	0.5405
PensoDejarCarrera	CómoSentísQueTeEstáYendo	-0.4580
CantVecesRecurso	CómoSentísQueTeEstáYendo	-0.5770
CantVecesRecurso	AnioIngreso	-0.4813
HorasTrabaja	AnioIngreso	-0.4519

Cuadro 2: Coeficientes de correlación moderada, para cada par de *features* listados en la columna 1 y 2. El listado completo de *Features* y su descripción se encuentra en el Anexo 3

[4] sin contar CBC

A partir del Cuadro 2, podemos ver que existe correlación entre 'CómoSentísQueTeEstáYendo' y 'PostergaCarrera' como esperábamos. Cuando la correlación es negativa, significa que el crecimiento de una de las variables puede ser explicado por el decrecimiento en la otra. En este caso, se corresponde que si 'PostergaCarrera' es positivo (valor 1) entonces el valor de 'CómoSentísQueTeEstáYendo' decrece.

También se relaciona 'PensoDejarCarrera' y otras variables. Sin entrar en muchos detalles, podríamos llegar a explicar intuitivamente estas relaciones. 'Trabaja' se relaciona con 'HorasTrabaja' y 'AnioDesdeTrabaja' dado que, en la encuesta, esas variables toman valores sólamente si 'Trabaja' es 'Sí'.

En cuanto a 'CantVecesRecursó' vs 'CómoSentísQueTeEstáYendo', es interesante ver que existe una tendencia en los alumnos a sentir que les va peor a medida que recursan con mayor frecuencia. Por último, 'AnioIngreso' vs 'CantVecesRecursó' y 'HorasTrabaja' puede verse como que a menor 'AnioIngreso' (lleva más años en la facultad), mayor la cantidad de horas que trabaja y la cantidad de veces que se recursaron materias.

Se decidió comparar los valores para 'PostergaCarrera' con los valores de 'CómoSentísQueTeEstáYendo', para observar la correlación más en detalle. Para ello se armó una matríz comparativa a fin de apreciar mejor las diferencias. La matriz se encuentra en el Cuadro 3 con su respectivo gráfico en la Figura 11

		PostergaCarrera	
		0	1
ComoSentisQueTeEstaYendo	1	0	18
	2	9	38
	3	32	42
	4	50	21
	5	43	11

Cuadro 3: Matriz comparativa entre los *features* 'CómoSentísQueTeEstáYendo' y 'PostergaCarrera'. En la columnas 0 y 1 de 'PostergaCarrera' se indican cuantos casos se encontraron dichos valores para cada posible valor del *feature* 'ComoSentisQueTeEstaYendo'

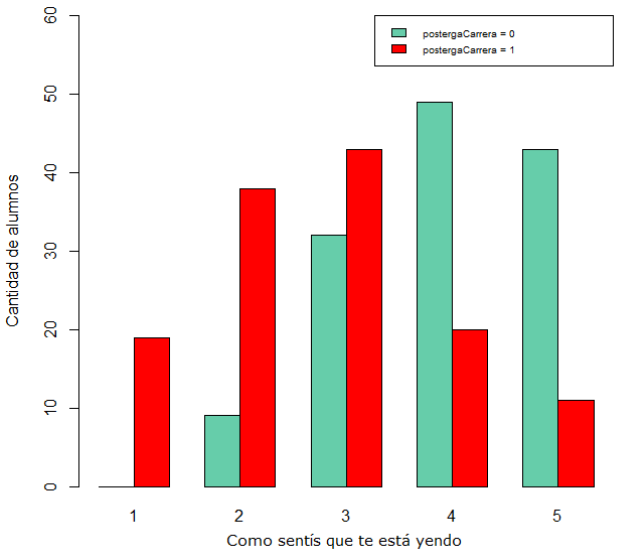


Figura 11: Gráfico de barras ilustrando los resultados expresados en el Cuadro 3. Los valores van del 1 al 5 donde 1 significa mal y 5 bien. Los alumnos que no postergan la carrera suelen optar por valores entre tres y cinco para expresar cómo sienten que les está yendo. Los alumnos que postergan la carrera, se inclinan por los valores dos y tres.

Para analizar ambos *features* de modo que tomen la misma cantidad de valores, decidimos binarizar los que tomaba el *feature* 'CómoSentísQueTeEstáYendo' (originalmente de 1 a 5, donde 1 significa que sienten que les fue mal y 5 que les fue bien). Con este fin, se consideró que aquellos alumnos que rinden menos de lo esperado son aquellos que se autocalificaron con 3 o menos en 'CómoSentísQueTeEstáYendo', mientras que los que rinden según lo esperado, con 4 o más. De esta forma, a aquellos que se calificaron con 4 o más les asignamos el valor 1 y al resto, el valor 0.

Cabe destacar que tomar esta decisión no es nada trivial dado que, como se observa en el Cuadro 3, los alumnos que calificaron 'CómoSentísQueTeEstáYendo' con valor 3 están bastante divididos entre los que creen que les va mal y los que creen que les va bien. Se consideró que 3 debía ser una medida que califique a un alumno que no está del todo convencido con su rendimiento. Aún así podría haberse optado por incluirla dentro de los que pensaron que les iba bien.

A continuación se muestra un resumen de la comparación del valor binario de 'CómoSentísQueTeEstáYendo' y el de 'PostergaCarrera'. Como comparten sus categorías de clasificación, podemos armar una matríz de confusión, la cual se encuentra en el Cuadro 4.

		PostergaCarrera	
		0	1
ComoSentisQueTeEstaYendo	0	93	32
	1	41	98

Cuadro 4: Matriz de confusión para los valores binarios de los *features* 'PostergaCarrera' y 'CómoSentísQueTeEstáYendo'. Los valores 1, 2 y 3 de 'CómoSentísQueTeEstáYendo' fueron traducidos al valor 0, mientras que los valores 4 y 5 se interpretaron con valor 0.

Como se observa de la matriz de confusión del Cuadro 4, hay muchos alumnos que sienten que les va bien según su propia percepción y, sin embargo, según el criterio del *feature* 'PostergaCarrera', esto no se cumple. También existen disparidades entre los alumnos que sienten que les va mal y, sin embargo, no postergan la carrera. Observamos los casos no coincidentes para tratar de detectar algunas características que causen este efecto.

Para los 32 casos en que los alumnos sintieron que les va bien y 'PostergaCarrera' indica lo contrario, detectamos algunas particularidades.

Notamos que este conjunto de alumnos realmente rindió menos de lo que se esperaba según el plan de estudios, basándonos en la cantidad de materias que les faltan cursar acorde a su año de ingreso a la facultad.

A modo de ejemplo, los alumnos con año de ingreso 2009 de este grupo deberían estar a un año de recibirse o quizá menos. Sin embargo, dichos alumnos adeudan la cursada de nueve materias o más, o no adeudan muchas cursadas pero sí muchos finales (siete en promedio).

Del año 2008, año para el cual ya deberían estar haciendo la tesis de licenciatura, encontramos alumnos que adeudan de cinco a siete cursadas, o bien adeudan siete u ocho finales.

Del año 2013, primer año de la carrera, encontramos alumnos que aprobaron dos o ninguna cursada hasta ahora, cuando deberían haber aprobado al menos tres siguiendo nuestro criterio.

Con la seguridad de que los casos fueron bien clasificados, nos detenemos a observar las particularidades del grupo que posterga la carrera. El 75 % trabaja y el 68,5 % trabaja 20 horas o más por semana. Por último, el 37,5 % vive en GBA.

En vistas de estas tres características se puede intuir que el trabajo y la distancia tienen gran incidencia en el rendimiento de estos alumnos.

Es comprensible que al vivir lejos de la facultad o trabajar una gran cantidad de horas, los objetivos a nivel académico de dichos alumnos no sean los mismos que los de uno que no tiene esta carga. Es por ello, que un alumno puede pensar que le va bien aunque no esté cumpliendo el plan de estudios, dada otros factores que inciden, como el trabajo y la distancia.

Pasamos a analizar al otro grupo de alumnos (para los que sentían que les iba mal y 'PostergaCarrera' indica lo contrario). Partimos de la premisa que aquellos alumnos que calificaron con el valor 3 en 'CómoSentísQueTeEstáYendo' buscaban catalogarse con un rendimiento promedio. Los casos restantes son 9 alumnos que se autocalificaron con 2. Según lo que se pudo observar de estos casos particulares, probablemente se deba a la cantidad de veces que recursaron o que tardaron un cuatrimestre más del esperado en superar el CBC. Aún así, muestran buen avance en la carrera.

Como pudimos ver a lo largo de esta sección, 'PostergaCarrera' resulta mejor indicador que 'CómoSentísQueTeEstáYendo' para evaluar si un alumno está por debajo de las expectativas del plan de estudios o no, dado que hay otros factores (trabajo, distancia, etc.) que influyen a la hora de calificar la autopercepción de los alumnos. Quizá el hecho de que un alumno se sienta atrasado tenga que ver con su grupo o 'camada' de alumnos, y sea un factor relativo a ese grupo. A la vez, un alumno puede sentir que le está yendo muy bien pese a tener un trabajo exigente (u otro factor similar) que le demanda atrasarse en sus estudios, por ejemplo. Por este motivo, se decidió intentar predecir el *feature* 'PostergaCarrera' en vez de 'CómoSentísQueTeEstáYendo'.

2.6.1. Selección de features

Realizamos una selección metodológica de *features*. La selección de *features* en Machine Learning se basa en eliminar *features* redundantes o ruidosas del conjunto de datos analizado. Estas *features* redundantes no aportan información y pueden generar problemas al momento de clasificar automáticamente los datos. Existen varias técnicas automáticas para la selección, como eliminar *features* con poca varianza (el *feature* no cambia mucho entre todos los casos del set de datos, por lo que no aporta información), usar regularización para penalizar *features* redundantes (utilizando la norma L1 o métodos como Lasso) o, en nuestro caso, basarse en la comparación de las variables en tests estadísticos. Esto lo lograremos utilizando análisis de la varianza (ANOVA) y χ^2 (Chi cuadrado) como funcion de scoring.

2.6.2. Descubriendo las features más significativas

Nuestro objetivo en la selección, es elegir las *features* relevantes para entrenar al método predictivo. De esta forma lograremos predecir con mayor exactitud el *feature* 'PostergaCarrera' utilizando regresión lineal.

En principio, 'PostergaCarrera' fue construido en base a otras *features* del set de datos. Debemos dejar fuera del clasificador a estos *features* porque son parte de la construcción de lo que queremos predecir. Recordemos que tampoco utilizaremos 'CómoSentísQueTeEstáYendo' dado que puede llevarnos a predecir erróneamente, como vimos en la sección 2.6, y tendrá gran peso en la decisión del clasificador dado que se parece (a grandes rasgos) y se encuentra correlacionado con 'PostergaCarrera'.

Los *features* que utilizamos como posibles predictores son:

- InfluenciaInternet
- HorasTrabaja
- CuatrimestresCbc
- PensóDejarCarrera
- PreguntaEnExamen
- SabíaProgramarAntes
- TieneGrupoTpFiel
- Trabaja
- TuvoBeca
- ZonaVive

Sabemos por el análisis exploratorio realizado en la sección 1.1 que 'SabiaProgramarAntes' no debería ser influyente, pero de todos modos observemos los resultados obtenidos por ANOVA para observar si se confirma esta hipótesis.

Se utilizó la función **glm** (Generalized Linear Model) del paquete estadístico R para construir un modelo utilizando regresión lineal (explicaremos regresión lineal en la sección 2.6.3) en principio utilizando todas las *features*. Luego, se

aplicó la función **anova** de R sobre dicho modelo para observar cuales eran los *features* más significativos. Los resultados fueron los siguientes:

Listing 1: ANOVA

```
Analysis of Deviance Table
Model: binomial, link: logit
Response: PostergaCarrera
Terms added sequentially (first to last)

                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                   263     365.97
InfluenciaInternet    1   6.8462       262     359.12  0.008883 **
HorasTrabaja          1  23.5448       261     335.58 1.220e-06 ***
CuatrimestresCbc      1  15.9492       260     319.63 6.507e-05 ***
PensoDejarCarrera     1  28.2864       259     291.34 1.046e-07 ***
PreguntaEnExamen      1   2.4654       258     288.87  0.116375
SabiaProgramarAntes   1   0.0406       257     288.83  0.840249
TieneGrupoTpFiel      1   8.5886       256     280.25  0.003383 **
Trabaja               1   0.8291       255     279.42  0.362530
TuvoBeca              1   4.1213       254     275.29  0.042347 *
ZonaVive              1   2.1380       253     273.16  0.143684
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

Como se observa en la Listing 1 las *features* más relevantes son las marcadas con el símbolo '*'. Se definen varios niveles de significación estadística, correspondientes a valores α comunes. Cuanto menor es el nivel, menor es la probabilidad de que el resultado haya sido obtenido por pura chance. A un valor α de 0.01, las *features* más significativas del análisis son:

- InfluenciaInternet
- HorasTrabaja
- CuatrimestresCbc
- PensoDejarCarrera
- TieneGrupoTpFiel
- TuvoBeca

Notemos que ANOVA dejó fuera de la lista al *feature* 'SabíaProgramarAntes' como esperabamos.

Una vez elegidas las *features*, se procedió a crear un método predictivo utilizando regresión logística y valiéndose, una vez más, de las herramientas proporcionadas por R.

2.6.3. Regresión lineal

La regresión lineal es una técnica utilizada para estudiar las relaciones entre dos o más variables o poder predecir algún tipo de respuesta cuantitativa. Es el estudio de cómo una de las variables cambia respecto a las otras. En cuanto al estudio de relaciones entre las variables, se puede obtener un valor que indica qué tanto una variable afecta a otra. La relación modelada entre las variables es

una relación lineal, pero es fácil obtener otros tipos de relaciones modificando minimamente esta técnica.

Es posible utilizar la técnica con predictores continuos o, como en nuestro caso, discretos. El objetivo es inferir una serie de coeficientes β_i que puedan explicar la relación linear de los predictores X_i con la variable de respuesta Y:

$$Y \approx \epsilon + \beta_0 X_0 + \beta_1 X_1 + ... + \beta_n X_n$$

La formula en general va a aproximar ($\approx$) a la variable Y y no lograr un resultado exacto, porque la recta es construida en base a una estimación de los datos. La recta va a ser la que mejor se ajusta a los datos según algún criterio. Nuestro criterio (y uno de los clásicos) va a ser que la recta tiene que minimizar la suma residual de cuadrados (*RSS, residual sum of squares*). Los β_i buscados van a ser los coeficientes que compongan esta recta.

Un *residual* está definido para cada X_i como la diferencia entre el valor real de la respuesta y_i y el estimado $\hat{y}_i = \epsilon + \beta_i X_i$:

$$e_i = y_i - \hat{y}_i$$

Entonces la suma residual de cuadrados está definida como:

$$RSS = e_1^2 + e_2^2 + ... + e_n^2$$

La recta que minimice el valor de RSS será la que obtendremos con regresión lineal.

2.6.4. Regresion logística

Nuestros datos a ajustar son categóricos. El *feature* 'PostergaCarrera' toma valores 'No' y 'Sí' (0 y 1). Para poder ajustar nuestros datos a la respuesta binaria, usaremos regresion logística. Usando solamente regresión lineal, los valores estimados podrían ser mayores que 1 o menores que 0, lo cual no es lo ideal. Este tipo de regresión, aplica una función logística sobre la relación lineal entre los X_i y la respuesta Y. La función logística entonces es:

$$\frac{e^t}{1 + e^t}$$

Donde t es el modelo lineal $\epsilon + \beta_0 X_0 + \beta_1 X_1 + ... + \beta_n X_n$.

2.6.5. Entrenando un método predictivo para PostergaCarrera

Valiéndonos de las herramientas del lenguaje R, y ya sabiendo que *features* utilizar, creamos un modelo utilizando la función glm.

Listing 2: GLM

```
logistic_fit <- glm(postergaCarrera ~ InfluenciaInternet
          + HorasTrabaja + CuatrimestresCbc
          + PensoDejarCarrera + TieneGrupoTpFiel
          + TuvoBeca,data=mydata,family = binomial("logit"))
```

Posteriormente, se utilizaron dos funciones predictivas de R para comprobar cuán bueno es el modelo obtenido.

La primera que se utilizó fue la función cv.glm de la biblioteca **boot** de R, la cual realiza lo que se conoce en machine learning como *cross validation*. La idea principal de esta función es dividir los datos aleatoriamente en M grupos.

Posteriormente, en diferentes iteraciones, se entrenará el modelo glm con los datos. En cada iteración, se omite (uno a la vez) a cada integrante de los M, el integrante excluido servirá como conjunto de testing.

Con el modelo ajustado al conjunto de datos parcial, se evalúa e intenta predecir el conjunto de testing obteniendo una estimación. Luego, se promedian las M estimaciones para obtener una única estimación del error. Utilizando los parametros por default de la función cv.glm, se realiza la variante llamada *Leave-one-out cross-validation*. Esta variante separa en cada iteración a sólo un sample del conjunto de datos, teniendo entonces que M es igual al tamaño del conjunto de datos.

```
cv.glm(mydata,logistic_fit)$delta # delta = 0.18 aprox
```

El error δ estimado según cv.glm fue de 0.18, resultando en casi un 82 % de efectividad.

Por otro lado, se utilizó la función *predict* de R. En este caso, fue necesario separar aleatoriamente los datos en dos grupos: un conjunto de entrenamiento, para el cual se ajustó el modelo a utilizar, y otro como conjunto de test. A continuación se exhibe parte del script utilizado, junto con la matriz de confusión correspondiente mostrando los resultados obtenidos.

```
# splitdf es una funcion devuelve una lista de conjuntos de train
# y test.
# Cada data frame contiene la mitad de la cantidad total de datos
splitdf <- function(dataframe, seed=NULL) {
  if (!is.null(seed)) set.seed(seed)
  index <- 1:nrow(dataframe)
  trainindex <- sample(index, trunc(length(index)/2))
  trainset <- dataframe[trainindex, ]
  testset <- dataframe[-trainindex, ]
  list(trainset=trainset,testset=testset)
}

# Aplicar la funcion a los datos con una seed arbitrario
splits <- splitdf(mydata, seed=810)


```

```
# Guardamos los conjuntos de test y training como data frames
train <- splits$trainset
test <- splits$testset

# Creamos el modelo con el conjunto de entrenamiento
logistic_fit <- glm(postergaCarrera ~ InfluenciaInternet
           + HorasTrabaja + CuatrimestresCbc
           + PensoDejarCarrera + TieneGrupoTpFiel
           + TuvoBeca,data=train,family = binomial("logit"))

# Ejecuta la funcion predict de R con el modelo obtenido
# para el conjunto de test
logistic_predict <- predict(logistic_fit,newdata=data.frame(test)
, type="response")

# Asignar 0 a los que predijo un valor inferior a 0.5
logistic_predict[logistic_predict < 0.5] <- 0

# Asignar 1 a los que predijo un valor superior o igual a 0.5
logistic_predict[logistic_predict >= 0.5] <- 1

# Asignamos las predicciones a un \textit{feature} 'predicted'
# del conjunto de test
test$predicted <- logistic_predict

postergaCarreravsPredicted <- test[,c("postergaCarrera",
      "predicted")]
```

Ahora ejecutamos el comando para ver la confusion matrix:

```
table(postergaCarreravsPredicted)
```

		predicted	
		0	1
PostergaCarrera	0	49	18
	1	13	52

Cuadro 5: Matriz de confusión de los valores originales del *feature* 'PostergaCarrera' y el valor predecido por el modelo utilizado. La gran mayoría de los valores fueron predecidos correctamente.

Como se observa en el Cuadro 5, se predijo correctamente 49 valores 0 y 52 valores 1 para 'PostergaCarrera', lo cual da un total de 101 predicciones correctas sobre un total de 132 casos. De esta manera, la efectividad es de aproximadamente 76,5 % en la evaluación.

Cabe destacar que este script es un caso particular para el cual se entrenó con la mitad de los datos y se predijo con el remanente, mientras que la función cv que fue usada anteriormente utiliza el método de 'leave-one-out'.

2.7. Factores del éxito académico: Análisis sobre PostergaCarrera vs noPostergaCarrera

A raíz de la distinción en dos grupos de alumnos a través del *feature* definido 'PostergaCarrera' vale la pena realizar algunos análisis de interés sobre estos. Observaremos que características presentan los alumnos de uno y otro grupo respecto a los demás *features*.

En esta sección veremos cómo las becas parecen influir positivamente en el rendimiento académico.

También observaremos cómo saber programar antes de comenzar la carrera tampoco influye en que los alumnos posterguen la carrera o no, como complemento a los análisis que se hicieron sobre este tema en particular para otros grupos de alumnos.

Podremos apreciar que las características como haber pensado pensar en dejar la carrera, trabajar, y no tener un grupo de trabajos prácticos que perdure por varios cuatrimestres, entre otras, influyen negativamente en el rendimiento académico.

Para los análisis de esta sección utilizaremos los tests de Student y Kolmogórov-Smirnov descriptos en la sección 1.2.

En primera instancia, decidimos ver si las becas que otorga la facultad está correlacionado de alguna manera para que los alumnos le dediquen más tiempo a la facultad y puedan desarrollarse mejor en la misma. Recordemos que en la encuesta se les preguntó a los alumnos si reciben o recibieron alguna vez una beca [5]. Parece una buena oportunidad para observar cuántos alumnos becados postergan la carrera y cuántos no.

Los resultados se pueden observar en la Figura 12

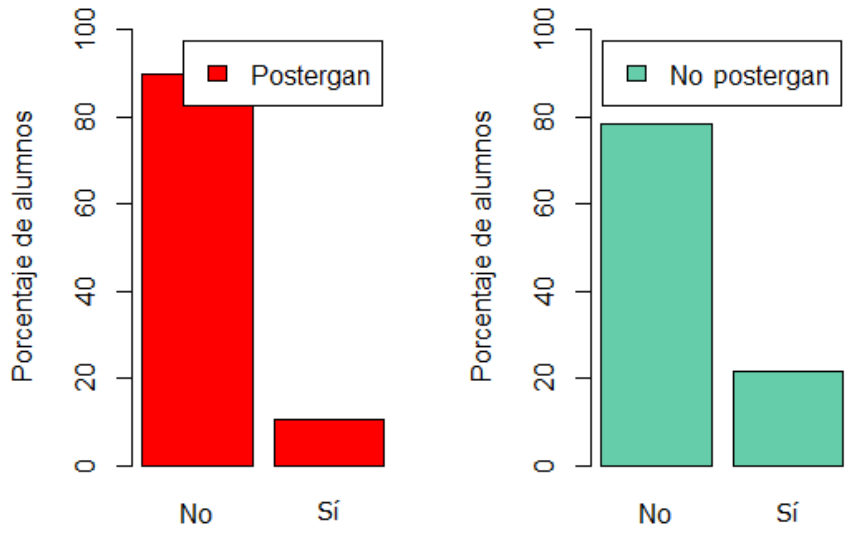


Figura 12: Porcentaje de alumnos con y sin beca, entre los que postergan y no postergan la carrera. No hay una diferencia significativa en cuanto a los becados de ambos grupos, a pesar de que el gráfico muestra que el porcentaje de becados que no postergan la carrera es levemente mayor que en los alumnos que sí la postergan.

[5] Otorgada o no por la facultad

Si bien no contamos con datos suficientes como para concluir que la beca ayude de manera significativa, en la figura podemos observar que para nuestra muestra en particular, el hecho de tener beca está correlacionado con postergar o no la carrera. Esto se ve reflejado en que hay un 22 % de alumnos becados que no postergan la carrera contra un 11 % que en los no becados.

El test de Kolmogorov–Smirnov indica que la diferencia de cuanto ayuda la beca a uno u otro grupo no es significativa, mientras que para el test de Student, sí lo es (p <0.3752 con $D_n = 0.1111$, y p <0.0125 , respectivamente).

Observemos que sucede con saber programar antes de comenzar la carrera para estos dos grupos, *feature* que ya fue centro de discusión en secciones anteriores, para observar si se mantiene la paridad para estos dos grupos de alumnos.

En la Figura 13 se puede observar como, una vez más, el *feature* se mantiene sin grandes diferencias entre ambos grupos: Aproximadamente un 58 % vs 42 % para los que postergan la carrera y un 45 % vs 55 % para los que no.

Los tests de Kolmogorov–Smirnov y Student comprueban que no hay diferencias significativas entre ambos grupos para este *feature* (p <0.1446 para Student y p <0.6604 con D_n= 0.0889 para Kolmogorov–Smirnov).

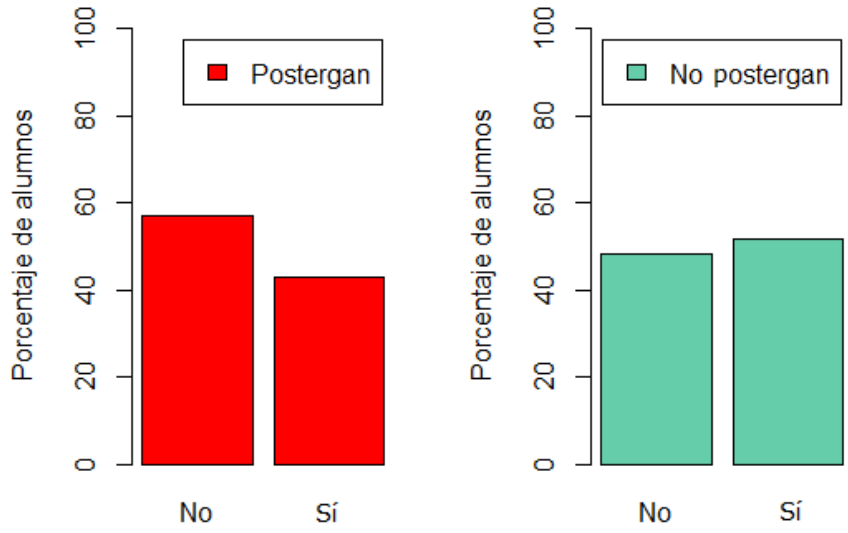


Figura 13: Porcentaje de alumnos que sabían o no programar antes de ingresar a la carrera, entre los que postergan y no postergan la carrera. No existen grandes diferencias entre los que sabían programar antes y los que no de cada grupo, los tests estadísticos confirman este resultado. Saber programar con anterioridad no influye en esta faceta del éxito académico de los alumnos.

Decidimos ver cuántos alumnos trabajan en ambos grupos. En la encuesta además de consultar por la cantidad de horas, consultamos a los alumnos si actualmente ocupaban un cargo laboral. Los resultados se observan en la Figura 14.

Como vemos, poco más del 80 % de los alumnos que postergan la carrera trabajan, mientras que dentro de los que no postergan la carrera, trabaja cerca del 60 %.

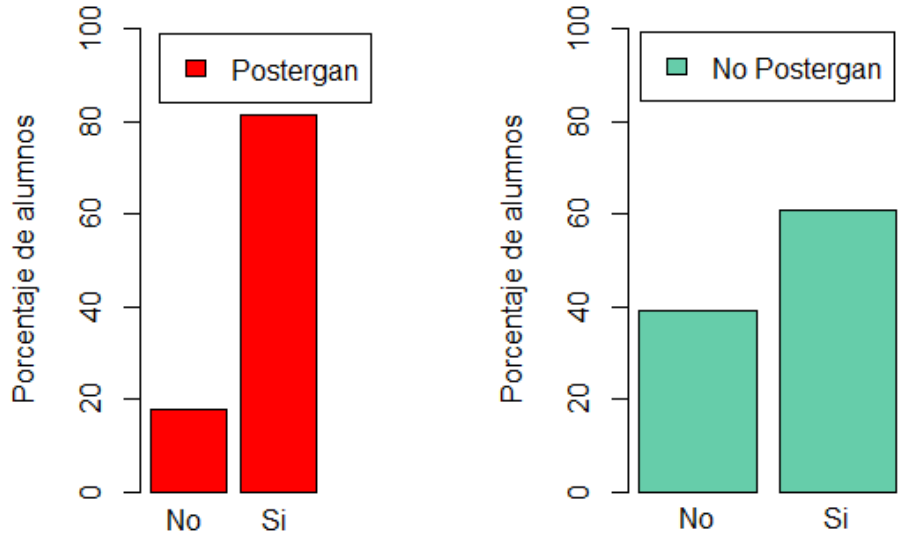


Figura 14: Porcentaje de alumnos que trabajan, entre los que postergan y no postergan la carrera. Una gran proporción de los alumnos que posterga la carrera trabajan (poco más del 80 %), mientras que los alumnos que no postergan la carrera y trabajan representan una proporción un poco menor dentro de su grupo (cerca del 60 %).

La diferencia es significativa, para los tests de Kolmogorov–Smirnov y Student ($p < 0.0060$ y $p < 0.0003$ respectivamente).

Podemos concluir que hay evidencia para respaldar el hecho que trabajar esta inversamente correlacionado con el éxito académico.

Siguiendo en la línea de lo laboral, otro análisis interesante fue ver la cantidad de horas que trabajan los estudiantes de ambos grupos, para observar si la carga laboral es un factor que influye en el éxito académico de los alumnos. En la encuesta, los alumnos debían seleccionar la cantidad de horas laborales que cumplían. En caso de no trabajar, se indica cero horas.

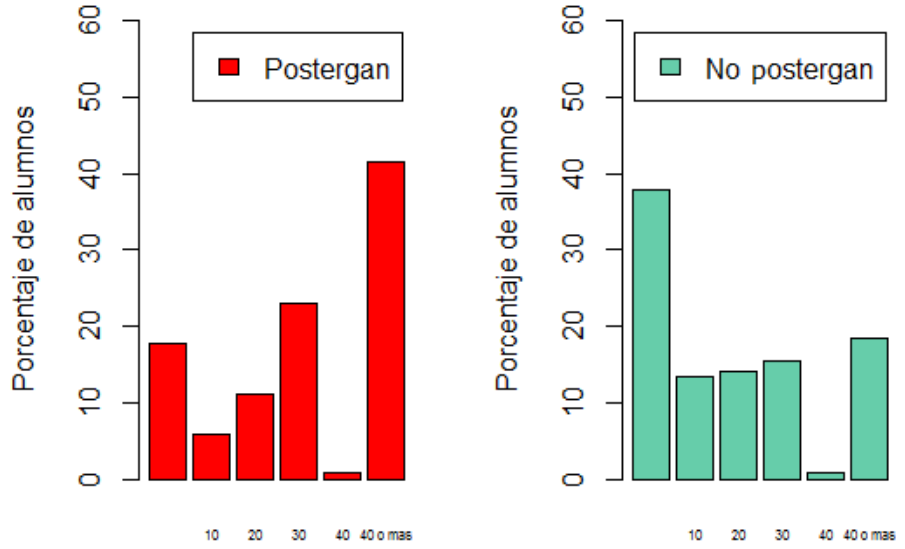


Figura 15: La cantidad de horas que trabajan los alumnos de uno y otro grupo (los que postergan y no, la carrera). Los alumnos que postergan la carrera trabajan muchas más horas que los que no lo hacen. El pico del gráfico para los que postergan la carrera se encuentra en las 40 horas o más semanales, mientras que el de los que no lo hace tiene su pico en cero horas (no trabajan) y cuentan con mayor cantidad de alumnos en las cargas horarias menos extensas.

Como se observa en la Figura 15, los estudiantes que postergan la carrera cuentan con cargas laborales mayores que los que están al día. Lo más llamativo son las dos cargas horarias extremas. En particular, el porcentaje de alumnos que postergan la carrera y trabajan 40 horas o más son cerca del 45 % contra apenas un 20 % dentro de los que no la postergan. Por otro lado, los que no trabajan (la barra más a la izquierda en la figura) dentro de los que postergan la carrera son el 17 % mientras que los que no la postergan casi un 40 %

Antes de realizar los tests, separamos los resultados en categorías '0,1,2,3,4,5' que se entienden como '10 horas, 20 horas, 30 horas, 40 horas y más de 40 horas', respectivamente. Los tests analizaron la media y la distribución sobre estas categorías creadas, según correspondía. Para este análisis, ambos tests de arrojan p-values significativos ($p < 0.0109$ para Kolmogorov–Smirnov y $p < 0.0002$ para Student).

Podemos concluir que la carga horaria laboral también influye en el éxito académico de los alumnos.

Para poder apreciar mejor la distribución de las horas en los dos grupos, podemos observar el gráfico de densidad de la Figura 16

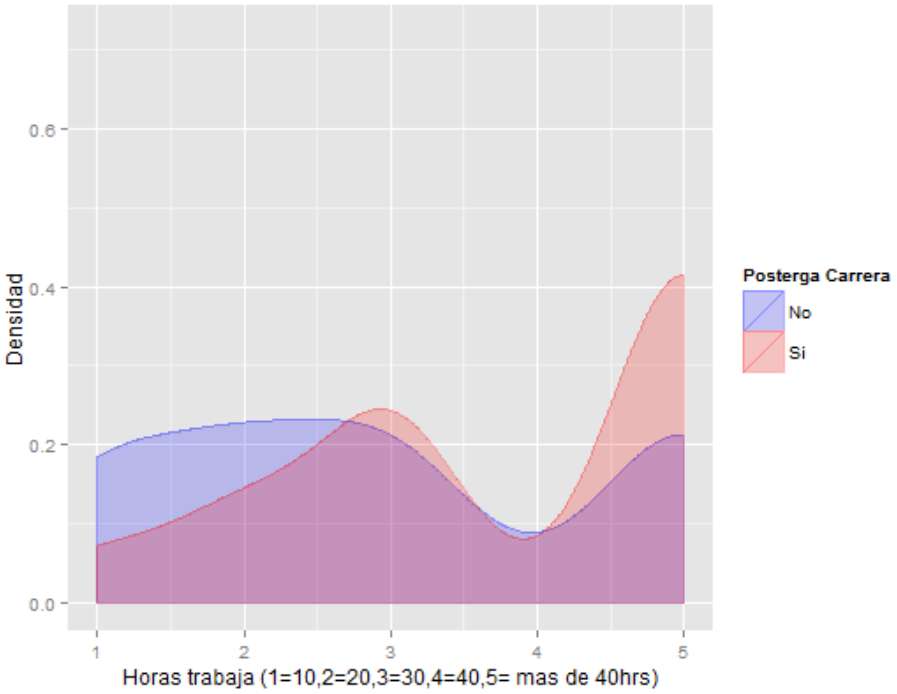


Figura 16: Grafico de densidad para la carga horaria laboral de los alumnos que postergan y no, la carrera. Los que no postergan tienen su moda sobre los valores más chicos de carga horaria. Para los que postergan la carrera ocurre lo opuesto.

Para continuar, otro de los análisis que arrojó una diferencia significativa para los tests de Kolmogorov–Smirnov y Student fue la influencia de Internet [6] en el éxito académico de ambos grupos. En la encuesta, los alumnos votaban cuánto creían que influía Internet en su éxito académico con un valor numérico entre 1 y 5.

Los resultados se exhiben en el gráfico de densidad de la Figura 17. Como podemos ver, la curva de densidad para los alumnos que no postergan la carrera es ascendente alcanzando su pico en el valor 5, mientras que la correspondiente a los alumnos que postergan la carrera se centra en los valores 3 y 4.

[6] El uso de Internet para complementar los estudios.

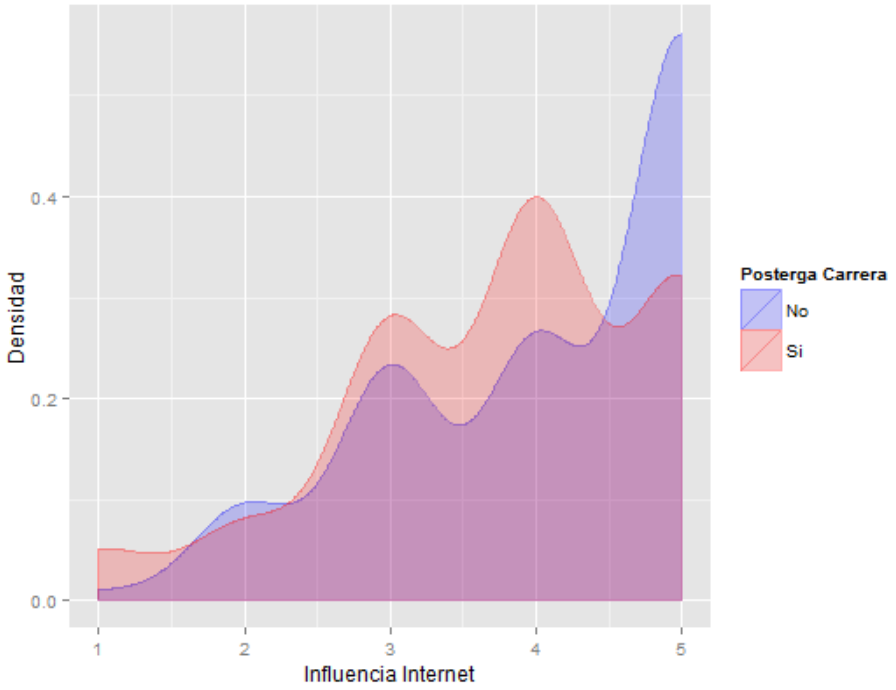


Figura 17: La percepción propia de cuanto influye Internet en el éxito académico sobre de los alumnos que postergan y no, la carrera. Los alumnos que no postergan la carrera afirman que Internet influye bastante en su éxito académico. Los que no postergan la carrera opinan que Internet es un poco menos influyente.

Los tests de Kolmogorov–Smirnov y Student arrojan un p-value significativo para resaltar la diferencia con respecto a este *feature* para ambos grupos (p $<$0.0079 y p $<$0.0097 respectivamente).

Podemos concluir que hay evidencia que muestra que los alumnos que no postergan la carrera utilizan decididamente Internet para complementar los estudios. Puede indicar también que tienen la percepción de que este hecho los ayuda de manera directa a conseguir el éxito académico y poder mantenerse al día con la carrera.

Continuando con los análisis, observamos la influencia del desempeño en el CBC en los alumnos de uno y otro grupo. Para ello, en la encuesta les consultamos a los alumnos la cantidad de cuatrimestres que les había tomado completar el Ciclo Básico Común para ingresar a nuestra carrera. La Figura 18 nos muestra que los que postergan la carrera suelen tardar entre dos o tres cuatrimestres en completar el CBC, mientras que a casi todos los alumnos que no postergan la carrera les toma sólo dos (que es lo que se espera). Los tests de Kolmogorov–Smirnov y Student confirman esta diferencia (p $<1.6842e^{-05}$ y p $<9.3963e^{-05}$ respectivamente).

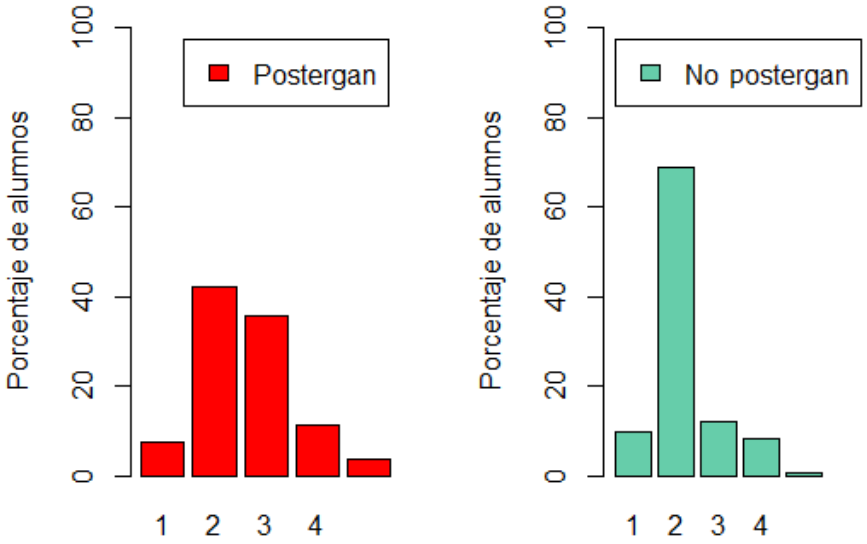


Figura 18: Cuantos cuatrimestres les lleva completar el CBC a los alumnos que postergan y no, la carrera. Los alumnos que no postergan la carrera suelen tardar dos cuatrimestres en su gran mayoría, mientras que los que postergan la carrera suelen tardar entre dos y tres.

Se pueden ver diferencias interesantes para los dos grupos con respecto a este *feature*. Notemos, además, que hay más alumnos a quienes el CBC les tomo tan sólo un cuatrimestre dentro de los que postergan la carrera que dentro de los que no lo hacen. Por este motivo, no nos atrevemos a afirmar por completo que postergar el CBC implique postergar la carrera. Sin embargo es probable que esto suceda si se tarda tres cuatrimestres o más en superar el ciclo, dada la diferencia de alumnos que cumplen esta característica en uno y otro grupo.

Tener un grupo de trabajos prácticos que perdure en el tiempo (o en varias materias) se suele creer que ayuda a consolidar al grupo, entenderse mejor, y aprobar con mayor facilidad los trabajos prácticos que demandan las diferentes materias de la carrera. Decidimos poner a prueba esta hipótesis utilizando el *feature* TieneGrupoTpFiel: en la encuesta se les consultó a los alumnos si tenían un grupo de trabajos prácticos que hayan mantenido por al menos cinco cuatrimestres. Los resultados se pueden observar en la Figura 19.

En su momento, los alumnos manifestaron que quizá cinco cuatrimestres era una cuota excesiva para mantener un grupo de trabajos prácticos, dada la diversidad de materias que eligen los alumnos. Esto se corresponde con los resultados que indicaban que gran parte de los alumnos no respetaba el plan de estudios estrictamente. Sin embargo, establecimos este valor con el objetivo de observar si el hecho de tener un grupo por un tiempo prolongado se correlacionaba con el éxito académico del alumnado.

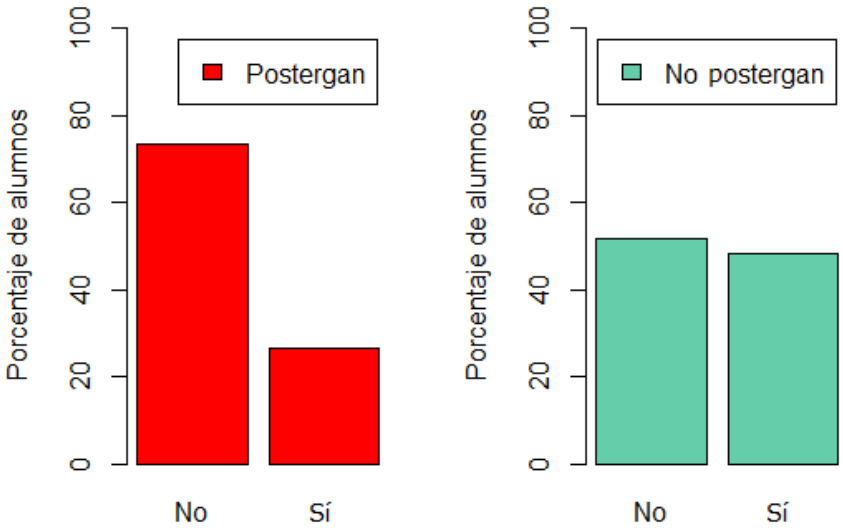


Figura 19: Porcentaje de alumnos que tienen un grupo de TPs "fiel"dentro de los que postergan y no, la carrera. Son bastante más en proporción los alumnos que no postergan la carrera y tienen un grupo de tp, que aquellos que lo tienen y postergan la carrera. Tener un grupo conocido con quien realizar los trabajos prácticos ayuda a mejorar esta faceta del éxito académico.

Como vemos en la Figura 19, la diferencia es considerable entre ambos grupos. Los alumnos que tienen un grupo de trabajos prácticos fiel son apenas un 23 % del grupo de los alumnos que postergan la carrera. Por otro lado, representan casi un 50 % dentro de los alumnos que no postergan la carrera.

Los tests de Kolmogorov–Smirnov y Student confirman cuán significativo es este análisis ($p < 0.0039$ y $p < 0.0002$ respectivamente).

Este resultado es bastante contundente. Hay evidencia fuerte que indica que tener un grupo de trabajos prácticos estable y compuesto de gente conocida, contribuye considerablemente al éxito académico de los alumnos ayudándolos a no atrasarse con el plan de estudios.

Durante la encuesta consultamos a los alumnos sobre su percepción de como les estaba yendo en la carrera, para ello debían contestar con un valor entre uno y cinco. Resulta interesante observar como varía este *feature* para ambos grupos de alumnos, por este motivo decidimos hacer un análisis también para 'CómoSentísQueTeEstáYendo'. Los resultados se pueden observar en la Figura 20

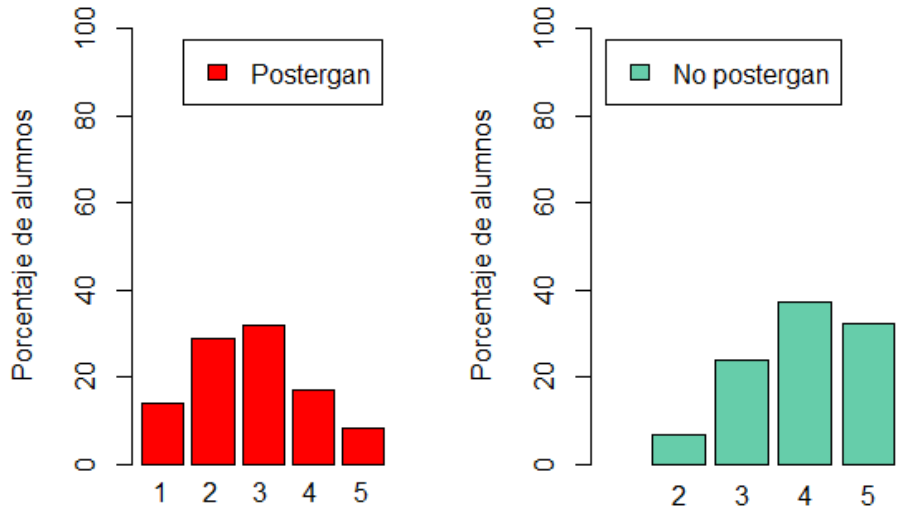


Figura 20: Como sienten que les va en la carrera a los alumnos que postergan y no, la carrera. Los alumnos que postergan la carrera escogen los valores más bajos para expresar como sienten que les va en la carrera. Los que no postergan la carrera, en cambio, se centran en los valores más altos.

Como vemos en la Figura 20, los alumnos que postergan la carrera eligieron valores centrados en 2 y 3, mientras que los que no postergan la carrera se centran en los valores 4 y 5. Para observar mejor esta distribución, observemos el gráfico de densidad de la Figura 21.

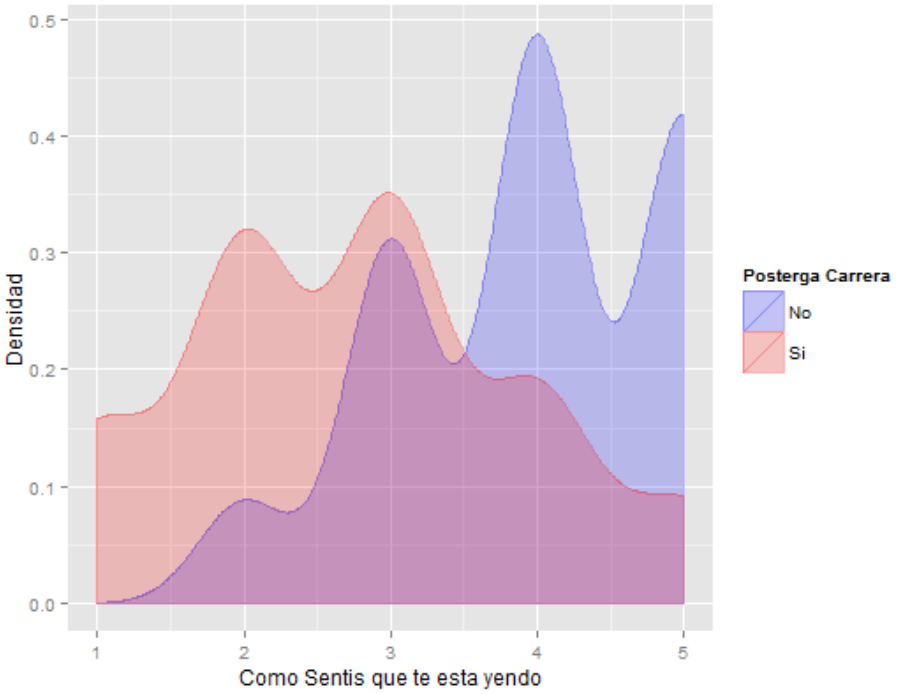


Figura 21: Distribución de valores de como sienten que les va en la carrera a los alumnos que postergan y no, la carrera. Se aprecia con mayor facilidad como los alumnos que postergan la carrera tienen sus modas en los valores más chicos, y ocurre lo inverso para aquellos alumnos que no postergan la carrera.

Los tests de Kolmogorov–Smirnov y Student arrojan indican que la diferencia entre los valores escogidos por los alumnos de ambos grupos son significativos (p $<7.5905e^{-12}$ y p $<2.9938e^{-18}$ respectivamente).

Podemos concluir que los alumnos de ambos grupos (en líneas generales) son concientes de su nivel de éxito académico. Sin embargo, como vimos en el análisis de la sección 2.6, existen algunos alumnos que sienten que les va bien y se encuentran atrasados en la carrera y viceversa. Los motivos de este fenómeno ya fueron analizados anteriormente en dicha sección, cuando comparamos los *features* 'PostergaCarrera' y 'CómoSentísQueTeEstáYendo'.

Por último, en la encuesta les consultamos a los alumnos si alguna vez habían pensado en dejar la carrera. Basándonos en este *feature*, elaboramos el análisis correspondiente para observar cómo se distribuían las respuestas para ambos grupos. Los resultados se observan en la Figura 22

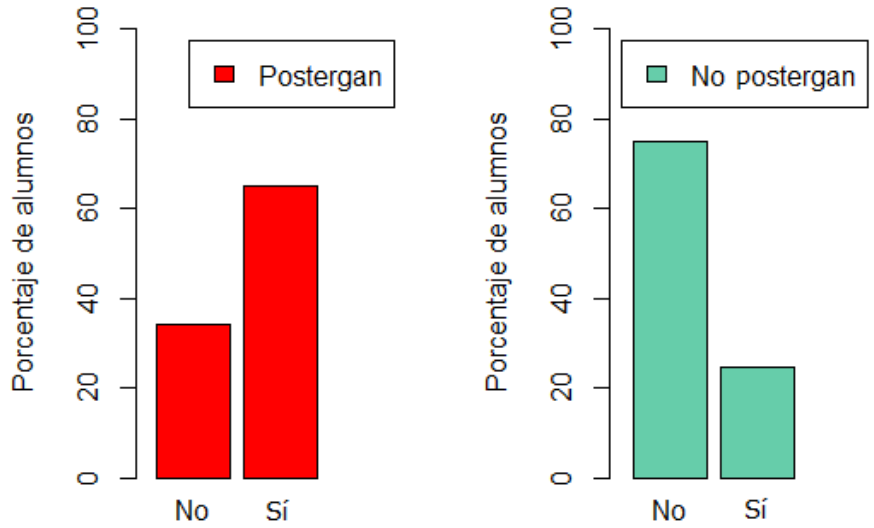


Figura 22: Porcentaje de alumnos que pensaron alguna vez en dejar la carrera para los grupos que postergan y no, la carrera. Pensar dejar la carrera es un razgo que se da con mucha más frecuencia en los alumnos que postergan la carrera que en los que no.

Claramente hay una inclinación a pensar en dejar la carrera para los alumnos que la postergan mientras sucede lo inverso para los que no postergan la carrera.

La diferencia es significativa, cerca del 70 % de los alumnos que postergan la carrera pensó alguna vez en dejarla, mientras que solo un 22 % lo hizo dentro de los alumnos que no postergan la carrera. Los tests de Kolmogorov–Smirnov y Student confirman los resultados (p $<3.7120e^{-10}$ y p $<1.0608e^{-11}$ respectivamente).

En resumen, en esta sección vimos que las becas parecerían ayudar levemente a los alumnos para no postergar la carrera, aunque los tests estadísticos indican que no existen diferencias estadísticamente significativas en este aspecto. Vimos, una vez más, como el saber programar antes de comenzar la carrera, no influye en que los alumnos posterguen la misma tiempo después. Notamos que los alumnos que actualmente trabajan y postergan la carrera representan un porcentaje bastante mayor a sus correspondientes en el grupo de los que no postergan la carrera. Más aún, los alumnos que postergan la carrera suelen tener una carga horaria laboral mayor a los que no, por lo que el trabajo influye de manera directa en el éxito académico de los alumnos.

Vimos como utilizar internet y tener un grupo de trabajos prácticos 'fiel' parece influir positivamente en el éxito académico. El contar con un grupo de trabajo conocido muchas veces facilita la distribución de tareas y acelera los tiempos en que se puede finalizar un trabajo.

Observamos que gran parte de los alumnos que postergan la carrera tardan entre tres y cinco cuatrimestres en completar el CBC, mientras que la mayoría de los que no la postergan indicaron que lo completaron en dos.

Por otro lado, los valores que tomó la variable 'ComoSentisQueTeEstáYendo' para estos dos grupos se correspondía bastante con su situación. Para los alumnos que postergan la carrera, los valores para esta variable se concentraron en dos y tres, mientras que para los que no, se concentraron en los valores tres,

cuatro y cinco.

Por último, vimos como cerca de un 70 % se los alumnos que posterga la carrera indicó que pensó alguna vez en abandonarla, mientras que apenas el 22 % de los alumnos que no postergan la carrera indicó esto mismo. Podemos concluir que un factor emocional como pensar dejar la carrera está correlacionado directamente en el éxito académico, quizás como un efecto colateral. El atrasarse en la carrera suele llevar a plantearse abandonar la misma y quizá esto termine causando que el alumno se atrase aún más.

3. Trabajando con los datos oficiales

En la siguiente sección analizaremos datos oficiales provistos por el Departamento de Alumnos de la Facultad de Ciencias Exactas y Naturales. Los datos contienen principalmente la condición de regularidad de los alumnos (Regular, Libre o Terminó) y las notas de los finales de algunas de las materias de la carrera para cada uno de ellos.

Los datos son anónimos y fueron obtenidos previa solicitud al Departamento de Alumnos. La elección de los datos estuvo sujeta a la disponibilidad y el requerimiento del anonimato de los mismos. Los datos de cada alumno comprenden:

- *FEC_NACIM*: la fecha de nacimiento del alumno.
- *ANY_LIB*: el año en que ingresó a la facultad.
- *COD_SEXO*: Género (M o F).
- *LUG_NACIM*: Lugar de nacimiento, una provincia (sólo tendremos en cuenta los valores Capital Federal y Buenos Aires).
- *CON_CAR*: Condición de regularidad del alumno (L, R o T).

Por otro lado, los datos sobre cada materia comprenden sólo información sobre exámenes finales: las notas en el acta y sus fechas correspondientes para cada alumno que haya realizado el examen. Los datos solicitados corresponden a un subconjunto de materias de la carrera, el cual creemos que es representativo para realizar nuestro análisis. Las materias solicitadas fueron siete: Algebra I, Análisis II, Algoritmos y Estructuras de Datos II, Algoritmos y Estructuras de Datos III, Organización del Computador II, Métodos Numéricos e Ingeniería del Software II.

Usando estos datos nos será interesante analizar principalmente qué factores influyen en la regularidad de los alumnos y qué patrones y conclusiones surgen de la información acerca de las materias. Esta base de datos es sustancialmente mayor a las obtenidas en nuestras encuestas, comprendiendo un total de 2279 casos de alumnos, y correspondiente a los ciclos lectivos desde el año 2000 hasta el 2014.

Similarmente a la encuesta anterior (2) los datos utilizados en esta sección son anónimos y por lo tanto los datos de esta base no pueden ser comparados con los de la encuesta. De lo contrario, estos datos hubieran enriquecido el estudio sobre la posteración de la carrera, permitiendonos asociarlos y analizarlos en conjunto.

En esta sección se realizarán análisis del tipo exploratorio sobre los datos y se aplicarán métodos de Machine Learning para analizar qué factores son predictores fuertes de la condición de regularidad de los alumnos.

3.1. Análisis sobre las materias

Empezaremos por el estudio exploratorio de las materias sobre las cuales tenemos información. Un primer análisis descriptivo indica que los alumnos desaprueban menos los exámenes finales en las últimas materias comparado con las primeras, como se observa en la tabla 6. A su vez, de la tabla se desprende

que la cantidad total de alumnos que rinden las últimas materias es significativamente menor.

Materia	Total Rendidos	Total Desaprobados	Media Notas	Mediana Notas
Algebra I	1707	519	5.193	5
Análisis II	1166	329	5.355	5
Algoritmos y Estructuras de Datos II	992	153	6.606	7
Algoritmos y Estructuras de Datos III	618	6	7.678	8
Métodos Numéricos	504	16	8.032	9
Organización del Computador II	615	2	8.289	9
Ingeniería del Software II	307	0	8.348	8

Cuadro 6: Información descriptiva de las materias

Otro análisis preliminar interesante es la distribución de notas teniendo en cuenta la separación de alumnos entre los que viven en Capital y los que viven en provincia.

Materia	Promedio Capital	Promedio Provincia
Algebra I	5.284	4.984
Análisis II	5.489	5.053
Algoritmos y Estructuras de Datos II	6.811	6.059
Algoritmos y Estructuras de Datos III	7.687	7.660
Métodos Numéricos	8.000	8.148
Organización del Computador II	8.284	8.247
Ingeniería del Software II	8.355	8.351

Cuadro 7: Promedios de notas entre capital y provincia.

Los promedios en la tabla 7 presentan diferencias significativas en las calificaciones correspondientes a las materias Algebra I, Análisis II y Algoritmos y Estructuras de Datos II ($p < 0.0378$, $p < 0.0132$ y $p < 0.00006$ respectivamente). La diferencia muestra que los alumnos provenientes de Capital Federal tienen un mejor promedio en las calificaciones para éstas materias.

En cuanto a las materias más avanzadas en el plan de estudios (como, por ejemplo, Ingenieria del Software II) no se encuentran diferencias significativas en los promedios. Esto último indica que la zona de donde el alumno proviene deja de ser relevante (y por consiguiente la educación secundaria recibida, también).

3.1.1. Análisis por género

Resulta interesante realizar un estudio basado en el género de los alumnos. Este tipo de estudios de género (en particular para el femenino) son interesantes en la actualidad para el área de informática. Si bien trabajaremos con los datos anónimos, abordaremos con mayor detalle lo que sucede con los alumnos de género femenino de nuestra carrera en la sección 4. Los datos actuales comprenden un total de 1962 alumnos de género masculino y 317 de género femenino.

Realizaremos, como primer análisis, una comparación entre las notas de cada materia distinguiendo a los alumnos por género.

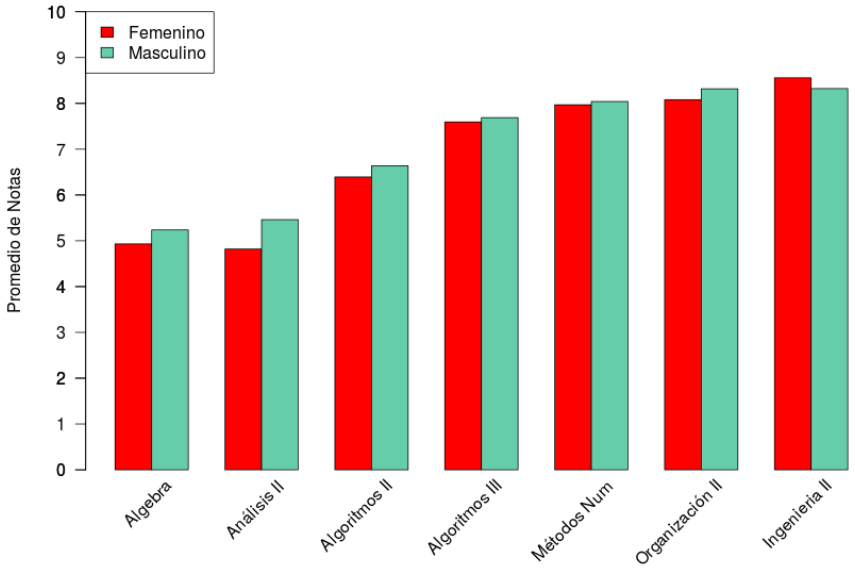


Figura 23: Promedio de notas para cada materia según el género. Pese a que se observe que el promedio de los hombres es mayor en general, la única diferencia significativa está en Analisis II.

En la figura 23 se observa una leve tendencia a que el promedio de los alumnos de género masculino sea mayor. Sin embargo, según los tests implementados, la única materia en dónde la diferencia es significativa corresponde a Análisis II ($p < 0.0017$ para Student y $p < 0.0317$ para Kolmogorov–Smirnov). Estudiaremos entonces esta materia en detalle.

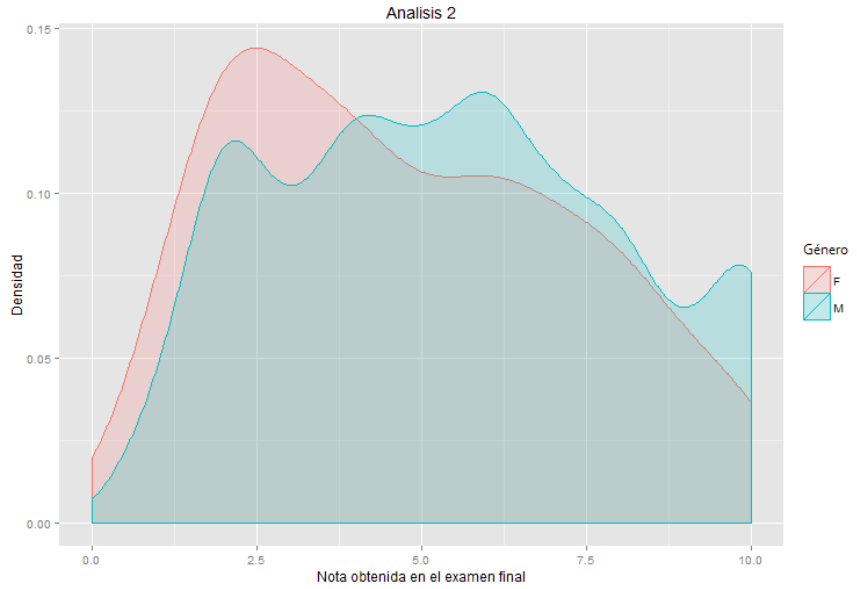


Figura 24: Densidad de notas para Análisis II según el género. Se observa que las mujeres tienden a recibir notas más bajas en esta materia.

En la figura 24 observamos que las mujeres tienden a recibir calificaciones menores en Análisis II con respecto a los varones. La mayor concentración de notas se encuentra por debajo del cuatro, lo que indica que la mayoria de las mujeres desaprueba el examen final. Por otro lado, los alumnos de género masculino presentan la mayor concentración de calificaciones por encima del cuatro.

Otro análisis interesante de realizar es poder saber si existe un crecimiento o decrecimiento en la cantidad de mujeres que inscriben en la carrera de Ciencias de la Computación con el paso de los años. Con este objetivo, realizamos un relevamiento de la cantidad de inscripciones para cada año. Como los datos se tomaron durante el ciclo lectivo 2014, se decidió omitir este año ya que pudieron haber existido más inscripciones durante el mismo y que éstas no se hayan llegado a plasmar en nuestros datos.

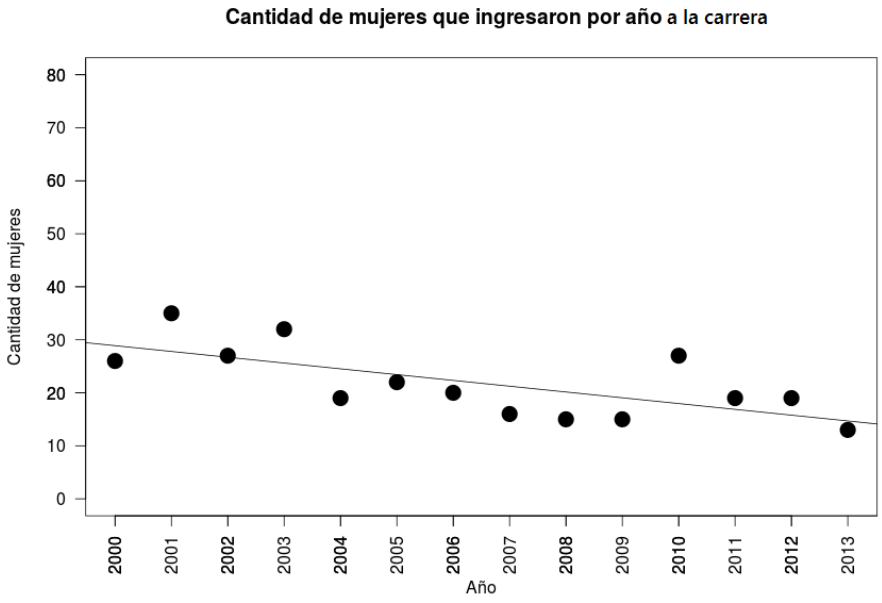


Figura 25: Cantidad de inscripciones de mujeres por año, comprendiendo desde el año 2000 al 2013. La linea de regresión muestra un decrecimiento en la cantidad de inscripciones.

En la Figura 25 se observa con una linea de regresión que entre el 2000 al 2013 existió un leve decrecimiento en la cantidad de inscripciones de mujeres. En promedio, se inscriben 22 mujeres por año, una taza que creemos por demás baja.

3.1.2. Plan de estudios

Como mencionamos, solicitamos información sobre un subconjunto de materias de la carrera. Podemos entonces comparar la propuesta de la facultad en cuanto al orden y tiempo en cuales cursar las materias (dado por el plan de estudios oficial de la carrera) con los datos reales (el orden y tiempo en que suelen cursar los alumnos). Según el plan de estudios, el orden en cual rendir las materias cuyos datos solicitamos es:

- Primer cuatrimestre: Álgebra I y Análisis II
- Tercer cuatrimestre: Métodos Numéricos
- Cuarto cuatrimestre: Organización del Computador II y Algoritmos y Estructuras de Datos II
- Quinto cuatrimestre: Algoritmos y Estructuras de Datos III
- Octavo cuatrimestre: Ingenieria del Software II

El árbol completo de correlatividades puede verse en el anexo 10. La carrera propone un límite de ocho cuatrimestres entre que se aprueba la cursada de una materia hasta que se apruebe el final correspondiente. Se puede suponer que el

plan de estudios pretende que los alumnos rindan el final el mismo cuatrimestre en que se aprobó la cursada.

Un estudio interesante es observar la diferencia, medida en cantidad de cuatrimestres, entre cuándo se espera que se rinda un final según el plan de estudios y cuándo lo rinden los alumnos según los datos obtenidos.

Materia	Cuatrimestre esperado	Cuatrimestre promedio
Algebra I	1	1.9
Análisis II	1	3.2
Métodos Numéricos	3	8.7
Organización del Computador II	4	6.9
Algoritmos y Estructuras de Datos II	4	5.2
Algoritmos y Estructuras de Datos III	5	7.1
Ingeniería del Software II	8	11.7

Cuadro 8: Comparación entre el cuatrimestre en el que se espera que se rinda el examen final de una materia según el plan de estudios y el promedio en que los alumnos suelen rendirlo.

El cuadro 8 muestra el cuatrimestre en que se espera que el alumno rinda un final según el plan de estudios y el promedio de cuatrimestres en que los alumnos rinden el final según nuestros datos. Se observa que todos los promedios superan al valor pretendido por el plan de estudios. Esto puede indicar que el plan de estudios es muy optimista en cuanto a sus expectativas. En principio, los alumnos tienen la elección de cursar las materias en el orden que ellos elijan, sin respetar el plan de estudios (siempre y cuando se respeten las correlatividades). Es posible que esto explique la diferencia observada en materias como Métodos Numéricos (que no tiene ninguna materia obligatoria que dependa de ella). Los alumnos puede optar rendir la materia más adelante en le carrera y no en el tercer cuatrimestre, como indica el plan de estudios.

Por otro lado, se puede observar la diferencia entre las fechas en las cuales se rindieron dos materias. Analizaremos esta diferencia medida en cuatrimestres. Casos interesantes incluyen a Álgebra I (la cual suponemos que es la primer materia que se rinde, en promedio) contra los casos de Ingeniería del Software II, Organización del Computador II y Métodos Númericos.

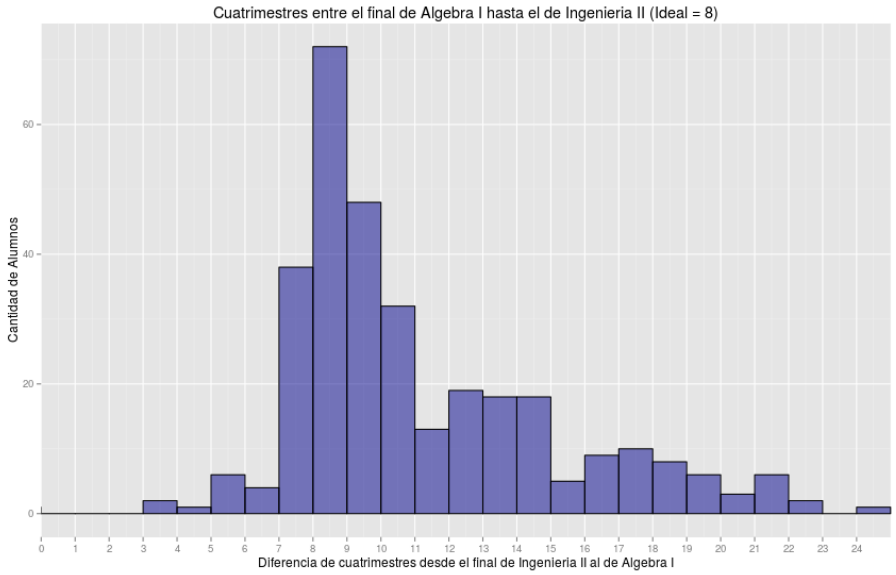


Figura 26: Diferencia medida en cuatrimestres entre el final de Ingeniería del Software II y Álgebra I. La mayoría respeta el tiempo ideal, pero hay muchos casos que tardan más de 10 cuatrimestres (5 años) en rendir Ingenieria II luego de rendir Algebra I.

En la figura 26 se observa que la mayoría elige rendir el final de Ingeniería del Software II ocho cuatrimestres después que el de Algebra I, como sugiere el plan de estudios. Sin embargo, existen muchos casos que lo hacen luego de los 10 cuatrimestres, incluso una llamativa cantidad que lo rinde luego de los 16 cuatrimestres (8 años de diferencia). Este último caso puede indicar alumnos que postergan la carrera, o abandonan para luego retomar en un futuro.

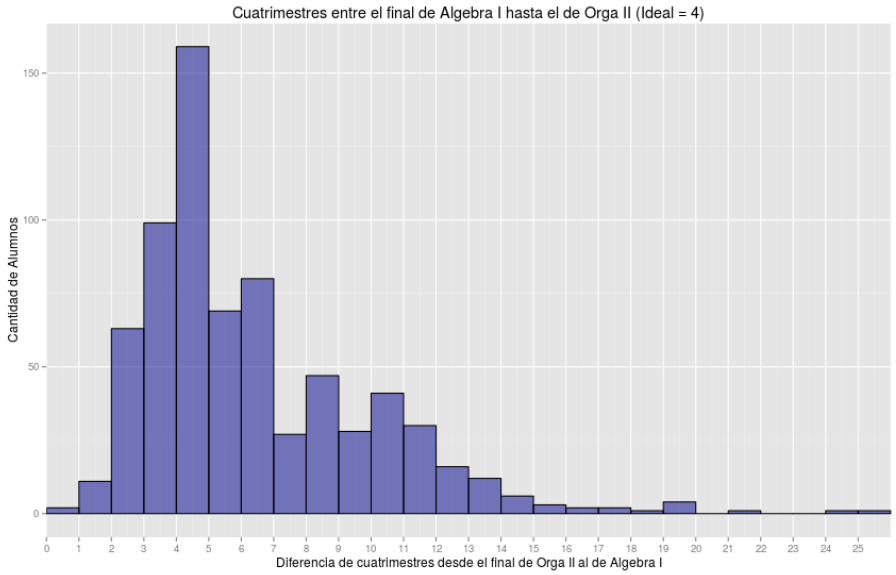


Figura 27: Diferencia medida en cuatrimestres entre el final de Organización del Computador II y Álgebra I. Se respeta el tiempo de cuatro cuatrimentres estimados entre los finales, aunque se observan numerosos casos que tardan seis, ocho o diez cuatrimestres.

Las diferencias entre Organización del Computador II y Algebra I mostradas en la figura 27 muestran que, como en el caso anterior, la mayoría rinde el final en el tiempo esperado de cuatro cuatrimestres de diferencia. Existe otro valor de relevancia a los seis cuatrimestres, otros a los ocho y a los diez. Durante varios cuatrimestres (muchos de los cuales fueron incluidos en este estudio), la materia Organización del Computador II ofreció como opción al final teórico, un trabajo práctico final. Estos casos en los que el final se rinde fuera de los términos pretendidos por el plan de estudios, se pueden deber a la postergación de la carrera de los alumnos que estudiamos en la sección 2. Sin embargo, el fenómeno también puede deberse al hecho de haber optado por realizar el trabajo práctico final y que éste lleve más tiempo del estimado por el alumno. En este análisis no podremos saber cual es el factor que influye en este fenómeno ya que no contamos con la información necesaria en los datos brindados por el Departamento de Alumnos. De todos modos, sería de interés comparar esta misma diferencia distinguiendo a los alumnos que rindieron el exámen teórico y los que eligieron el trabajo práctico, de contar con este dato, en algún estudio a futuro.

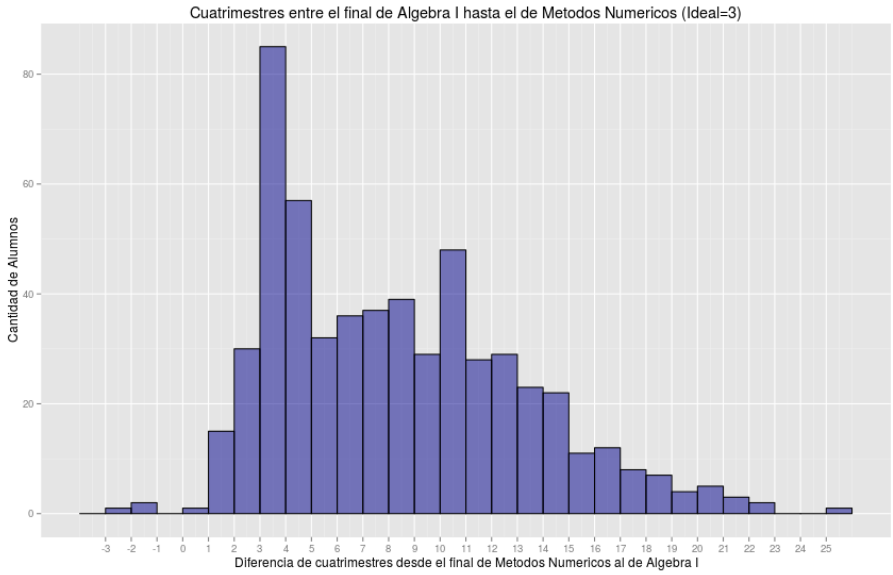


Figura 28: Diferencia medida en cuatrimestres entre el final de Métodos Numéricos y Álgebra I. En este caso observamos que, aunque la mayoría respeta el tiempo ideal de tres cuatrimestres, hay otro foco acumulado importante a los diez cuatrimestres. Los valores negativos indican casos en los cuales el final de Métodos Numéricos se rindió antes que el de Algebra I.

La materia Métodos Númericos debería rendirse idealmente tres cuatrimestres después de Algebra I. En la figura 28, observamos que el punto de mayor concentración de alumnos se encuentra dentro de este parámetro. Sin embargo, la cantidad de gente que elige rendir el final de Métodos Numéricos más adelante en la carrera, los supera en número. Un pico llamativo aparece a los diez cuatrimestres (5 años de diferencia). Como mencionamos antes, en el programa de la carrera, sólo materias optativas son correlativas a Métodos Numéricos. Esto puede explicar por qué la materia se suele postergar hasta el final de la carrera, para concentrar el tiempo en finales de otras materias que tienen como correlativas otras materias obligatorias.

3.1.3. Orden en que se rinden los finales

Contando con la información acerca de las fechas en que fueron rendidos los exámenes finales, se puede realizar un estudio para conocer el orden en que los alumnos suelen rendir los exámenes finales de las distintas materias, y cuales anteceden a otras. Pudimos realizar un seguimiento del orden en que fueron rendidas las siete materias de las cuales tenemos datos, dando como resultado los caminos más comunes elegidos por los alumnos.

								Frecuencia
1	ALGI	–	–	–	–	–	–	201
2	ANAII	ALGI	–	–	–	–	–	73
3	ALGI	ANAII	–	–	–	–	–	51
4	ALGI	ALGOII	–	–	–	–	–	43
5	ANAII	ALGI	ALGOII	–	–	–	–	32
6	ANAII	ALGI	ALGOII	ALGOIII	ORGAII	MET NUM	INGII	30
7	ANAII	ALGI	ALGOII	MET NUM	ALGOIII	ORGAII	INGII	29
8	ANAII	ALGI	ALGOII	MET NUM	ORGAII	ALGOIII	INGII	26
9	ANAII	ALGI	ALGOII	ALGOIII	MET NUM	ORGAII	INGII	24
10	ALGI	ALGOII	ANAII		–	–	–	21

Cuadro 9: Los diez órdenes más frecuentes en que los alumnos rinden las materias, sin requerir que se hayan rendido todas. El camino más frecuente es haber rendido solo Algebra I, seguido por haber rendido ambas Análisis II y Algebra I en algún orden. No es frecuente haber rendido sólo Análisis II. Todos los caminos más frecuentes que incluyen las siete materias empezaron con Análisis II.

Para realizar este análisis, consideraremos por cada alumno, el orden en cual se rindieron las diferentes materias y contaremos la frecuencia con la que se elije cada posible camino. También, tendremos en cuenta los casos en los que puede que no se hayan rendido el final de las siete materias.

En el cuadro 9 se observa que el caso más frecuente es haber rendido solo Álgebra I, seguido por haber rendido sólo Álgebra I y Análisis II en cualquier orden. Esto último indica que lo más común es intentar rendir las dos primeras materias antes de avanzar por alguna de las ramas de la carrera (de todas formas, es posible que se hayan rendido materias en el medio, de las cuales no tenemos información). Notamos que entre los caminos más frecuentes no se encuentra el caso en que se rindió sólo Análisis II.

Los alumnos que cumplan haber rendido todas las materias presentadas son casos que se encuentran en el final de la carrera, o ya recibidos. Los órdenes más frecuentes, considerando que rindieron las siete materias, se muestras en el cuadro 10. En estos casos cabe destacar que Análisis II se rinde entre las primeras materias, siempre antes que Algoritmos III, y junto a Álgebra I (a excepción de uno de los caminos más frecuentes). Una creencia extendida por el alumnado es que Análisis II es una materia que, por su dificultad, es rendida comúnmente al final de la carrera. Sin embargo, nuestros datos contradicen esta creencia ya que los caminos más frecuentes analizados, la ubican entre las primeras que se rinden.

Estudiando los datos de Análisis II con más detalle encontramos que casi un tercio del alumnado (31.78 %) la posterga más de dos cuatrimestres. El promedio en cuatrimestres para aprobar la materia de ese tercio del alumnado, es 7.25. En otras palabras, rinden casi 4 años después de cursar, justo antes de que se terminen las instancias para rendir el final. Entonces, la sensación de que la materia es postergada, proviene de que los alumnos que la postergan lo hacen por mucho tiempo. Este efecto se propaga hacia las materias correlativas haciendo que Métodos Numéricos se rinda en promedio 9.93 cuatrimestres luego de anotarse en la carrera, mucho mayor a los 3 cuatrimestres esperados.

Destacamos también a los alumnos que aprueban Ingeniería del Software II, siendo ésta la última materia a aprobar en la mayoría de los casos. Encontramos que el 60.58 % tarda en promedio 14.08 cuatrimestres en aprobar la materia

comparado con los 8 cuatrimestres esperados.

								Frecuencia
1	ANAII	ALGI	ALGOII	ALGOIII	ORGAII	MET NUM	INGII	30
2	ANAII	ALGI	ALGOII	MET NUM	ALGOIII	ORGAII	INGII	29
3	ANAII	ALGI	ALGOII	MET NUM	ORGAII	ALGOIII	INGII	26
4	ANAII	ALGI	ALGOII	ALGOIII	MET NUM	ORGAII	INGII	24
5	ANAII	ALGI	ALGOII	ORGAII	ALGOIII	MET NUM	INGII	13
6	ALGI	ANAII	ALGOII	ALGOIII	MET NUM	ORGAII	INGII	8
7	ALGI	ALGOII	ANAII	ALGOIII	ORGAII	MET NUM	INGII	8
8	ANAII	ALGI	MET NUM	ALGOII	ALGOIII	ORGAII	INGII	7
9	ALGI	ANAII	ALGOII	ALGOIII	ORGAII	MET NUM	INGII	7
10	ANAII	ALGI	ALGOII	ALGOIII	ORGAII	INGII	MET NUM	6

Cuadro 10: Los diez ordenes más frecuentes en que los alumnos rinden las materias, requiriendo que se hayan rendido todas. Según el plan de estudios, de las materias presentadas, Métodos Númericos debería ser la tercer materia en rendirse. Éste no es el caso en la mayoría de los ordenamientos, los alumnos prefieren postergar Métodos. Ninguno de los caminos más frecuentes respeta el orden planteado por el plan de estudios.

Otra materia destacable es Métodos Númericos. Según el plan de estudios y las materias que consideramos, debería ser la tercera en rendirse. A excepción del camino 8, esto no ocurre. Se suele postergar hasta despúes de Algoritmos II e incluso hasta después de Algoritmos III. En el cuadro 10 se observa que es frecuente rendirla justo antes que Ingenieria II y en uno de los caminos, luego de ésta. Esto podría estar indicando que la materia resulta más complicada que lo previsto por plan de estudios.

Por último, se observa que Organización del Computador II suele postergarse ya que debería rendirse antes de Algoritmos y Estructuras de Datos III según el plan de estudios y esto no suele ocurrir en los caminos del cuadro 10. Según éste, es una de las materias que suele rendirse al final, junto con Ingeniería II.

3.2. Acerca del estado de regularidad de los alumnos

Los datos obtenidos del Departamento de Alumnos nos permiten conocer el estado de regularidad de cada alumno. El estado de regularidad se refiere a la condición del alumno frente a la universidad. Este último se divide entre regulares, libres y alumnos que terminaron la carrera. En la base de datos de la facultad (y en nuestro análisis) estos estados son referidos como R, L y T respectivamente. Es importante destacar que un alumno que está en condición de libre si no cumple los requisitos obligatorios para mantener la regularidad como no haber completado los censos obligatorios que propone la facultad o no haber votado, caso contrario puede figurar en estado regular pese a no estar del todo 'activo'.

En las siguientes secciones estudiaremos la distribución y comportamiento del estado de regularidad de los alumnos. Veremos qué variables son características de cada grupo y aplicaremos técnicas de Machine Learning para construir un predictor del estado de un alumno dado. Dicho predictor puede ser util para conocer las características particulares de cada estado de regularidad.

Los grupos se dividen en:

- *Regulares (R)*: 1486 alumnos (65.20 %).

- *Libres (L)*: 637 alumnos (27.95 %).
- *Terminaron (T)*: 156 alumnos (6.84 %).

Como mencionamos anteriormente, los datos comprenden a los alumnos que iniciaron la facultad entre el año 2000 y 2014. Sin embargo, los años de ingreso con los que nos interesa trabajar comprenden el período 2000 al 2010, ya que en este último año se registra la última persona recibida. A raíz del simple conteo de alumnos en cada condición se observa, en principio, que la cantidad de alumnos recibidos es muy baja y el índice de deserción es alto (gran cantidad de alumnos libres). En cuanto a los alumnos que figuran como regulares, si bien no figuran formalmente como libres, es posible que mantengan la regularidad cumpliendo los requisitos mínimos (votaciones o censos obligatorios) pero que no estén actualmente dedicándole tiempo a la facultad. A partir de los datos que poseemos es imposible saber ésto con certeza, y la universidad tampoco puede saberlo oficialmente. Analizaremos, en secciones posteriores y utilizando nuestro predictor, qué alumnos regulares tienen características similares a los libres.

3.2.1. Relación de las materias con la condición del alumno

Es posible realizar un estudio del estado de los alumnos (L, R o T) y comparar las notas y materias agrupandolos bajo esa condición. El objetivo es encontrar características y propiedades que surjan de la comparación entre los diferentes grupos, y la existecia de variables indicativas que los diferencien. A priori, algunas hipótesis posibles podrían ser que las notas tienen relevancia en la condición del alumno y que existe alguna materia, que luego de rendido su examen final, la probabilidad de que un alumno quede libre es baja. Nos encargaremos de ver si esto efectivamente se cumple.

Se consideró calcular, para cada materia, cuál es la probabilidad de que el alumno quede libre luego de rendirla. Para este cálculo se incluyeron los casos en los cuales el alumno no rindió ninguna materia antes de quedar libre.

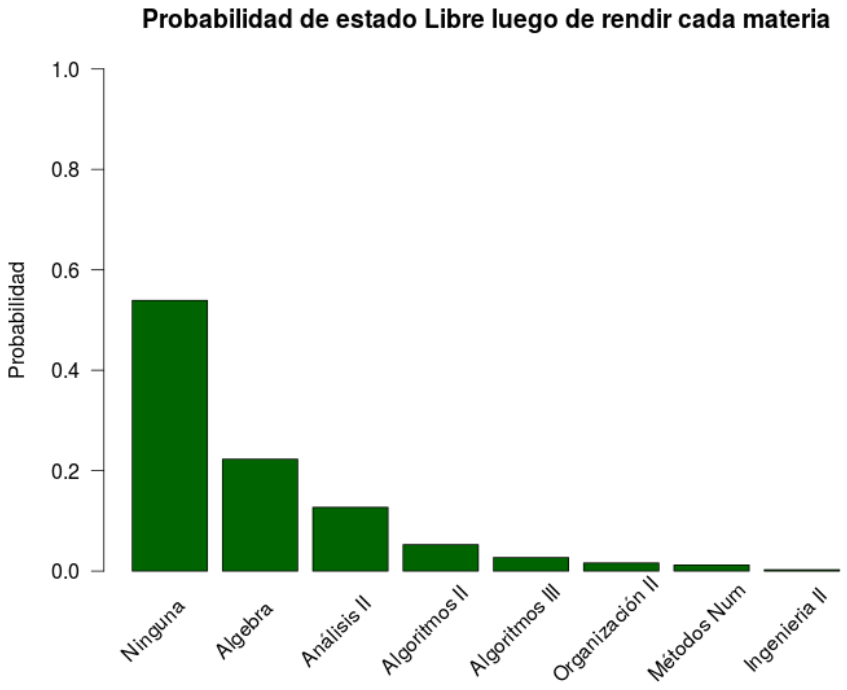


Figura 29: Probabilidad de que un alumno quede libre luego de rendir cada materia. Más de la mitad de la densidad está en los casos en que el alumno queda libre sin aprobar ninguna materia. Luego de aprobada alguna materia, la probabilidad de quedarse libre disminuye notablemente. Luego de aprobar Algoritmos III, la densidad acumulada de los demas casos es menor a 0.1.

Se observa en la Figura 29 que el momento más probable en que un alumno quede libre es cuando no cuenta con ningún examen final aprobado. Es importante entonces poder brindar ayuda y apoyo a los alumnos en esta etapa para evitar una posible deserción de este grupo. Luego de aprobadas Álgebra I y Análisis II, la densidad de la distribución decrece rapidamente. Más aún, luego de aprobar Algoritmos III, es improbable que un alumno quede libre. Una prueba de esto último es que la suma de las probabilidades para Organización del Computador II, Métodos Numéricos e Ingeniería del Software II que se observan en la Figura 29 es menor a 0.1.

Similarmente, contando con la información de cuál es la última materia aprobada de cada alumno y la condición en la carrera del mismo, pudimos estimar el promedio que tardan los alumnos que se recibieron en completar todos los exámenes finales de las materias solicitadas. En promedio, desde que empiezan a cursar hasta que rinden el último examen, los alumnos tardan 5.29 años. No contamos con la información de la fecha exacta en la cual se recibió el alumno, por lo que el cálculo que realizamos corresponde sólo la fecha en que los alumnos terminan de aprobar el examen final de todas las materias. Esto implica que al promedio de años anteriormente calculado, hay que sumarle el tiempo en que se tardaría en hacer la tesis de licenciatura y las materias optativas. De todas formas, este tiempo promedio calculado (5.29 años) supera ampliamente al estimado por el plan de estudios para rendir los examenes finales de todas las

materias (4 años), como también al tiempo que se estima para rendir los finales de las materias optativas y finalizar la tesis de licenciatura (5 años).

Otra comparación interesante que podemos realizar es el promedio de notas de cada materia entre los alumnos que tienen estado Libre y Terminó.

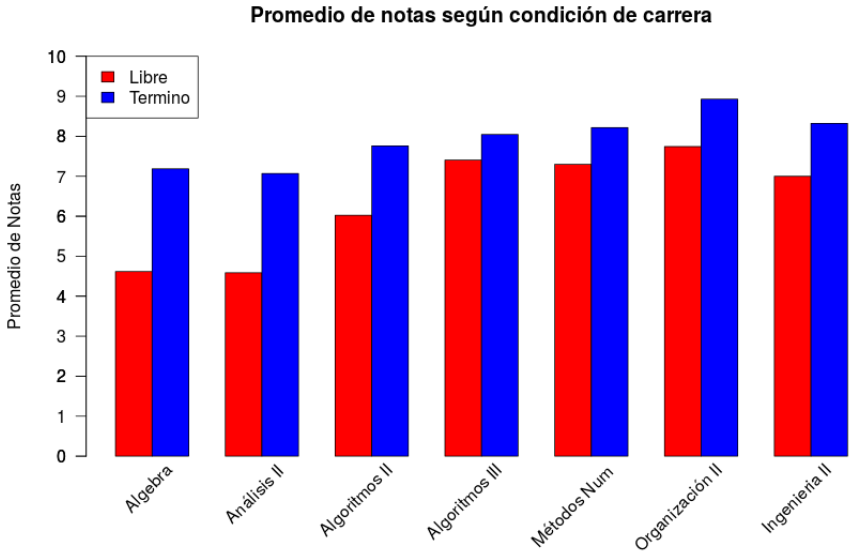
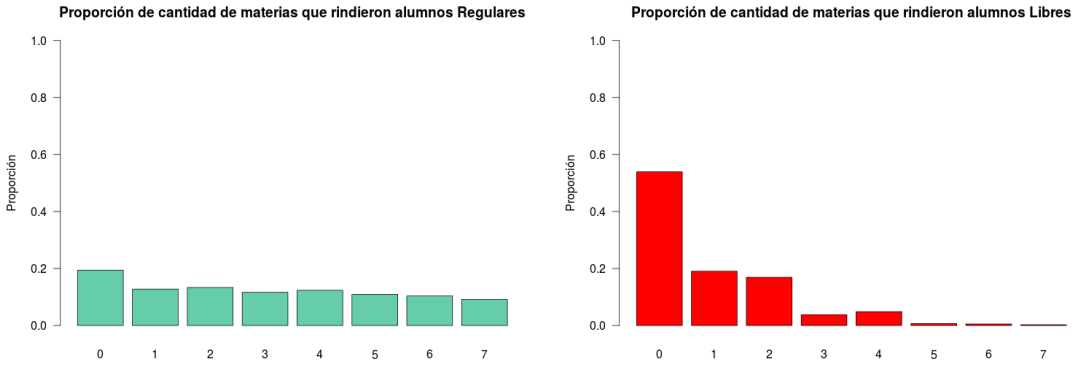


Figura 30: Promedio de notas para cada materia según la condición de la carrera. Las notas de los alumnos libres son en promedio menores a las de los alumnos que terminaron la carrera. Las diferencias son notables en las primeras tres materias.

En la Figura 30 se observa que, en promedio, las notas de los alumnos con estado Libre son menores a los alumnos con estado Terminó. En las materias Análisis II, Álgebra I y Algoritmos II, la diferencia es sustancial. Esto puede ser indicativo que la deserción esté relacionada con el rendimiento del alumno en estas primeras materias. Es importante, entonces, que se acompañe al alumno y se lo incentive durante estas etapas iniciales. En cuanto a las demás materias, más avanzadas en la carrera, no se observan diferencias importantes. Es posible, entonces, que el alumno quede libre en estos casos por otras razones ajenas a la nota que recibe en estas materias. El caso de la materia Ingeniería II la cantidad de alumnos que la rindieron no son suficientes para que la diferencia sea significativa.



(a) Cantidad de materias que rindieron alumnos regulares.

(b) Cantidad de materias que rindieron alumnos libres.

Figura 31: Proporción de la cantidad de materias que rinden los alumnos regulares y libres. Los alumnos regulares estan mayormente distribuidos equitativamente. Más de la mitad de los alumnos libres no aprobaron ninguna materia.

Se contó la cantidad de exámenes finales que cada alumno rindió en total para observar su distribución. Este análisis tiene sentido sólo para las condiciones de Libre y Regular, ya que los alumnos con condición Terminó, cuentan con todos los exámenes finales ya aprobados. En la Figura 31a, se observa que los alumnos regulares se distribuyen equitativamente a grandes rasgos, esto indica que hay alumnos regulares en todas las etapas de la carrera. Hay levemente mayor proporción entre los alumnos que no rindieron ningún final todavía para ambos grupos. Esto puede ser explicado debido a la presencia de alumnos con año de ingreso en 2014, los cuales son nuevos y probablemente no hayan rendido ningún examen final a la fecha.

En cuanto a los alumnos libres, en la Figura 31b, se observa que más de la mitad de ellos jamás rindió ningún final. Más aún, entre los alumnos que rindieron una o dos materias, suman el 40 % aproximadamente de la proporción. Estos datos se condicen con los resultados que obtuvimos anteriormente, los alumnos libres suelen presentarse sobre el principio de la carrera y no llegan a rendir demasiados exámenes.

3.2.2. Prediciendo la condición en la carrera

En esta sección nos enfocaremos en la condición de la carrera en sí y qué variables son más relevantes en cada estado. Buscaremos introducir un método predictivo que nos provea, con algún nivel de certeza, la condición de la carrera de cada alumno. El objetivo que buscamos es poder generar un predictor que clasifique nuestros datos en tres clases: L, R y T. Si bien se está prediciendo el estado de un alumno, esto, en sí, no es del todo interesante, ya que bastaría con observar el estado real del alumno en el sistema de inscripciones. Sin embargo, si un alumno con estado actual Regular es clasificado como Libre o Terminó por nuestro predictor, esto puede ser indicativo de que ese alumno está cercano a cambiar su condición de regularidad. Esto sí es interesante desde el punto de vista de combatir la deserción o incentivar a los alumnos cercanos a recibirse.

3.2.3. Métodos utilizados

Para esta sección del trabajo, se utilizó el paquete `scikit-learn` [15] de Python. Este paquete se especializa en brindar herramientas para el análisis de datos y Machine Learning que utilizaremos posteriormente. Se elaboró un script en Python que, a partir de la base de datos actual, preprocesa los datos, genera nuevas features, las procesa y ejecuta automáticamente varios métodos predictivos para que después se pueda analizar cuáles fueron los de mejor rendimiento (definiremos una métrica para medir el rendimiento más adelante). Dentro de los métodos testeados se encuentran Random Forests, Gradient Boosting, SVM, Decision Trees, Perceptrones, AdaBoost entre otros. Fueron elegidos dos métodos, Random Forests y Gradient Boosting, por ser los de mejor rendimiento sobre el set de datos generado.

3.2.4. Random Forests

Random Forests es un método dentro de los llamados métodos de ensamble, los cuales combinan varios algoritmos de aprendizaje para poder lograr un mejor rendimiento de la clasificación. En el caso de Random Forests, tienen como base árboles de decisión. Estos árboles dividen los datos de acuerdo a cada predictor, para poder llegar a una clasificación en las hojas. Random Forests genera muchos árboles de decisión, cada uno limitado a un subconjunto aleatorio de los predictores originales. Luego, por cada caso en el set de datos, cada árbol propone una clasificación. Estas clasificaciones sirven como votos para poder llegar a la clasificación final, que estará dada por la mayoría de todos los votos. Este método suele asegurar una mejora en el rendimiento de la predicción reduciendo la varianza total manteniendo el sesgo.

En general es un método que funciona bien sin tener que ajustar mucho sus parámetros [10] y tiene la ventaja de poder brindar información acerca de cómo se decide la clasificación. Esto es interesante en nuestro caso ya que podremos descubrir que variables nos interesan del set de datos observando los árboles de decisión generados.

3.2.5. Gradient Boosting

El método de Gradient Boosting [11] toma un enfoque diferente al de Random Forest. Este último intentaba mejorar el modelo a partir de la votación de modelos más débiles (árboles de decisión), pero Gradient Boosting utiliza una estrategia constructiva tratando de mejorar el modelo paso a paso. La idea principal es que intenta iterativamente entrenar los modelos débiles, buscando que cada nuevo modelo mejore la precisión de todo el conjunto. Para lograr la mejora se busca agregar un modelo que maximize el gradiente negativo de la función de pérdida, donde la función de pérdida puede ser elegida para el problema en particular de entre varias opciones clásicas (como por ejemplo normas L1 y L2, binomial, entre otras). En nuestro caso particular se utiliza una regresión logística como función de pérdida.

Es un método que, en general, obtiene buenos resultados pero a diferencia de Random Forests, tiene menor interpretabilidad, tornándose en un método de caja negra en la práctica.

3.2.6. Procesamiento de los datos del Departamento de Alumnos

Para construir los clasificadores, podemos usar como features los campos que nos presentan los datos que obtuvimos del Departamento de Alumnos. Estos campos son generales a todos los alumnos (como vimos al principio de la sección 3) *FEC_NACIM*, *ANY_LIB*, *COD_SEXO* y *LUG_NACIM* y los provenientes de los datos de las materias *FEC_ACT_RE* y *NOT_EXAM* (fecha y nota de cada exámen respectivamente). Separaremos del set de datos a la feature *CON_CAR*, ya que es la que queremos predecir.

A partir de estas features se pueden armar otras nuevas. Esto resulta útil para los clasificadores ya que se les proporciona mayor información sobre el dominio, y para el análisis de los resultados ya que ayuda a la interpretación. Las features que agregamos son:

- *DESAPROBO_X*: Donde *X* refiere a alguna de las siete materias solicitadas al Departamento de Alumnos. Esta feature binaria indica si el alumno desaprobó el final de la materia X. Hay 7 features de este tipo, una por materia.
- *CANT_MATERIAS_APROBADAS*: Esta feature indica la cantidad de materias aprobadas por el alumno.
- *DIF_FEC_X_FEC_Y*: Con *X* e *Y* son materias distintas. Cuenta la diferencia entre las fechas de dos examenes finales medida en años. Cada par de examenes aparece una sola vez. En otras palabras, si existe el feature *DIF_FEC_X_FEC_Y* entonces no se incluye el *DIF_FEC_Y_FEC_X*. Hay 21 features de este tipo.
- *DIF_NACIM_FEC_X*: Diferencia entre la fecha de nacimiento del alumno y cada materia, medida en años. Hay 7 features de este tipo, una por materia.
- *DIF_ANY_LIB_FEC_X*: Diferencia entre el año de inscripción del alumno (año de la libreta) y cada materia, medida en años. Hay 7 features de este tipo, una por materia.
- *EDAD_INSCRIPCION*: Edad con la que se inscribió el alumno a la carrera. Se calculó a partir de la diferencia en años entre la fecha de nacimiento del alumno y el año que figura en su libreta universitaria.

Se agregaron un total de 44 features nuevas que, sumadas a las 17 features originales, hacen un total de 61 features para utilizar en nuestros clasificadores. Un detalle no menor, a tener en cuenta, son los datos faltantes. Existen features que son combinaciones de dos materias, pero no todos los alumnos rindieron los examenes finales de todas las materias. Corresponde asignar un valor a estos datos faltantes para poder generar los clasificadores. Se decidió asignar a las notas y fechas faltantes el valor -1.

El paquete `scikit-learn` que utilizamos para todo lo referido a esta sección, facilita el preprocesamiento del set de datos usando `Pipelines`. Esto permite agregar y modificar facilmente las features que va a tener nuestro set de datos y poder experimentar con diferentes combinaciones de features y procesos.

Listing 3: Pipeline

```
1
2  Pipeline([
3      ('preprocess', Preprocessor()),
4      ('desaprobado', DesaproboMateria()),
5      ('cant_aprobo', ContarMateriasAprobadas()),
6      ('features', FeatureUnion([
7          ('iden', IdentityTransform()),
8          ('dif_fec_mat', DiferenciaFechasMaterias()),
9          ('dif_fec_nacim', DiferenciaMateriasFechaNacim()),
10         ('dif_any_lib', DiferenciaMateriasAnyLib()),
11         ('edad_inscripcion', EdadInscripcion()),
12     ])),
13     ('dropCols', DropCols()),
14     ('scaler', StandardScaler()),
15 ])
```

En el Listing 3 se encuentra el Pipeline que utilizamos en nuestro procesamiento. El mismo, ejecuta en forma secuencial todas las funciones definidas, empezando por `Preprocessor()` que transforma todos los datos originales a un formato adecuado (reemplazando strings por números y rellenando datos faltantes). Luego, dentro de `FeatureUnion()` se agregan todas las funciones de las cuales se espera que generen features nuevas. Finalmente, `DropCols()` filtra ciertas columnas que no son útiles (como ser ID del alumno) y `StandardScaler()` realiza lo que se conoce como Feature Scaling, normalizar las features para que tengan media cero y varianza unitaria. Es un proceso común en machine learning y es particularmente útil en la práctica para nuestro Gradient Boosting.

3.2.7. Fase experimental

Una vez procesado el conjunto de datos y generadas las nuevas features, se puede utilizar para probar diferentes métodos de machine learning sobre él. La biblioteca `scikit-learn` ofrece varias alternativas tanto para regresión como clasificación para aprendizaje supervisado, con la facilidad de que todos los métodos cuentan con la misma interfaz. Los métodos sólo necesitan como entrada el conjunto de datos y el vector de respuesta para entrenar y predecir (el cual en nuestro caso, dicho vector es la condición de carrera *CON_CAR*) sumado a los parámetros propios que pueda llegar a necesitar cada método.

Considerando que se puede tener una misma interfaz para todos los métodos, se creó un script que ejecuta varios métodos elegidos sobre el mismo set de datos, para poder observar cual es el que tiene mejor rendimiento para nuestro problema.

Dentro del aprendizaje supervisado, es una práctica común separar el conjunto de datos en dos partes: set de entrenamiento y set de test. El objetivo de esta separación es poder reducir el overfitting al entrenar cada método. El ovefitting ocurre cuando un modelo resulta muy específico a los datos con los cuales se entrenó y empieza a seguir el ruido o errores propios de los datos, en vez de las relaciones reales entre las variables. Este hecho ocasiona que el modelo no se ajuste bien a la realidad del problema. Al separar en dos el conjunto de datos se puede entrenar el modelo sólo con uno de ellos mientras que el otro es utilizado únicamente con el fin de observar los resultados, simulando el rendimiento del

modelo en un caso real. Nuestra elección para la separación fue generar el conjunto de entrenamiento y test en base a filas aleatorias del conjunto de datos original, con 33 % de los datos totales para el conjunto de test y 66 % para el de entrenamiento.

El proceso consiste en intentar todos los métodos entrenándolos con el set de entrenamiento (utilizando *cross-validation*), y luego predecir el set de test. En el momento de entrenamiento, se necesita que el set de datos esté clasificado (en nuestro caso, cada fila tiene que tener su correspondiente condición de carrera) pero en el momento de predicción no se utiliza la clasificación, ya que el modelo predecirá una por cada fila. Dado que contamos con la clasificación real para cada fila del set de test, utilizaremos ésta para comparar la predicción contra la respuesta real. Esta comparación es la que nos dará el rendimiento de cada modelo, siendo mejor el modelo si la predicción está más cerca de la realidad. Esta noción de cercanía entre la predicción y la respuesta real es variable de problema en problema, hay varias métricas posibles con diferentes ventajas. Nuestra elección de métrica para medir el rendimiento de cada modelo, es la métrica F1. La métrica está definida en términos de *precision* y *recall*, otras dos medidas de relevancia y es el promedio ponderado de estas dos. Las métricas se definen como:

$$Precision = \frac{vp}{vp + fp}$$

$$Recall = \frac{vp}{vp + fn}$$

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Donde *vp*, *fp* y *fn* corresponden a Verdaderos Positivos, Falsos Positivos y Falsos Negativos respectivamente. El rango de la función es de 0 (peor rendimiento) a 1 (mejor rendimiento). La elección de esta métrica se basó en que otorga igual importancia a la *precision* tanto como al *recall*, teniendo así una predicción robusta.

Los métodos que fueron puestos a prueba para ser utilizados en esta etapa experimental, además de Random Forests [4] y Gradient Boosting [11] , incluyeron: SVM [6], Decision Trees [16], Perceptrones [14], Stocastic Gradient Descent [2](varias variantes, utilizando diferentes *loss functions*), Naive Bayes [13], Bagging [3] (con Decision Trees como estimador básico), Adaboost [9] y Extremely Randomized Trees [12]. Cada método utiliza diferentes parámetros, y se usaron varias combinaciones de valores elegidas arbitrariamente en principio y luego ajustadas según el rendimiento observado.

Al correr el script, la salida muestra el puntaje F1 de cada método, junto con una matriz de confusión para observar el comportamiento general de la clasificación.

Método	Puntaje F1
Gradient Boost	0.8596
Random Forest	0.8537
Bagging	0.8419
SVM	0.8220
AdaBoost	0.8140
Perceptron	0.8027
SGD	0.8024
Decision Trees	0.7969
Extra Tree	0.7946
Naive Bayes	0.6465

Cuadro 11: Algoritmos de machine learning aplicados al set de datos junto con su puntaje F1. Los mejores resultados fueron obtenidos por Gradient Boosting, Random Forests y Bagging.

En la tabla 11 se observa el rendimiento de todos los algoritmos aplicados por el script con su respectivo puntaje F1. Los métodos sobresalientes son Random Forests, Gradient Boosting y Bagging. Analizaremos los primeros dos resultados. Cabe destacar que no existe un algoritmo que sea superior en todos los casos, sino que el rendimiento de cada método depende del problema específico [19].

3.2.8. Resultados con Random Forests

Empezaremos analizando Random Forests, que en nuestro caso logró un puntaje F1 de 0.8537. Dentro de los parámetros configurables de la función, se encuentra la cantidad de árboles del bosque y la profundidad máxima de cada uno. Elegimos tener 70 árboles en el bosque con un máximo de 12 nodos de profundidad. Para observar específicamente cual fue la clasificación en este caso, se puede generar una matríz de confusión.

		Predicción		
		R	L	T
Verdad	R	472	37	9
	L	44	203	0
	T	18	0	39

Cuadro 12: Matriz de confusión para la clasificación de condición de carrera usando Random Forests. Los errores clasifican los estados L o T como R, o viceversa. No hay errores que clasifiquen L como T, teniendo buena separación entre estos estados. Los colores del cuadro se ajustan a la concentración de casos en el resultado, en degradé. En negro podemos ver que la mayoría de los casos se logra predecir correctamente la condición de los alumnos. En gris y gris claro, algunos casos aislados en los que el predictor falló al predecir la condición de los alumnos.

Los resultados de la matriz de confusión mostrada en 12 se basan en el conjunto de test. Este conjunto no fue utilizado para el entrenamiento del clasificador, por lo que refleja el comportamiento de éste en un caso similar a lo que se encontraría en la realidad. La matriz muestra que la clasificación es mayormente correcta, con errores sólo al diferenciar los estados L y T de los R. En ningún caso se clasificó erróneamente un estado L como T o viceversa, indicando

que estos estados son altamente diferenciables entre sí. Los casos en que se clasificó erroneamente pueden provenir de la dificultad que existe para diferenciar un alumno que está libre o terminó la carrera de uno regular, por el hecho de que la regularidad se puede mantener un tiempo aunque se haya abandonado la carrera (completando el censo) o que el alumno haya rendido los finales de los cuales tenemos información pero no haya completado la tesis aún.

Con Random Forests es posible obtener detalles interpretables acerca de la clasificación. Al entrenarse con varios árboles de decisión, se pueden observar las features significativas y qué decisiones se toman para llegar a la clasificación.

Las diez features más significativas, sobre las 61 totales, resultaron ser:

- *FEC_ACT_RE_ALGO2*
- *ANY_LIB*
- *FEC_ACT_RE_ALGEBRA*
- *FEC_NACIM*
- *NOT_EXAM_ALGO2*
- *DIF_NACIM_FEC_ALGO2*
- *DIF_ANY_LIB_FEC_ALGO2*
- *DIF_NACIM_FEC_ING2*
- *EDAD_INSCRIPCION*
- *FEC_ACT_RE_ANALISIS*

La importancia de las features es específica a la ejecución del método y depende tanto de los parámetros particulares como de la elección aleatoria de features que utiliza cada estimador de Random Forests. Aún así, en general las features más importantes son las mostradas, aunque el orden de la lista no se respete entre diferentes ejecuciones. Se observa que hay varias de éstas features que hacen referencia a Algoritmos II. Se tiene en cuenta como feature más importante la fecha en la cual se rindió esta materia, luego la nota obtenida por el alumno y la edad del alumno al rendirla. De la lista se observa, también, la importancia general de la edad del alumno en las features *ANY_LIB*, *FEC_NACIM* y *EDAD_INSCRIPCION*. Por último se da importancia a las fechas en las cuales se rindieron Algebra I y Análisis II y la edad del alumno al rendir Ingeniería II (la última materia obligatoria a rendir, según el plan de estudios).

Concluimos que según Random Forests, tener datos sobre Algoritmos II es crucial para la separación entre las clases a clasificar (R, L y T). Creemos que éstas features son buenos separadores entre las clases R y L, mientras que la feature que hace referencia a Ingeniería II separa entre los alumnos que terminaron la carrera (T) y los que no (R y L).

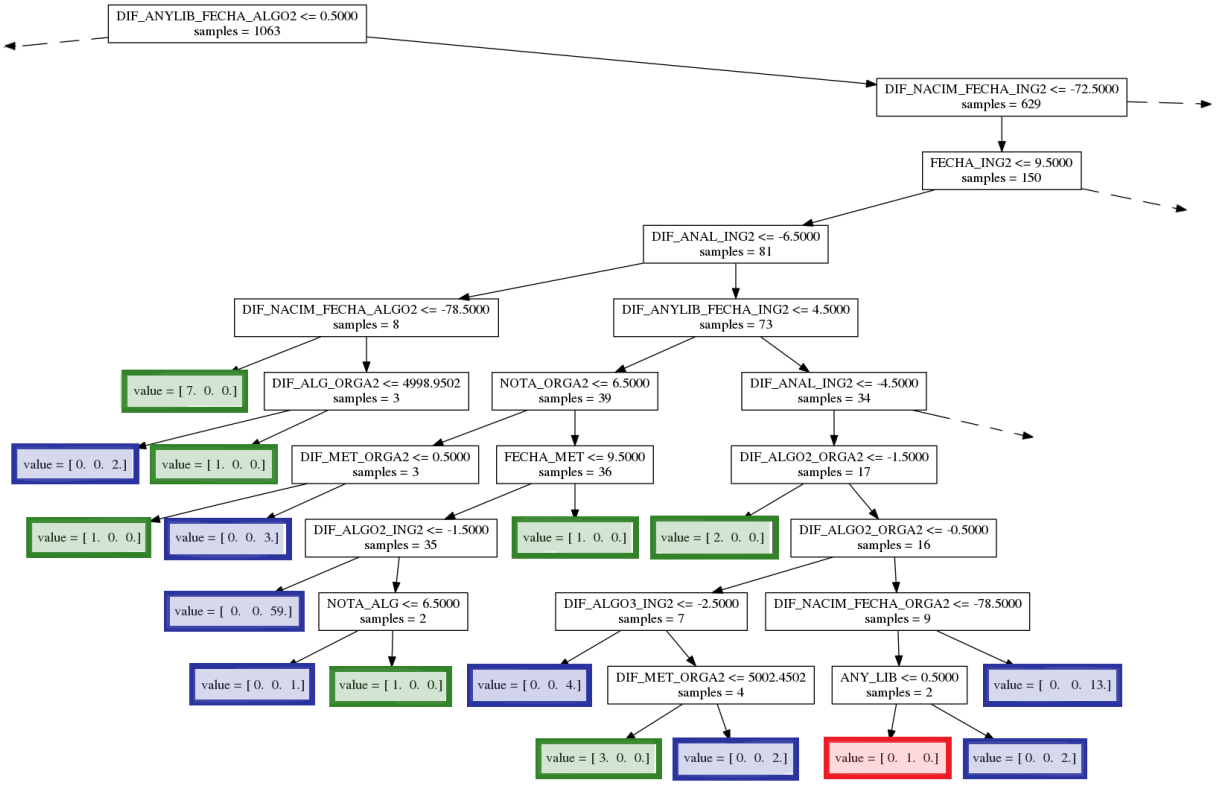


Figura 32: Representación de uno de los árboles estimadores del clasificador, recortado por brevedad. Muestra el proceso de decisión para algunas de las ramas, con los nodos internos mostrando condiciones sobre las features y las hojas mostrando la decisión para cada caso. Las clases R, T y L toman en las hojas los colores verde, azul y rojo respectivamente.

En la figura 32 se presenta uno de los árboles internos del clasificador, con el objetivo de mostrar la interpretabilidad del método de Random Forests. Fue recortado para poder visualizarlo ya que el original tiene hasta 12 nodos de profundidad. Se muestra la decisión tomada por el clasificador por las clases R, T y L con los colores verde, azul y rojo respectivamente. En este subárbol se observa que busca distiguir principalmente entre las clases R y T. Éste es sólo uno de los árboles de todo el ensamble; el método tiene en cuenta la información aportada por todos los árboles estimadores para llegar a una clasificación final.

3.2.9. Resultados con Gradient Boosting

En el caso de Gradient Boosting, los resultados obtenidos fueron similares, con un puntaje F1 de 0.8596. Los parámetros que toma la implementación del método en *scikit-learn* son la cantidad de etapas de Boosting a realizar y la profundidad alcanzada por cada estimador individual dentro del proceso de boosting. En nuestro caso, los resultados fueron obtenidos con 230 etapas de boosting y una profundidad de 14.

		Predicción		
		R	L	T
Verdad	R	475	33	10
	L	52	195	0
	T	19	0	38

Cuadro 13: Matriz de confusión para la clasificación de condición de carrera usando Gradient Boosting. Este caso es similar al de Random Forests, con errores de clasificación solo entre la clase R contra las clases L y T.

Una vez más en la matriz de confusión 13 observamos que los errores de clasificación se encuentran sólo al intentar diferenciar la clase R de las demás. En cuanto a las clases T y L, el método puede diferenciarlas sin problemas, al igual que Random Forests. Debido a esto, se puede inferir que las propiedades de un alumno que está por recibirse o que está por dejar la carrera son facilmente distinguibles y que el clasificador podría utilizarse para saber a cual clase se asemeja más el alumno.

En cuanto a las importancias, las diez features en éste caso fueron:

- *FEC_NACIM*
- *ANY_LIB*
- *EDAD_INSCRIPCION*
- *FEC_ACT_RE_ALGO2*
- *DIF_NACIM_FEC_ING2*
- *FEC_ACT_RE_ALGEBRA*
- *NOT_EXAM_ALGEBRA*
- *DIF_NACIM_FEC_ALGEBRA*
- *DIF_ANY_LIB_FEC_ALGO2*
- *CANT_MATERIAS_APROBADAS*

Las features son mayormente similares a las obtenidas en Random Forests, pero dándole más importacia a la edad del alumno (*FEC_NACIM*, *ANY_LIB*, *EDAD_INSCRIPCION*). Como en el caso anterior, se considera fuertemente la fecha en las que se rindieron Algoritmos II e Ingeniería II, la primera para distinguir la clase L de las demás y la última para distinguir T de las demás.

Se observan más features relacionadas con la materia Álgebra I, que consideramos que separa la clase L de las demás, ya que mucha gente deja la carrera antes de rendir el primer final como observamos en secciones anteriores. También parece importante la cantidad de finales aprobados por el alumno, pero no tiene tanto peso como la edad y la situación en las materias Álgebra I, Algoritmos II e Ingeniería II.

3.2.10. Experimentos adicionales

Teniendo en cuenta que nuestros clasificadores pueden realizar una buena separación entre las clases Libre y Terminó, analizaremos cuáles features específicas son determinantes para lograr esta separación. Utilizando un dataset sólo compuesto por alumnos de clase L y T, se entrenó el clasificador de Random Forests. La clasificación resultó en un puntaje F1 de 1.0, indicando que hay un conjunto de features determinante para distinguir una clase de la otra. Observando las cinco features más importantes obtenemos las siguientes:

- *DIF_ANY_LIB_FEC_ING2*
- *NOT_EXAM_ING2*
- *FEC_ACT_RE_ING2*
- *DIF_NACIM_FEC_ING2*
- *DIF_FEC_ALGO3_FEC_ING2*

Todas estas features refieren a la materia Ingeniería II. Esto indica, como es de esperarse, que todos los alumnos con clasificación T aprobaron la materia. Pero a su vez, indica que ningún alumno con condición libre llegó a rendir la materia.

Decidimos eliminar las features que refieran a Ingeniería II, para intentar obtener resultados más significativos. Es probable que un alumno regular promedio no haya llegado a cursar la materia todavía, por lo que resulta interesante ver que otros factores hacen que éste alumno esté más cerca de quedar libre o terminar la carrera. Al ejecutar el clasificador en esta instancia obtuvimos un puntaje F1 de 0.98 y las features más importantes fueron:

- *DIF_FEC_ALGO3_FEC_ORGA2*
- *CANT_MATERIAS_APROBADAS*
- *DIF_FEC_METODOS_FEC_ORGA2*
- *NOT_EXAM_METODOS*
- *DIF_NACIM_FEC_METODOS*

Se observa que, además de la cantidad de materias aprobadas, es importante que el alumno haya rendido Organización del Computador II y Métodos Numéricos (y en tercera instancia, Algoritmos III) para considerarse más cerca de recibirse que de quedar libre.

Por último, ejecutamos el clasificador entrenado con las clases L y T sobre un dataset que contiene sólo alumnos de clase R. El dataset utilizado fue el que no contiene las features referentes a Ingeniería II. El objetivo es identificar alumnos regulares que, por sus características, están cerca de recibirse o quedar libres. El clasificador puede indicar para cada caso, la probabilidad con la que cree que pertenece a cada clase. Decidimos quedarnos con los casos en los cuales el clasificador indica una probabilidad mayor a 0.9 en alguna de las clases objetivo. Fueron clasificados 924 alumnos como L y 218 como T. Dentro de los alumnos regulares que fueron clasificados como L, se encuentran alumnos

que abandonaron la carrera pero su condición de carrera no fue actualizada y alumnos que siguen cusando con cierta regularidad pero estan necesitando una ayuda adicional o motivación para continuar la carrera. En principio, con los datos con los que contamos, es imposible distinguir entre esos dos grupos. De los alumnos clasificados como T, se puede inferir que es un grupo cercano a recibirse y se los puede incentivar y motivar para poder finalizar la carrera en menos tiempo.

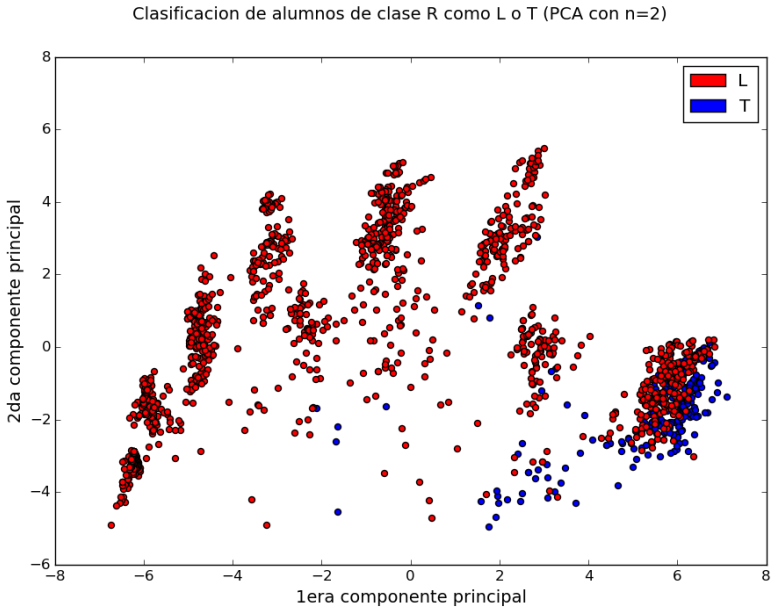


Figura 33: PCA con n=2 para los alumnos originalmente de clase R, pero clasiicados como L o T. Se observa que la diferencia entre las clases es marcada, aún en dos dimensiones.

Sólo a modo de visualización, en la Figura 33 se muestra una representación de las dos primeras componentes principales. Los datos mostados corresponden a los alumnos actualmente regulares clasificados como L o T, utilizando el dataset que no contiene features sobre Ingeniería II.Aunque las componentes mostradas no tienen una interpretación semántica de por sí, se puede observar una diferencia marcada entre las clases (aún en dos dimensiones).

4. Analisis sobre alumnas de género femenino

En esta sección abordaremos un estudio sobre las alumnas de la carrera. Realizamos una encuesta enfocada a este grupo específico que contiene preguntas sobre el rendimiento en la carrera y las posibles razones que pudieron haber llevado a la alumna a elegirla.

Antes de comenzar con nuestro análisis en particular, no queremos dejar de mencionar el aporte de la Fundación Sadosky con respecto a este tema.[7] La Fundación Sadosky realizó un estudio [17] para analizar, también, los diferentes factores que influyen en que las mujeres no elijan carreras relacionadas a la computación. El público objetivo de su estudio eran chicos/as del secundario, de género indistinto, los cuales debían completar una encuesta. Dado el alcance que pudo lograr la Fundación y la magnitud del público objetivo, se lograron recolectar cerca de 630 respuestas, en el conurbano.

Queremos destacar ciertos resultados interesantes que se desprenden del estudio de la fundación, ya que nos sirvieron como guía para definir algunos temas a la hora de realizar nuestro estudio. Estos resultados también resaltan la importancia y lo interesante de realizar un estudio de este tipo.

Potencialidad para la informática:

Un 22 % de la muestra (138 individuos), manifiestan alguna potencialidad para desarrollarse en la informática. Es decir, se trata de jóvenes que manifiestan diversos deseos, vocaciones, prácticas y/o actividades que permiten suponer que pueden devenir en informática. Entre estos estudiantes encontramos casi al cuádruple de varones que de mujeres (109 y 29 respectivamente). Específicamente entre quienes manifiestan la vocación de trabajar o estudiar informática la desproporción se agudiza: 9 mujeres y 67 varones. Y, más aún, entre quienes, adicionalmente, tienen alguna práctica cercana a la informática se cuentan 17 varones pero sólo 1 mujer.

Trabajo:

Sólo un 34 % de los entrevistados entiende que los salarios de las actividades informáticas son altos. Sin embargo, los salarios aparecen como un interés secundario para elegir una actividad laboral en el caso de los varones, y de mucha menor importancia aún para las mujeres.

Respecto de la actividad laboral deseada ambos géneros coinciden en priorizar a las actividades profesionales. Luego, en el caso de las mujeres, las actividades del arte y del espectáculo tienen un lugar preferencial. La asistencia a personas (el llamado trabajo afectivo), las actividades de belleza y estética, el trabajo informacional y la docencia también se destacan. En el caso de los varones, resulta remarcable la inclinación a mencionar a la producción de software, que ocupa el segundo lugar en las preferencias, aunque con un porcentaje de apenas un 12 % de las preferencias.

[7] Google realizó un trabajo muy interesante sobre este tema, pero fue publicado durante la etapa final de esta tesis, por este motivo no fue tenido en cuenta en este trabajo. El mismo se puede consultar en http://static.googleusercontent.com/media/www.wenca.cn/en/us/edu/pdf/women-who-choose-what-really.pdf

Software:

Un 48 % de los varones y un 63 % de las mujeres declara no saber qué es un programa de computadora o software.

Educación superior:

En términos de carreras de educación superior deseadas, entre las mujeres, las carreras de informática ocupan el anteúltimo lugar, y son elegidas sólo por un 2,3 % de las entrevistadas. Por el contrario, en el caso de los varones, computación e informática aparecen como las carreras con más potencial favoritismo, con un 19,4 % de las respuestas.

Habilidades:

Algunas habilidades asociadas a las que se utilizan en los procesos productivos de software tienden a estar más (o faltar menos) entre los varones que entre las mujeres: armar y desarmar objetos, aprender autónomamente, hacer tareas de matemática y lógica, estar sentados frente a una computadora por un tiempo prolongado.

Contacto con programadores:

Los entrevistados presentan un escaso contacto con programadores: alrededor de un 60 % no conoce a ninguno. Las mujeres tienden a conocer menos que los varones.

Uso diario (promedio) de la computadora:

La media del uso diario de la computadora es de 4 horas, lo que supone una vinculación importante con esa tecnología digital. Sin embargo, la distinción entre los sexos muestra que los varones pasan en promedio 4 horas y 20 minutos, mientras las mujeres dedican 3 horas y 40 minutos.

Sentido de pertenencia con la carrera:

Los resultados de las encuestas indican que ambos sexos consideran a las mujeres distantes de las carreras informáticas. Las mujeres se perciben más aún distantes de como las perciben los varones. En otras palabras, la informática es una actividad laboral que no sólo ambos sexos tienden a disociar de las mujeres; sino que es una en la que las propias mujeres se juzgan aún más ajenas de lo que las consideran los hombres. No obstante, se puede relativizar y contextualizar las afirmaciones previas, ya que en el caso de la programación hay, de cualquier modo, un 55 % de los entrevistados que considera que es una actividad para cualquiera de los dos sexos.

De los diferentes resultados expuestos pudimos observar características que no influyen a las mujeres a elegir una carrera con fines informáticos, pero sobre todo pudimos apreciar (ya desde el secundario) que las mujeres tienden a no elegir estas carreras. En el informe de la Fundación Sadosky se pueden observar más en detalle estos y otros resultados, con tablas y gráficos.

Volviendo a nuestro estudio, como mencionamos y pudimos apreciar en la

sección 2.3 , no contamos con una gran presencia de alumnos de género femenino en la carrera de Licenciatura en Ciencias de la Computación de nuestra facultad. Este fenómeno no sólo se produce particularmente para nuestra carrera, sino que es de público conocimiento que sucede en gran parte de gran parte de las carreras relacionadas con las ciencias de la computación en todo el mundo.

Nuestro objetivo en esta sección es poder aportar desde nuestro lugar, algunos resultados que ayuden a entender porque se produce este fenómeno. También buscaremos poder detectar, al menos de forma particular, algunos factores que influyan en que las mujeres elijan nuestra carrera y tengan éxito en la misma, a fines de poder fomentar e incrementar la participación femenina en la misma.

4.1. Obtención de datos

No existen datos públicos en nuestra facultad a partir de los cuales se pueda realizar el tipo de análisis que pretendemos en esta sección. Las fuentes a las que podemos remitirnos, como ser el sistema de inscripciones (filtrando a los alumnos por género femenino) no cuentan con ningún dato que pueda responder el propósito de nuestro análisis (detectar factores que influyan en que las mujeres elijan la carrera o tengan éxito en la misma), sino más bien datos exclusivamente académicos.

Si bien se han analizado algunos resultados que se produjeron trabajando con los datos de secciones previas, y por lo dicho anteriormente, para esta sección trabajaremos exclusivamente con datos obtenidos de una encuesta que realizamos al alumnado de género femenino de la carrera de Licenciatura en Ciencias de la Computación de la FCEN.

Al igual que en la seccion 2.3, una vez conformada la encuesta, incentivamos al público objetivo a llenarla. Esto se llevó a cabo difundiendo la encuesta vía e-mail mediante la lista de distribución de la carrera y diferentes grupos de alumnos creados en las redes sociales a los que los alumnos de la carrera frecuentan. Se obtuvieron un total de 59 respuestas.

4.2. Elaboración de la encuesta

Las preguntas a incluidas en esta encuesta se separan en dos secciones: Una enfocada en detectar factores que puedan influir en el éxito académico de las alumnas, y la segunda en conocer factores que puedan haber influido para que las alumnas eligiesen estudiar Licenciatura en Ciencias de la Computación en la UBA.

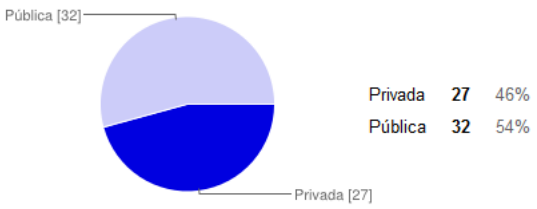
Debido a la poca cantidad de alumnas que contamos en nuestra carrera, la cantidad de datos no es significativa como para realizar un análisis del que se puedan sacar conclusiones fuertes. Sin embargo, el objetivo de esta sección es poder analizar particularmente a las estudiantes de género femenino de nuestra facultad y mostrar resultados que podrían, o no, darse a nivel global.

Uno de los factores positivos de esta encuesta fue que al ser confeccionada posteriormente a la encuesta de la sección 2 pudimos formular nuevas preguntas para hacer foco sobre ciertos aspectos que quizá omitimos incluir en la anterior.

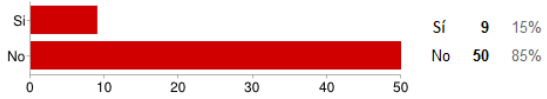
La encuesta completa se puede consultar en el Anexo 2.

4.3. Análisis exploratorio

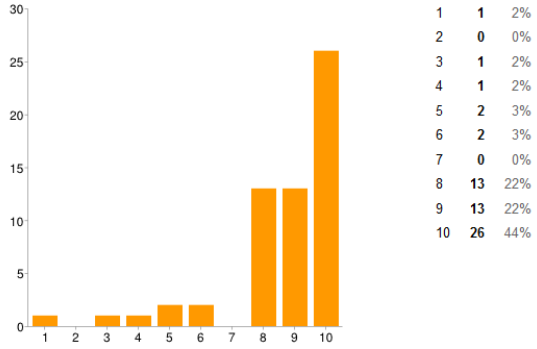
A continuación se exhiben los resultados más relevantes obtenidos de las alumnas que respondieron la encuesta (la lista completa de resultados se puede ver desde aquí)



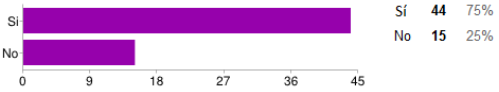
(a) ¿Tu escuela era privada o pública?



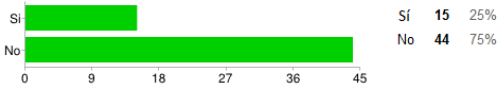
(b) ¿Estudiaste en una escuela técnica?



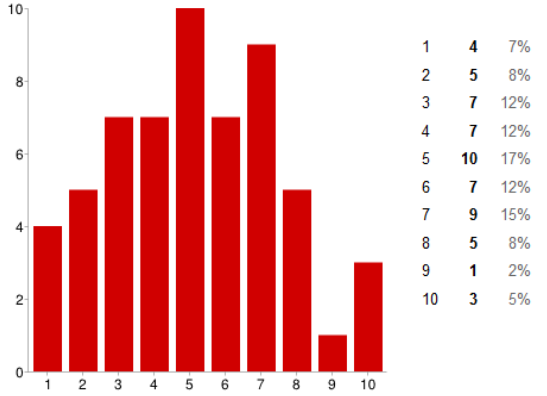
(c) En tu opinión ¿Cómo te iba en matemáticas en la secundaria?



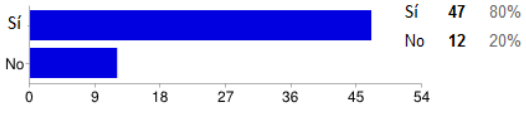
(d) ¿Mostrabas interés en computación antes de entrar a la carrera?



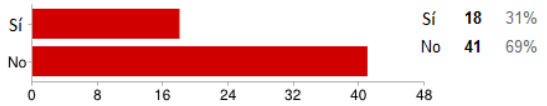
(e) ¿Tenías temas de programación en la secundaria?



(f) ¿Cuánto te costó superar el CBC?



(g) ¿Asististe a alguna charla del DC antes de elegir la carrera?



(h) ¿Sabías inglés antes de entrar a la carrera?

Figura 34: Algunos resultados de la encuesta realizada a alumnas de género femenino de la carrera Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales de la UBA.

Para comenzar, en la figura 34a podemos observar que casi la mitad de las alumnas encuestadas asistieron a una escuela pública y la otra mitad a una privada, con un mínimo predominio de las que asistieron a escuelas públicas.

No sucede lo mismo cuando se preguntó si la escuela a la que asistieron era técnica, apenas el 15 % de las encuestadas cumple ese característica (Figura 34b).

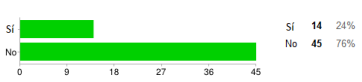
En la Figura 34c observamos un rasgo que puede ser interesante, el 88 % de las encuestadas considera que le iba bien en matemáticas en el secundario, con un 44 % afirmando que le iba excelente. Es decir, que los conocimientos matemáticos pueden ser un factor interesante para que las alumnas elijan carreras a fines a sistemas.

Por otro lado, como se ve en la Figura 34e, apenas el 15 % de las encuestadas pudo tener contacto con algo relacionado a la programación en el secundario. Sin embargo, el 75 % afirma haber tenido interés en computación antes de comenzar la carrera (Figura 34d).

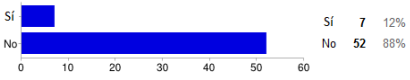
En cuanto al esfuerzo realizado para superar el Ciclo Básico Común que exige la Universidad de Buenos Aires a los alumnos de la carrera como requisito para ingresar a la carrera, las opiniones estan divididas bastante equitativamente en los diferentes posibles valores de difícultad que las encuestadas podían seleccionar (Figura 34f). Si bien se observan dos picos en los valores cinco y siete, descartamos que este sea un factor que influya determinantemente en el éxito académico de las estudiantes.

De la Figura 34g vemos que apenas el 31 % de las encuestadas asisitió a alguna charla impulsada por el departamento de computación antes de decidirse por estudiar Licenciatura en Ciencias de la Computación, sin embargo, en la lista completa de resultados se puede observar que aquellas que sí asistieron señalaron (en su mayoría) que la charla había influido considerablemente en su decisión para elegir la carrera.

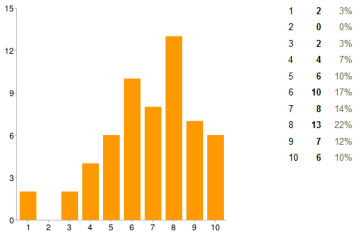
Consideramos que el idioma inglés es bastante importante para una carrera con la salida laboral que tiene la carrera de Ciencias de la Computación, dada la gran demanda nacional e internacional. Aprovechamos para incluir la pregunta sobre cuantas de nuestras alumnas saben dicho idioma. En la Figura 34h se ve como el 80 % de las encuestadas ya dominaba el idioma antes de ingresar a la carrera.



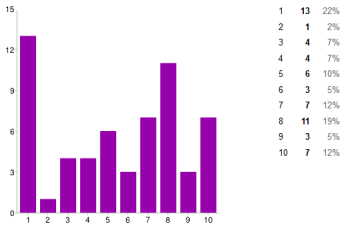
(a) ¿Tuviste algún referente femenino en el DC?



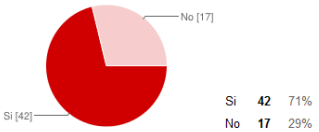
(b) ¿Trabajabas en algo relacionado con sistemas antes de elegir estudiar en el DC?



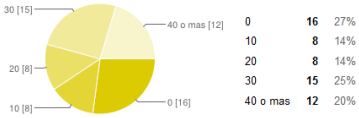
(c) ¿Cómo calificarías tu nivel de inglés actual?



(d) ¿Cuánto influyó la amplia salida laboral en tu decisión de estudiar en el DC?



(e) ¿Trabajás actualmente?



(f) ¿Cuantas horas por semana trabajas?

Figura 35: Algunos resultados de la encuesta realizada a alumnos de género femenino de la carrera Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales de la UBA.

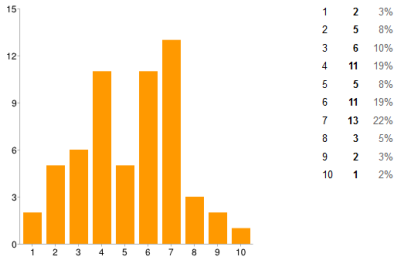
También decidimos preguntar por el nivel de inglés actual que cada alumna tiene según su propia percepción, para conocer la realidad actual en este aspecto. En la Figura 35c podemos observar que en general las alumnas se califican con 6 o más, lo cual refleja que en general el nivel de inglés es bueno (de nuevo, desde el punto de vista de las alumnas).

Cuando preguntamos si las alumnas tenían algún referente femenino (alguien en quien se vean reflejadas o que las inspire a estudiar la carrera) el 76 % respondió que no, por lo que este no parece ser un factor que influya en el éxito académico ni que lleve a las alumnas a elegir la carrera (Figura 35a). En este último sentido, los resultados pueden tener que ver con lo mencionado en el informe de la Fundación Sadosky donde el 60 % de los chicos de colegio no conoce a ningún programador/a.

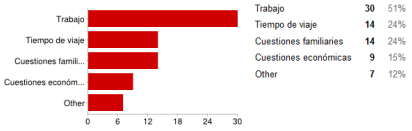
En la Figura 35b vemos como el 88 % de las encuestadas no trabajaba en algo relacionado a computación antes de decidir estudiar computación. Por otro lado, en la Figura 35e también vemos que el 71 % de las encuestadas trabaja actualmente, y en su mayoría lo hacen con una carga horaria de 30 horas por semana o más (Figura 35f). Más adelante en esta sección veremos como el trabajar y la carga horaria mostrada en estas figuras influyen en el éxito académico en cuanto a la falta de tiempo para estudiar.

Otra pregunta que se hizo a las encuestadas fue que indicaran (con un puntaje de 1 a 10) la influencia de la salida laboral que tiene la carrera a la hora de decidir estudiar computación. En la Figura 35d vemos que, en general, las opiniones estan bastante divididas. Un 22 % de las encuestadas afirma que la

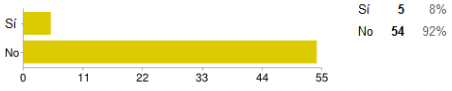
salida laboral no influyó en nada en su decisión, mientras que un 10 % votó 5, el 12 % votó 7, el 19 % calificó con 8 y el 12 %, 10. Es decir, para algunas alumnas la salida laboral influyó considerablemente, mientras que para otro porcentaje respetable no tuvo incidencia alguna. Por lo que la salida laboral es un aspecto interesante de la carrera, pero que no siempre es el principal factor por lo que las alumnas la eligen.



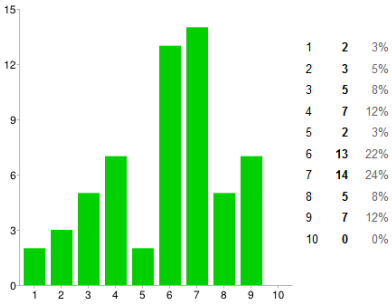
(a) ¿Cuánto sentís que te alcanza el tiempo que le dedicás a la facu?



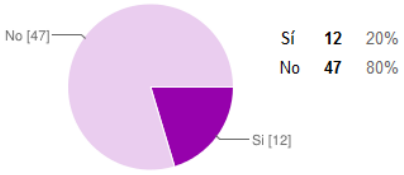
(b) ¿Qué factores hacen que no te alcance el tiempo?



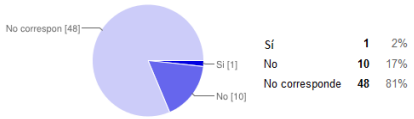
(c) ¿Pusiste en duda estudiar sistemas por la mayoría masculina en las clases?



(d) ¿Estás contenta con tu rendimiento académico?



(e) ¿Estudiaste otra carrera antes de empezar en el DC?



(f) ¿La otra carrera tenía que ver con Sistemas?

Figura 36: Algunos resultados de la encuesta realizada a alumnas de género femenino de la carrera Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales de la UBA.

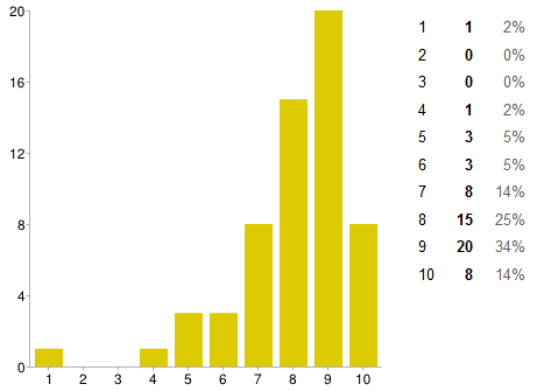
Continuando con los resultados, en la figura 36a podemos ver que las alumnas se dividen casi en un 50 y 50 entre las que consideran que les alcanza el tiempo que le dedican a la facultad en función de sus expectativas académicas y las que no. La pregunta permitía calificar cuanto sentía la alumna que le alcanzaba el tiempo con un puntaje de 1 a 10. El 48 % calificó con 5 o menos, mientras que el 52 % votó 6 o más. Podemos ver que si las alumnas consideran que el tiempo no les alcanza acorde a sus expectativas, entonces hay ciertos

factores que influyen en su éxito académico que hacen que el tiempo no alcance para (casi) la mitad de las encuestadas. Intentamos encontrar cuales son estos factores con otra pregunta de la encuesta, proponiendo algunas opciones (Figura 36b). Esta pregunta la debían contestar únicamente aquellas alumnas que hayan calificado con 6 o menos la pregunta anterior. Notar que para esta pregunta, las encuestadas podían votar más de una opción, es por eso que nos centraremos en la cantidad de votos de cada una de las opciones y no en el porcentaje. De este subgrupo de alumnas, 30 indicaron al trabajo como factor de que no les alcance el tiempo, 14 indicaron que un factor influyente era el tiempo de viaje hasta la facultad, 14 acusaron cuestiones familiares, 9 cuestiones económicas, y 7 indicaron que también había otros factores que no se encontraban entre las opciones que influyen en la falta de tiempo. Como habíamos mencionado anteriormente en esta sección, dado que había una gran cantidad de alumnas encuestadas que trabajan, era esperarse que el trabajo sea uno de los factores principales, y sin dudas es un factor importante en el éxito académico (al menos, desde el punto de vista de las alumnas).

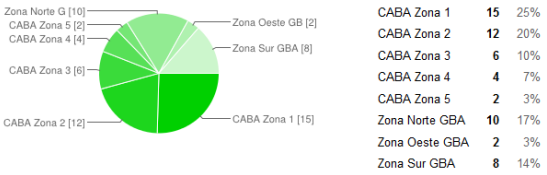
Continuando con el análisis, en la Figura 36c podemos ver como apenas un 8 % de las encuestadas afirma haber puesto en duda estudiar la carrera sabiendo que iba a haber una mayor presencia masculina en las clases. Podemos descartar, entonces, este hecho como factor de influencia en estudiar la carrera.

En cuanto a la percepción propia de las alumnas en cuanto al rendimiento académico de cada una, encontramos que el 22 % y 24 % calificaron con 6 y 7 respectivamente su satisfaccion por el rendimiento que tienen en la carrera (Figura 36d). Sin embargo, llama la atención que el 31 % haya calificado su satisfacción con 5 o menos.

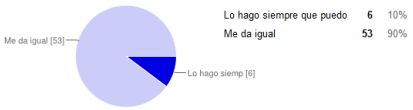
Otra pregunta consultaba si las alumnas habían estudiado otra carrera previamente a escoger la Licenciatura en Ciencias de la Computación. En la Figura 36e vemos que el 20 % lo hizo, lo cual es un porcentaje respetable con respecto a lo que esperábamos. Luego consultamos (a ese 20 %) si la carrera escogida anteriormente tenía que ver con computación (Figura 36f). El resultado fue que casi la totalidad indicó que la carrera anterior no tenía relación alguna (apenas una votó que si).



(a) ¿Te gusta lo que ves en la carrera?



(b) ¿En que zona vivís?



(c) ¿Buscás integrarte con mujeres para los TPs siempre que podés?

(d) ¿Sentís que rendís mejor estudiando sola o con un grupo de estudio?

Figura 37: Algunos resultados de la encuesta realizada a alumnas de género femenino de la carrera Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales de la UBA.

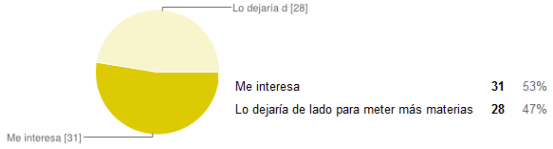
Finalmente, en la figura 37a podemos observar que el 92 % de las encuestadas votó 6 o más en una escala de 1 a 10 para la pregunta '¿Te gusta lo que ves en la carrera?', lo cual confirma rotundamente la satisfacción de casi la totalidad del alumnado femenino para con el plan de estudios y contenidos de las materias de la carrera.

El mapa (para conocer las distintas numeraciones asignadas a las zonas de CABA) correspondiente al gráfico de la pregunta '¿En que zona vivis?' se puede consultar en el Anexo 2 donde se encuentran todas las preguntas que fueron incluídas en la encuesta. En la Figura 37b podemos apreciar que el 25 % de las encuestadas pertenece a Caba Zona 1, seguida por un 20 % perteneciente a Caba Zona 2, y un importante aporte de Zona Norte y Sur de Buenos Aires (con un 17 % y 14 % respectivamente). Notamos que las zonas que más alumnas aportan son las mismas que predominaban cuando hicimos el análisis utilizando la encuesta dirigida a todos los alumnos, de ambos géneros. Zona Norte también se destacaba con un importante aporte luego de las dos principales. Sin embargo, el aporte de alumnas de la Zona Sur se destaca más en este caso, que en los resultados de la encuesta mixta. Estos datos son importantes para la divulgación de la carrera, entendiendo de que zonas provienen los alumnos de la

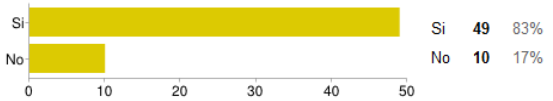
facultad se puede elegir mejor donde promover y donde no las carreras del área de computación.

Analizamos que tan común es que una alumna forme grupo de trabajos prácticos con otas alumnas. En la Figura 37c se puede observar que apenas el 10 % lo hace siempre que puede. El 90 % de las alumnas no tiene preferencias en cuanto a compartir grupo con hombres.

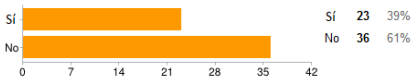
A partir de la Figura 37d vemos que las mujeres prefieren estudiar de a grupos, que solas. El 73 % de las encuestadas opinó que percibe que tiene un mejor rendimiento estudiando en grupo, mientras que el 27 % afirma que su rendimiento es más alto si estudia por su cuenta.



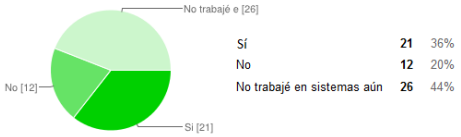
(a) ¿Te interesa el promedio o lo dejarías de lado si podes meter más materias en menos tiempo?



(b) ¿Te interesan las cuestions matemáticas de los temas que ves en la carrera?



(c) ¿Algún familiar tuyo estudió/trabajó de algo relacionado con sistemas?



(d) ¿Los conocimientos del trabajo te sirvieron para que alguna cursada se te haga más simple?

Figura 38: Algunos resultados de la encuesta realizada a alumnas de género femenino de la carrera Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales de la UBA.

Las alumnas fueron consultadas sobre si preferían aprobar materias para obtener rápidamente el titulo de grado sin importar las calificaciones u obtener buenas calificaciones aunque implique demorarse un poco en obtener el título (Figura 38a). Las opiniones estuvieron bastante divididas. El 53 % afirmó que prefiere obtener buenas calificaciones y darle importancia al promedio, mientras que el 47 % admitió que dejaría de lado las buenas calificaciones a cambio de obtener el título más rápidamente.

Consultamos a las alumnas si las cuestiones matemáticas de la carrera eran de su interés. Los resultados siguen siendo contundentes en este campo. En la Figura 38b podemos ver como el 83 % opinó que las cuestiones matemáticas son de su interés, mientras que apenas un 17 % votaron que no.

Otro resultado interesante se encuentró cuando se consultó si algún familiar de las encuestadas trabajó u estudió algo relacionado a la computación (Figura

38c). El 61 % contestó que no, y un 39 % que sí. A priori se podría decir que tener un familiar relacionado con el área de computación es un factor que influye para que las alumnas hayan elegido la carrera.

Por último y cerrando este análisis exploratorio, las alumnas fueron consultadas sobre si los conocimientos que se adquieren al trabajar les sirven para alguna materia que se dicta en la carrera. En la Figura 38d podemos ver que el 36 % afirmó que los conocimientos laborales le sirvieron, el 20 % dijo que no le sirvieron en la carrera, mientras que el 44 % no tuvo aún un trabajo relacionado con la computación como para dar su veredicto.

Aún con estos resultados, resulta útil separar el alumnado en ciertos grupos de interés para observar sus diferencias y similitudes, al igual que hicimos en la sección 2.4. En la siguiente sección analizaremos características de diferentes grupos en los que se dividió al alumnado.

4.4. Análisis exploratorio por grupos

Como dijimos anteriormente, se realizaron varios tipos de divisiones dentro del alumnado:

- Las que trabajan y las que no trabajan
- Las de escuela técnica y no técnica
- Las de escuela pública y privada
- Las que les interesa el promedio y las que lo dejarían de lado para obtener el título más rápidamente

Elegimos realizar una distinción entre las alumnas que trabajan y los que no, dado que el hecho de trabajar demanda tiempo, y vimos en la sección anterior como el trabajar aparecía como uno de los factores más señalados por las encuestadas por la que no cuentan con todo el tiempo que quisieran para estudiar.

Por el lado de las de escuela técnica y no técnica, buscamos ver si el hecho de haber concurrido a una escuela técnica incide en el éxito académico de las estudiantes. Sabiendo que las escuelas técnicas suelen incluir ciertos conocimientos que pueden aplicarse a la informática, analizaremos si existe una diferencia interesante en algún aspecto académico entre estos dos grupos.

Siguiendo la misma línea, otra distinción de grupos que nos resultó interesante realizar fue según si las alumnas estudiaron en una escuela pública o si lo hicieron en una privada. Si bien los conocimientos impartidos en una u otra no tienen porque diferir, analizaremos si haber estudiado en uno u otro tipo de escuela influye en el éxito académico de las encuestadas.

Por último, en la encuesta se consultó sobre si las alumnas dejarían de lado el promedio para obtener el título más rápidamente o si el promedio les interesaba más allá de que pueda demorarlas en la carrera, separaremos a las alumnas en dos grupos según hayan optado por una u otra respuesta.

Al realizar las comparaciones entre las propiedades de los diferentes grupos, se aplicarán los tests estadísticos de Student y Kolmogorov–Smirnov (mencionados en la sección 1.2) que nos indicarán si los resultados obtenidos son estadísticamente significativos respecto a alguna medición en particular.

En los siguientes análisis haremos uso de la palabra 'significativo' para describir diferencias 'estadísticamente significativas' arrojadas por los tests utilizados, pudiendo existir diferencias significativas (en su interpretación coloquial) visuales en los gráficos.

Recordemos que dado que en nuestra muestra contamos una pequeña cantidad de datos, los análisis que se realizaran a continuación son anecdóticos sobre lo que sucede puntualmente para las estudiantes de nuestra carrera.

4.4.1. Trabajan vs no trabajan

En esta sección analizaremos las propiedades de las alumnas que trabajan comparándolas con las que no lo hacen. Veremos que a las alumnas que trabajan les suele ir mejor en algunas materias de la carrera (por las que se preguntó en la encuesta) que a aquellas que no lo hacen. Sin embargo las alumnas que trabajan, como vimos en la sección 4, afirman que les alcanza menos el tiempo que a aquellas que no lo hacen. En este análisis podremos ver este aspecto más en detalle.

Un punto importante para aclarar en esta sección es que no se compararon las calificaciones de las alumnas para la materia Algoritmos y Estructura de datos III (las cuales fueron consultadas en la encuesta) debido a que en la muestra no hay alumnas que cumplan no estar trabajando y haber rendido este final. Como mencionamos a lo largo de esta tesis, la amplia salida laboral de la carrera hace que los alumnos de la facultad se inserten rápidamente en el mercado, y dado que Algoritmos y Estructuras de Datos III es una materia del tercer año de la carrera, no resulta del todo atípico que esto suceda a esa altura.

En principio decidimos comparar las calificaciones que indicaron las alumnas de uno y otro grupo para la materia Álgebra I, dictada en el primer año de la carrera. En el gráfico de densidad de la Figura 39 se observa cómo las estudiantes que trabajan (actualmente) parecen obtener mejores calificaciones para esta materia. Las que trabajan se centran en siete (con un pequeño pico en diez), mientras que las que no trabajan se centran entre cuatro y cinco.

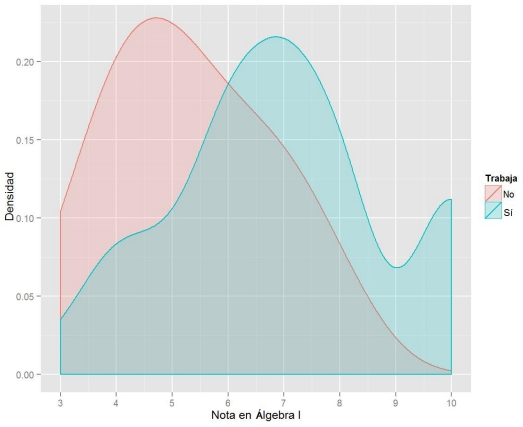


Figura 39: Gráfico de densidad de las calificaciones de la materia Álgebra I para los alumnos de género femenino, distinguiendo aquellas a que trabajan y los que no. Se observa como las alumnas que no trabajan parecieran obtener mejores calificaciones que aquellas que lo hacen, para esta materia en particular (modas = 7 y 4.5 respectivamente).

Realizamos los tests estadísticos para comparar estos dos grupos y observar si efectivamente existían diferencias estadísticamente significativas entre ambos. El test de Student sugirió que sí existen diferencias en las medias de las notas entre los dos grupos ($p < 0.0057$). Sin embargo, para el test de Kolmogorov–Smirnov para este mismo grupo se obtuvo un resultado no significativo ($p < 0.1447$).

Continuando con el análisis para este grupo, decidimos observar qué sucedía para una materia de segundo año como Algoritmos y Estructura de Datos II. El gráfico de distribución correspondiente a este análisis se puede observar en la Figura 40. La distribución para ambos grupos parece ser bimodal. Las estudiantes que trabajan tienen dos picos entre cinco y seis y otro en nueve, las que no trabajan los tienen en cuatro y entre siete y ocho.

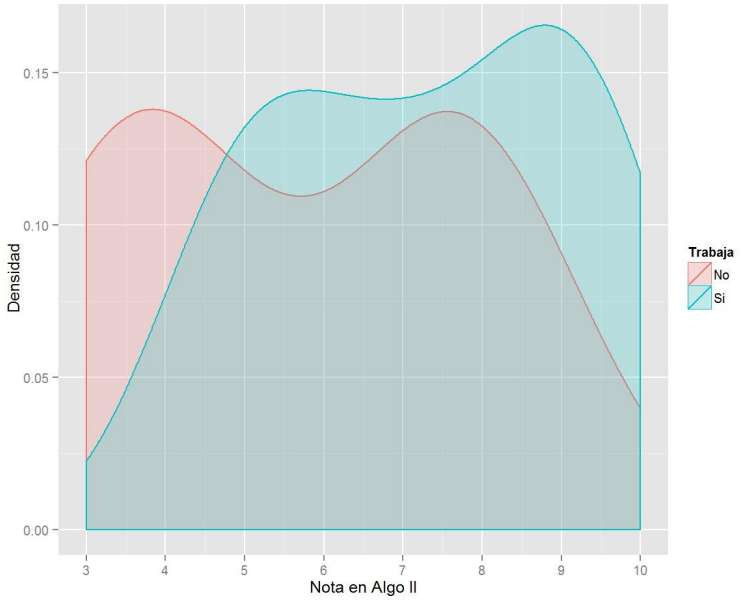


Figura 40: Gráfico de densidad de las calificaciones de la materia Algoritmos y Estructura de Datos II para los alumnos de género femenino, distinguiendo aquellos a que trabajan y los que no. Se observa que si bien las modas se hallan en puntos distintos para cada grupo, los tests estadísticos afirman que no existen diferencias estadísticas significativas entre las calificaciones de ambos.

Los tests de Student y Kolmogorov–Smirnov indican que para estos grupos no existen diferencias estadísticamente significativas en aspecto ($p < 0.149$ y $p < 0.314$ respectivamente), aunque en principio, mirando el gráfico de la figura parece haberlas. Casi la totalidad de las mujeres de ambos grupos logró aprobar la materia.

Por último, decidimos ver más en detalle la pregunta sobre '¿Cuánto sentís que te alcanza el tiempo que le dedicás a la facu?' para estos dos grupos, en el que debían calificar con un puntaje entre uno y diez según cuánto creían que les alcanzaba el tiempo en base a sus expectativas académicas. Recordamos que en la sección 4.3 vimos como el hecho de trabajar era uno de los factores de falta de tiempo más señalados por las encuestadas.

En el gráfico de densidad de la figura 41 se puede observar claramente como las alumnas que no trabajan se centran en el valor siete (y algunas pocas en cinco), mientras que las que trabajan tienen dos picos definidos, uno entre tres y cuatro y otro en seis.

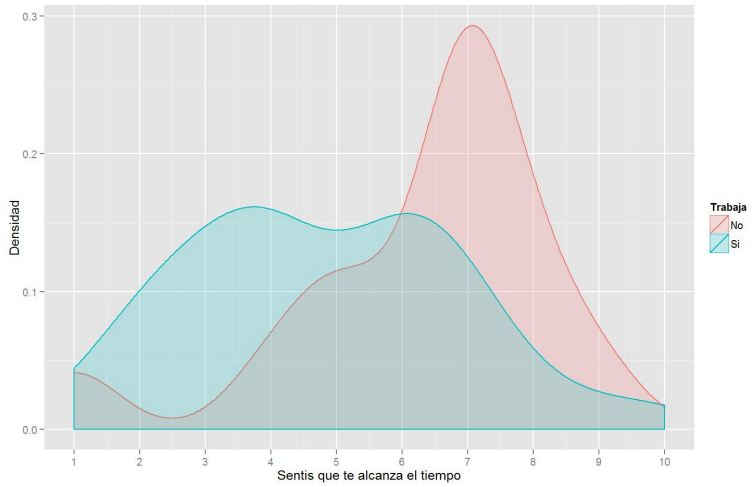


Figura 41: Gráfico de densidad de la percepción de cuanto les alcanza el tiempo para tener un rendimiento óptimo a los alumnas de género femenino, distinguiendo aquellas a que trabajan y los que no. Las alumnas que no trabajan opinan que les alcanza más el tiempo con respecto a aquellas que lo hacen.

Para observar mejor las calificaciones de uno y otro grupo, decidimos generar un gráfico de barras porcentual en el que distinguiremos los votos de uno y otro grupo. En la figura 42 se puede observar con mayor claridad como poco más del 40 % de las alumnas que no trabajan votó siete como respuesta, mientras que las que no trabajan se encuentran bastante divididas entre cuatro y seis.

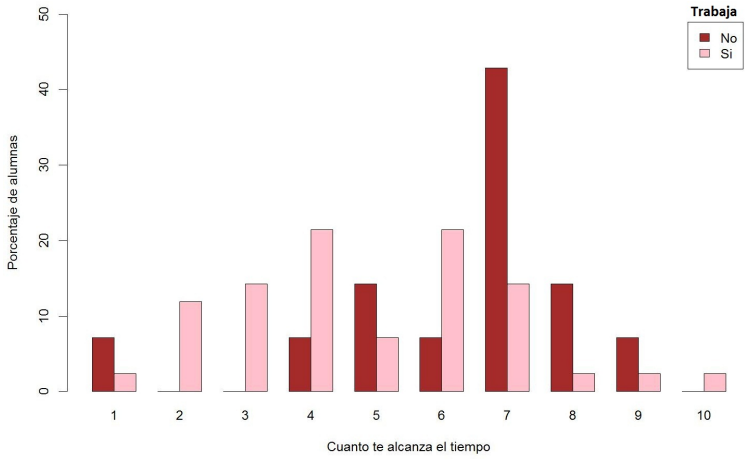


Figura 42: Gráfico de barras ilustrando de forma más detallada los valores de cuánto les alcanza el tiempo indicados por las alumnas de cada uno de los grupos (las que trabajan y las que no).

Los tests de Student y Kolmogorov–Smirnov indican que efectivamente existen diferencias estadísticas a tener en cuenta en este aspecto ($p < 0.0297$ y $p < 0.0422$ respectivamente). Por lo tanto, confirmamos una vez más la influencia de trabajar en la percepción del tiempo con el que cuentan las alumnas para estudiar y tener un buen rendimiento académico.

4.4.2. Escuela técnica vs no técnica

En esta sección analizaremos las propiedades de las alumnas que asistieron a una escuela secundaria técnica comparándolas con las que asistieron a una secundaria con otra orientación. Decidimos realizar esta distinción considerando que las escuelas técnicas suelen impartir conocimientos que muchas veces son utilizados o sirven como base para comprender mejor o tener cierta facilidad al aprender los temas que se aprenden en las diferentes materias de una carrera como la nuestra, de ciencias exactas. Se busca, entonces, observar si este factor influye en el éxito académico de las encuestadas.

A lo largo de esta sección, veremos que los tests estadísticos que utilizamos en esta tesis indican que no existen diferencias a tener en cuenta para las calificaciones de distintas materias, ni tampoco en cuanto a dificultades para superar el Ciclo Básico Común exigido por la UBA separando a las alumnas en estos dos grupos.

También observaremos que las estudiantes de escuela técnica suelen obtener conocimientos sobre programación en el secundario, algo que ocurre rara vez con las que asistieron a una secundaria con otra orientación. Con esto confirmamos, una vez más, lo observado en análisis anteriores (cuando realizamos la encuesta para ambos géneros en la sección 2.3) en los que vimos como el hecho de saber programar con anterioridad no influye determinantemente en el éxito académico de los alumnos de nuestra facultad (en este caso, de género femenino particularmente).

Al igual que para el grupo anterior, comenzamos analizando las calificaciones en la materia Álgebra I. En el gráfico de densidad de la Figura 43 se observa que las alumnas que provienen de escuelas con otras orientaciones parecerían obtener mejores calificaciones, teniendo su pico en 7. Las alumnas que asistieron a escuelas técnicas, por otro lado, tienen su pico en 5.

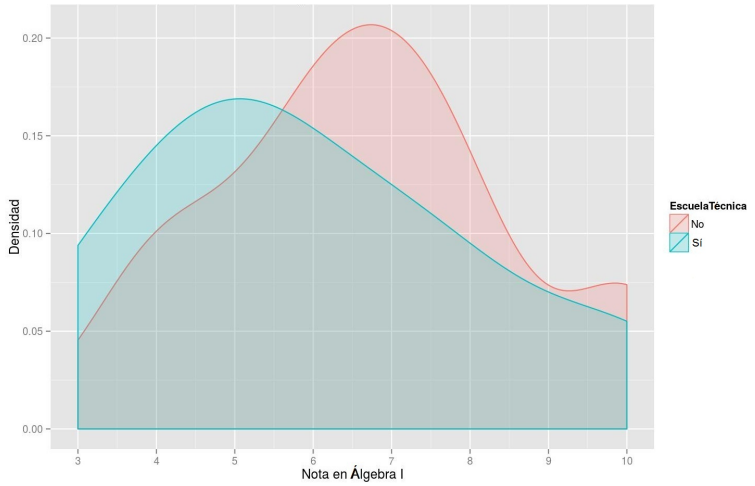


Figura 43: Gráfico de densidad de las calificaciones de la materia Álgebra I para las alumnas de género femenino, distinguiendo aquellas a las que provienen de escuela técnica y los que no. Se observa como las alumnas que no provienen de escuelas técnicas parecerían obtener mejores calificaciones que aquellas que lo hacen, para esta materia en particular (modas = 7 y 5 respectivamente).

A pesar de lo que se ve en el gráfico, los tests de Student y Kolmogorov–Smirnov nos indican que no existen diferencias estadísticamente significativas tanto para las medias como para la distribución de las calificaciones obtenidas por ambos grupos ($p < 0.449$ y $p < 0.816$ respectivamente).

Veamos que sucede para Algoritmos y Estructura de Datos II. El gráfico correspondiente se ve en la Figura 44.

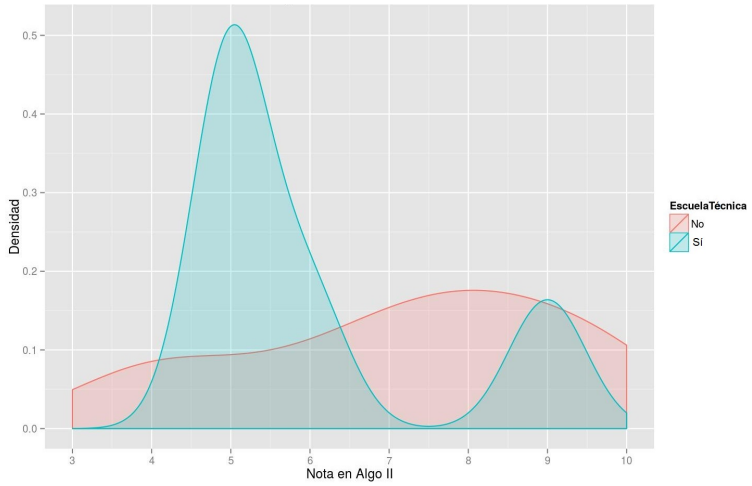


Figura 44: Gráfico de densidad de las calificaciones de la materia Algoritmos y Estructura de Datos II para las alumnas de género femenino, distinguiendo aquellas a las que provienen de escuela técnica y los que no. A pesar de lo que parecen ser diferencias importantes entre las calificaciones de ambos grupos, los tests estadísticos afirman que no las hay.

Esta vez la distribución de las alumnas de escuela técnica muestra ser bimodal con dos picos en los valores cinco y nueve. Por el lado de las de escuelas con otras orientaciones, la distribución es bastante equitativa para varias notas, oscilando por valores entre cuatro y nueve, con un pequeño pico en ocho. La diferencia de notas entre ambos grupos aquí parece ser un poco contrastante si miramos el gráfico aunque los tests de Student y Kolmogorov–Smirnov vuelven a indicar que no existen diferencias estadísticamente significativas entre las notas de los grupos ($p < 0.223$ y $p < 0.308$ respectivamente).

Siguendo la línea de calificaciones en materias de la carrera, realizamos el análisis correspondiente para Algoritmos y Estructura de Datos III, materia del tercer año. El gráfico de la Figura 45 se observa que la distribución para las alumnas provenientes de escuela técnica tiene claros picos en los valores siete, ocho y nueve. Para el caso de las alumnas que estudiaron en escuelas con otras orientaciones, los valores oscilan entre cuatro y diez, con un pequeño pico en 9.5.

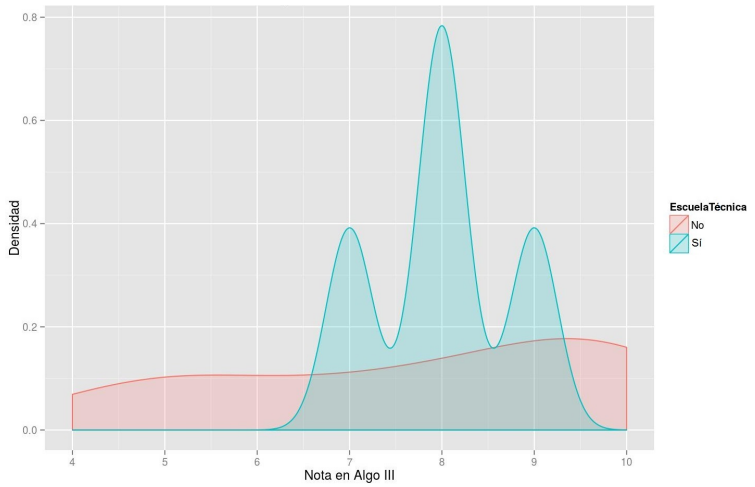


Figura 45: Gráfico de densidad de las calificaciones de la materia Algoritmos y Estructura de Datos III para las alumnas de género femenino, distinguiendo aquellas a las que provienen de escuela técnica y los que no. A pesar de lo que parecen ser diferencias significativas entre las calificaciones de ambos grupos, los tests estadísticos afirman que no las hay.

En cuanto a los tests estadísticos para este caso indican que, al igual que para las otras materias, no hay diferencias estadísticas significativas para estos dos grupos en esta materia ($p < 0.724$ para Student y $p < 0.942$ para Kolmogorov–Smirnov).

Otro punto interesante para observar es cuanto afirman las alumnas que les costó superar el Ciclo Básico Común para cada uno de estos dos grupos, ya que podríamos asumir que las alumnas de escuela técnica contarían con mayor facilidad dado el tipo de conocimiento impartido según su orientación. En la encuesta se les solicitó a las encuestadas que indiquen el nivel de dificultad (según su propia percepción) del CBC exigido por la Universidad de Buenos Aires para ingresar a la carrera. En la figura 46 se puede observar que si bien ambos grupos de alumnas votaron (en su mayoría) entre 4 y 5, las alumnas provenientes de escuelas técnicas parecen votar más estos valores en particular que las que no. De todas maneras, no parece haber grandes diferencias. El hecho de venir de una escuela con una orientación técnica o no, no influye en este aspecto. Los tests de Student y Kolmogorov–Smirnov confirman esto último ($p < 0.635$ y $p < 0.935$ respectivamente).

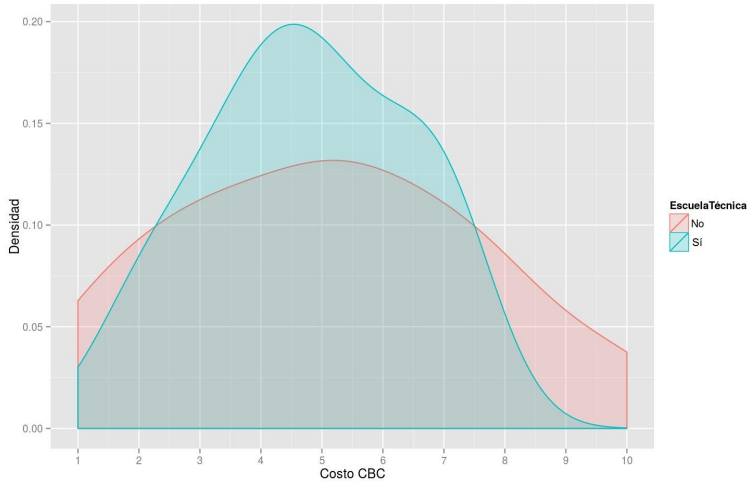


Figura 46: Gráfico de densidad de los valores que representan la dificultad en superar el CBC para los alumnos de género femenino, distinguiendo aquellas que provienen de escuela técnica y las que no. Se puede observar como las alumnas de escuela técnica llegan al CBC con una mejor preparación dada su orientación, aunque no hay diferencias estadísticas significativas entre ambos grupos.

Por último, en la encuesta se preguntó '¿Tenías temas de programación en la secundaria?'. Veamos qué sucede con las respuestas a esta pregunta para ambos grupos de alumnas. En el gráfico de densidad de la figura 47 podemos observar una diferencia bastante notoria entre ambos grupos. Pocos casos aislados de las alumnas de secundarias con otras orientaciones pudieron tratar temas de programación en sus escuelas. Por otro lado, las alumnas provenientes de secundarias con orientación técnica son mucho más propensas a haber visto temas de programación durante la duración de la misma.

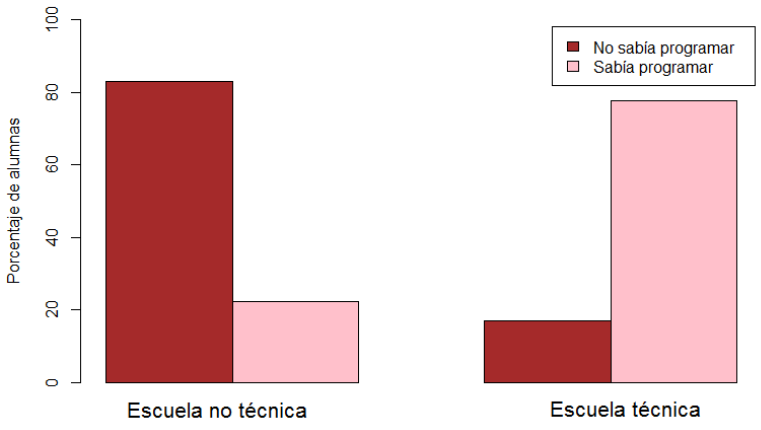


Figura 47: Porcentaje de alumnas que sabían programar desde el secundario, distinguiendo aquellas a las que provienen de escuela técnica y los que no. La diferencia es notoria: gran parte las alumnas que provienen de escuelas técnicas sabían programar antes de ingresar a la carrera. Para aquellas que provienen de escuelas con otras orientaciones ocurre lo opuesto.

Los tests de Student y Kolmogorov–Smirnov confirman estadísticamente la diferencia notoria entre las alumnas que tuvieron temas de programación en la secundaria de ambos grupos ($p < 0.0029$ y $p < 0.00757$ respectivamente). Por lo tanto, como esperábamos, podemos afirmar que las alumnas provenientes de escuelas técnicas suelen ingresar a la carrera sabiendo programar o habiendo realizando algo similar, aunque ya vimos que esto no necesariamente implica que su éxito académico vaya a ser mejor que aquellas que no contaban con este conocimiento previo.

4.4.3. Escuela pública vs privada

En esta sección analizaremos las propiedades de los alumnas que asistieron a una escuela pública comparándolas con las que asistieron a una privada. Realizamos esta distinción con el fin de detectar si existe alguna diferencia significativa a la hora de analizar diferentes aspectos del éxito académico, como hicimos en las secciones anteriores, como por ejemplo las calificaciones en algunas materias de la carrera.

Veremos que los tests estadísticos confirman que no existen diferencias a tener en cuenta para las calificaciones de distintas materias, ni tampoco en otros aspectos como la dificultad para superar el Ciclo Básico Común exigido por la UBA.

Comenzamos analizando las calificaciones en la materia Álgebra I para estos dos grupos. En el gráfico de densidad de la Figura 48 se observa que las alumnas que provinenen de escuelas privadas parecen tener, en general, calificaciones más centradas entre seis y ocho. Las estudiantes de escuelas públicas parecen presentar un promedio bastante equitativamente distribuido por los valores cuatro y ocho. A pesar de esta distinción que podemos hacer tan solo observando el gráfico, los tests estadísticos de Student y Kolmogorov–Smirnov indican que no existen diferencias estadísticamente significativas en este aspecto ($p < 0.925$ y $p < 0.802$ respectivamente).

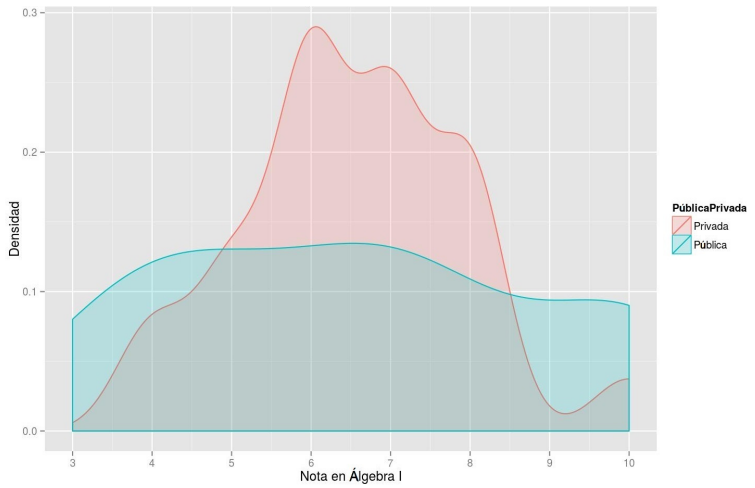


Figura 48: Gráfico de densidad representando las calificaciones en la materia Álgebra I, distinguiendo aquellas que provienen de escuela pública y las de escuela privada. Los tests estadísticos indican que no existen diferencias estadísticamente significativas entre las calificaciones obtenidas por cada uno de estos dos grupos.

Continuando con el análisis de calificaciones en materias, veamos que sucede para la materia Algoritmos y Estructura de Datos II. En la Figura 49 se puede observar como parece que las alumnas de escuela pública presentan una moda entre los valores ocho y nueve, mientras que las de escuela privada tienen una mucho más amplia, que abarca los valores entre cinco y ocho. A pesar de esto, al correr los tests estadísticos, tambien obtuvimos resultados no significativos para esta materia ($p < 0.250$ para el test de Student y $p < 0.852$ para el de Kolmogorov–Smirnov).

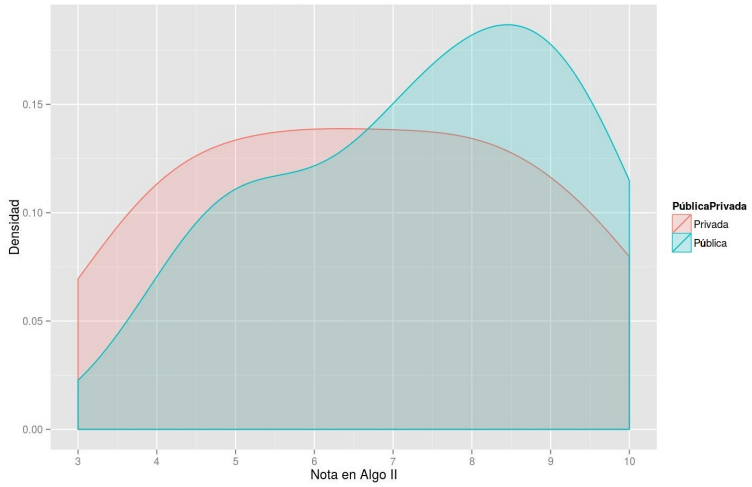


Figura 49: Gráfico de densidad representando las calificaciones en la materia Algoritmos y Estructura de Datos II, distinguiendo aquellas a las que provienen de escuela pública y las de escuela privada. A simple vista las alumnas provenientes de escuelas publicas obtienen mejores notas. Los tests estadísticos indican que no existen diferencias significativas entre las calificaciones obtenidas por cada uno de estos dos grupos.

Con respecto a la materia Algoritmos y Estructura de Datos III. En la Figura 50 se puede observar cómo las alumnas de escuelas públcias presentan una distribución bimodal con modas en los valores 5 y 8. Por su lado, las alumnas provenientes de escuelas privadas, también presentan una distribución bimodal con modas en los valores 6 y 9.5. Sin embargo, los tests estadísticos indican que existan diferencias estadísticamente significativas entre las calificaciones de estos dos grupos ($p < 0.606$ para el test de Student y $p < 0.840$ para el de Kolmogorov–Smirnov).

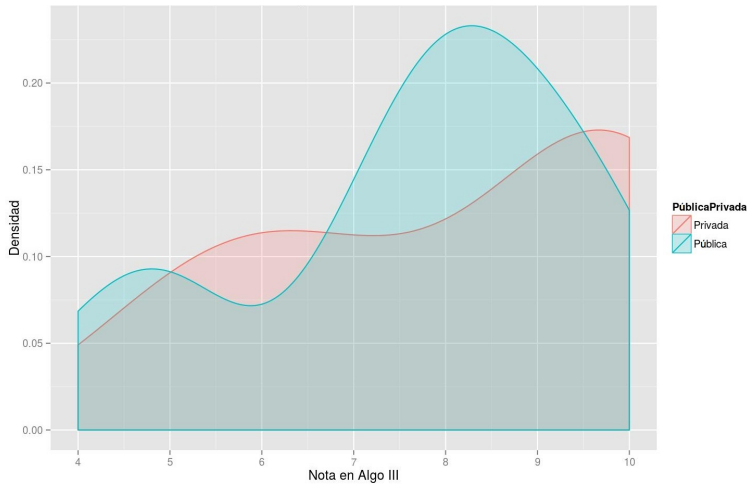


Figura 50: Gráfico de densidad representando las calificaciones en la materia Algoritmos y Estructura de Datos III, distinguiendo aquellas que provienen de escuela pública y de escuela privada. Los tests estadísticos indican que no existen diferencias estadísticamente significativas entre las calificaciones obtenidas por cada uno de estos dos grupos.

Veamos qué sucede para estos dos grupos en cuanto a cuanto piensan las encuestadas que les costó superar el CBC.

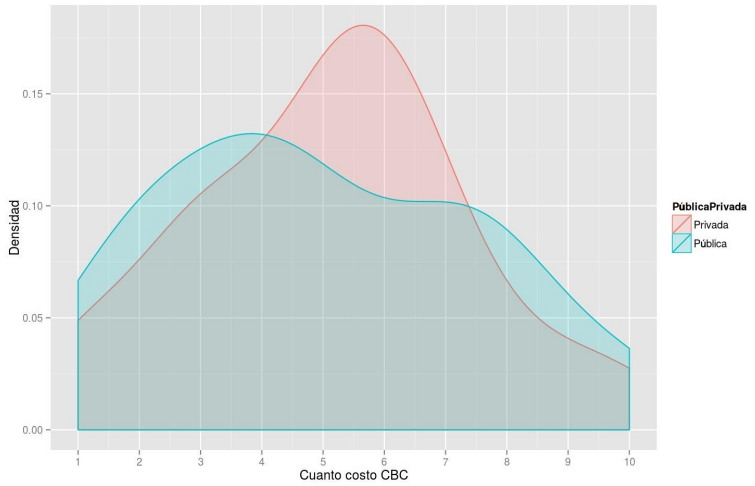


Figura 51: Gráfico de densidad representando cuanto creen las alumnas que les costó el CBC, distinguiendo aquellas que provienen de escuela pública y las de escuela privada. Los tests estadísticos indican que no existen diferencias significativas entre los valores escogidos por cada uno de estos dos grupos.

En la Figura 51 podemos observar que en la distribución de los votos para las alumnas de escuela pública hay una moda en el valor 3.5 y otra pequeña en 7.5. Por el lado de las alumnas de escuela privada, la moda en su distribución se encuentra en el valor 5.5. Como sucedió para las calificaciones de las diferentes materias de estos dos grupos, los tests de Student y Kolmogorov–Smirnov no reflejan diferencias significativas para la media ni la distribución en este aspecto

(p <0.86 para el test de y p <0.96 respectivamente).

Sabemos que de las entidades públicas se suele decir que se debe saber administrar mejor el tiempo, dado que en las privadas se realiza un seguimiento más personalizado y guiado de cada uno de los alumnos. Teniendo esto en cuenta decidimos observar que sucedía con los valores votados por las alumnas en cuanto a cuánto sienten que les alcanza el tiempo en función de sus expectativas académicas. Siguiendo la línea de lo dicho anteriormente, esperábamos encontrar que las alumnas de escuelas públicas hayan optado por elegir valores más altos que aquellas alumnas que estudiaron en escuelas privadas (los valores iban de 1 a 10, con 1 representando que el tiempo le alcanza muy poco y 10 que le alcanza perfectamente).

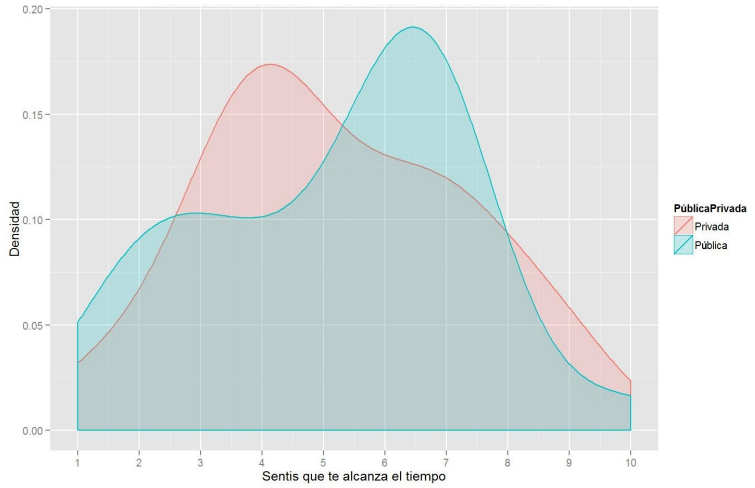


Figura 52: Gráfico de densidad representando cuánto creen las alumnas que les alcanza el tiempo en función de sus expectativas académicas, distinguiendo aquellas que provienen de escuela pública y las de escuela privada. Los tests estadísticos indican que no existen diferencias significativas entre los valores escogidos por las alumnas de cada uno de estos grupos. Sin embargo, pareciera que las alumnas de escuelas publicas piensan que les alcanza mejor el tiempo.

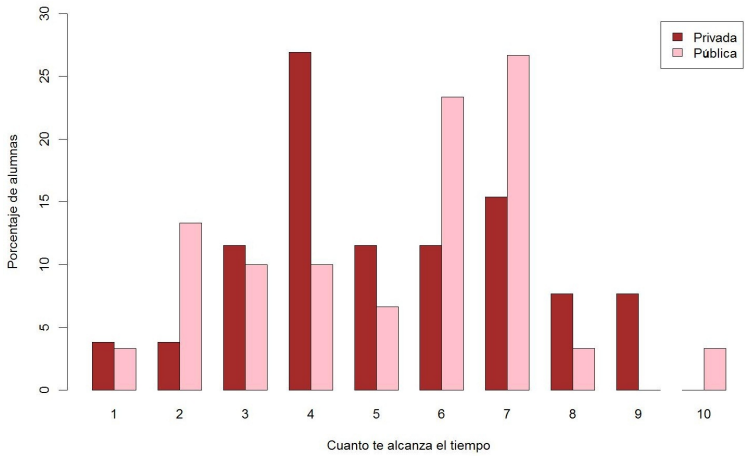


Figura 53: Gráfico de barras mostrando más en detalle los valores escogidos por las alumnas de cada uno de estos grupos para calificar cuánto sienten que les alcanza en tiempo en función de sus expectativas académicas. Se observa como, al igual que el gráfico de densidad de la Figura 52, las alumnas de escuela pública eligieron valores 6 y 7 en su mayoría, mientras que el valor 4 predomina para las de escuela privada.

En la Figura 52 podemos observar como las alumnas de escuelas públicas presentan, en su distribución de votos, su moda en el valor 7.5, mientras que el valor 4 es el de mayor frecuencia para las alumnas de escuela privada. A simple vista, parecería que las alumnas de escuela pública saben administrar mejor el tiempo que tienen para alcanzar sus expectativas académicas. También realizamos un gráfico de barras para analizar más en detalle los valores escogidos por las alumnas de uno y otro grupo (Figura 53). Sin embargo, los tests estadísticos afirman que tal diferencia no es estadísticamente significativa ($p < 0.989$ para el test de Student y $p < 0.936$ para el de Kolmogorov–Smirnov).

4.4.4. 'Le interesa el promedio' vs 'Lo dejaría de lado para recibirse en menos tiempo'

En esta sección analizaremos las propiedades de las alumnas a las que les interesa el promedio comparándolas con aquellas que lo dejarían de lado si esto implica recibirse más rápido. Analizaremos únicamente como estas particiones se corresponden con las calificaciones obtenidas en las materias por las que se preguntó en la encuesta correspondiente. Observamos que las alumnas a las que les interesa el promedio obtienen mejores calificaciones, y confirmaremos estos resultados con los tests estadísticos que estuvimos utilizando.

Comenzaremos comparando las calificaciones obtenidas por cada grupo de esta partición en la materia Algebra I. Como vemos en la Figura 54, el grupo de alumnas a las que le interesa el promedio presenta una moda notoria en el valor 7. Por el lado de las alumnas que dejarían de lado el promedio para obtener el título más rápido, la moda se encuentra entre los valores 5.5 y 6. El test estadístico de Student indica que la diferencia de las medias de las calificaciones entre ambos grupos es significativa ($p < 0.022$). Mirando el gráfico, la diferencia no parece tan notoria, esto puede deberse a que Álgebra I es una de las materias que se deben cursar durante el primer cuatrimestre de la carrera (acorde al plan de estudios y correlatividades), y la postura sobre preocuparse o no por

el promedio de las alumnas pudo haber cambiado en el tiempo para aquellas que hayan completado la encuesta tiempo después de rendir el final de dicha materia.

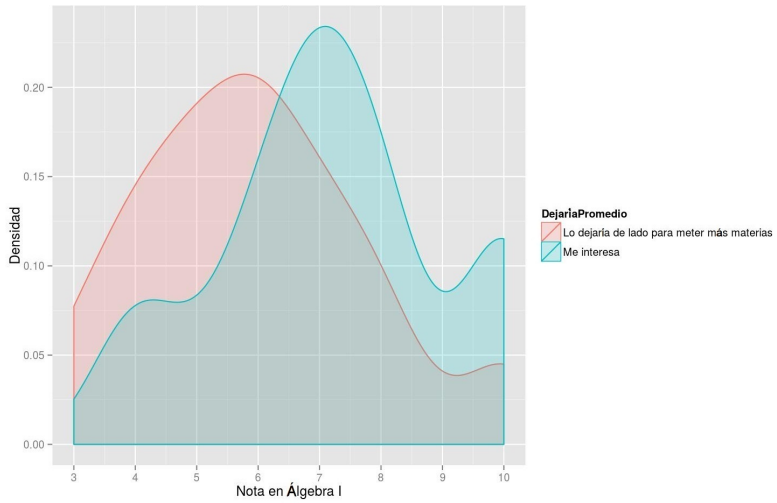


Figura 54: Gráfico de densidad las calificaciones obtenidas en la materia Algebra I, distinguiendo aquellas a las que indicaron que les interesa el promedio y las alumnas que lo dejarían de lado para obtener el título más rápidamente. La moda de la distribución del primer grupo se centra en el valor 7, mientras que para las que dejarían de lado el promedio , su moda se centra entre los valores 5.5 y 6.

Continuamos analizando calificaciones, esta vez de la materia Algoritmos y Estructura de Datos II. En este caso, también se observan diferencias entre los dos grupos, esta vez un poco más notorias en cuanto a los valores de las calificaciones. En la Figura 55 se observa cómo la distribución para las calificaciones de las alumnas a las que les interesa el promedio es bimodal, con modas en los valores 5 y 9. Por el lado de las alumnas que dejarían de lado el promedio para recibirse antes, la distribución también es bimodal con modas en los valores 4 y 7. Notamos también que la moda en 9 de las alumnas a las que les interesa el promedio es mucho mayor que aquella centrada en 5, mientras que para el otro grupo no parece haber mucha diferencia entre las dos modas de su distribución. Esta vez, los tests estadísticos de Student y Kolmogorov–Smirnov coinciden en que la diferencia entre las calificaciones de uno y otro grupo es significativa ($p < 0.000365$ y $p < 0.00906$).

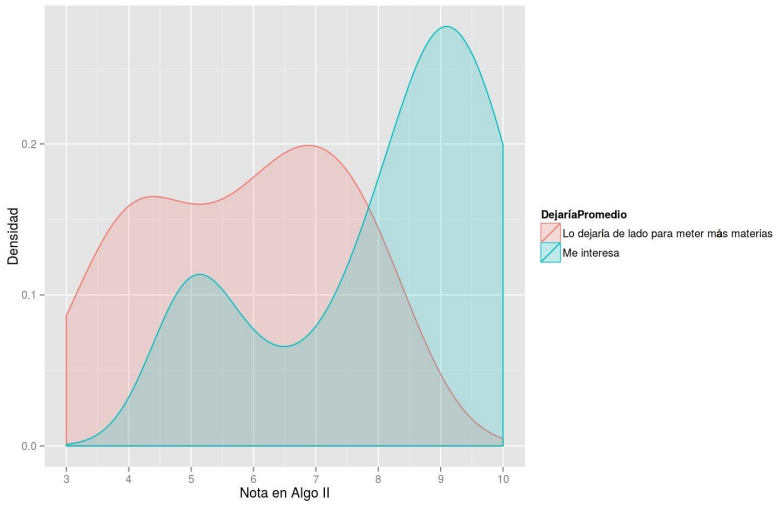


Figura 55: Gráfico de densidad las calificaciones obtenidas en la materia Algoritmos y Estructura de Datos II, distinguiendo aquellas a las que indicaron que les interesa el promedio y las alumnas que lo dejarían de lado para obtener el título más rápidamente. El primer grupo presenta una distribución bimodal en los valores en los valores 5 y 9. Las alumnas que dejarían de lado el promedio presentan una distribución bimodal en los valores 4 y 7.

Por último observemos que sucede para las calificaciones de la materia Algoritmos y Estructura de Datos III para estos dos grupos. En este caso la diferencia es más notoria aún que para las dos materias analizadas anteriormente. En la Figura 56 podemos ver como aquellas alumnas a las que les interesa el promedio presentan una distribución bimodal en los valores 8 y 9.7, mientras que aquellas que dejarían de lado el promedio tienen una moda en 5.5 para su distribución. Ambos tests significativos confirman estas diferencias entre las calificaciones de uno y otro grupo ($p < 0.00421$ para Student y $p < 0.0410$ para Kolmogorov–Smirnov).

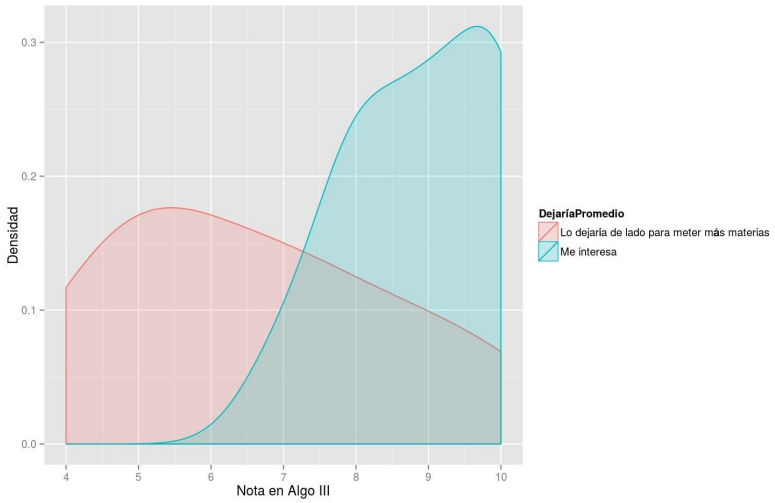


Figura 56: Gráfico de densidad las calificaciones obtenidas en la materia Algoritmos y Estructura de Datos III, distinguiendo aquellas a las que indicaron que les interesa el promedio y las alumnas que lo dejarían de lado para obtener el título más rápidamente. El primer grupo presenta una distribución bimodal en los valores en los valores 8 y 9.7. Por su parte, las alumnas que dejarían de lado el promedio presentan una distribución con moda en el valor 5.5.

4.5. Análisis de correlación

Realizamos un análisis de correlación entre las variables obtenidas en la encuesta. Cada correlación entre dos variables, tiene asociado un p-value proveniente del test de hipótesis correspondiente. La hipótesis nula en el caso de correlaciones es que las variables son independientes (en otras palabras, no existe correlación entre ellas). Dado que tenemos sólo 56 casos en nuestro dataset, es interesante realizar un ajuste a la significancia de la correlación. El ajuste elegido fue la corrección de Bonferroni, el cual ajusta los p-values de la correlacción de todo par de variables. El ajuste realizado por el método es muy conservador, y toma como nuevo α la división de nuestro $\alpha = 0{,}05$ original por la cantidad de casos del dataset. Luego, nos quedaremos sólo con las correlaciones que son significativas (aquellas cuyo su p-value es menor a nuestro *alpha* ajustado). Una vez que tenemos el conjunto de correlaciones significativas, también decidimos quedarnos sólo con aquellas cuyo valor de correlación es mayor a 0.3 en valor absoluto.

Feature 1	Feature 2	Correlación
EscuelaTecnica	ProgramacionSecu	0.599
InglesAntesDeEntrar	InglesActual	0.671
InglesAntesDeEntrar	CostoCBC	-0.544
HorasTrabaja	Trabaja	0.814
HorasTrabaja	TiempoAlcanza	-0.581
ContentaRendimiento	TiempoAlcanza	0.811
ContentaRendimiento	FactoresTiempo	-0.647
CharlaDC	InfluyoCharla	0.837
OtraCarreraDeSistemas	OtraCarrera	-0.624

Cuadro 14: Correlaciones moderadas entre features

Podemos observar fuerte correlación positiva entre 'ContentaRendimiento' y 'TiempoAlcanza', asumimos que puede ser debido a que una mejor organización lleva a un mejor desempeño y esto mejora tanto el rendimiento como el aprovechamiento del tiempo. Otra correlación interesante es la que existe entre 'CharlaDC' y 'InfluyoCharla', indicando que las charlas que se dan en el Departamento suelen tener efecto en motivar, informar e incentivar a las mujeres a seguir una carrera de computación.

Dentro de las correlaciones negativas podemos encontrar que a las que sabían inglés antes de entrar a la carrera, les costó menos el CBC ('InglesAntesDeEntrar' comparado contra 'CostoCBC'). Este factor quizá esté ligado a que las aquellas que tuvieron inglés en el secundario posiblemente tuvieron otra preparación mental, o una dificultad extra en esta instancia, que hizo que se les dificulte menos la etapa del Ciclo Básico que propone la Universidad de Buenos Aires.

Otra correlación interesante es que las mujeres que cursaron total o parcialmente otra carrera antes de inscribirse en computación, cursaron una carrera no relacionada a Sistemas. Esto puede indicar un cambio de intereses de la persona, posiblemente debido a que conocieron la carrera de computación luego de inscribirse a la primera carrera.

5. Conclusiones

En el presente trabajo se presentó una investigación sobre los alumnos y graduados de la carrera de Ciencias de la Computación del Departamento de Computación del FCEyN. El trabajo se dividió en tres partes: el estudio acerca de una encuesta al alumnado general, el estudio de datos del Departamento de Alumnos y el estudio de una encuesta especificamente enfocada a alumnas.

En el estudio hecho a partir de la encuesta al alumnado en general, logramos comprobar algunas hipótesis que se suelen plantear en las charlas de divulgación de la misma y preguntas comunes de futuros ingresantes. Dentro de las dudas comunes de los ingresantes se encuentra la creencia que hay que tener conocimientos previos de programación antes de ingresar a la carrera para tener éxito en la misma. En nuestro análisis pudimos ver como ésto no es cierto ya que **la falta de conocimientos previos no influye significativamente en la postergación de la carrera por parte de los alumnos**. También pudimos comprobar algunos resultados que suelen mencionar los docentes de las diferentes materias, como el hecho de que **el realizar los ejercicios de las guías prácticas influye positivamente en el éxito académico de los alumnos logrando que un porcentaje menor de alumnos postergue la carrera**. También pudimos comprobar **la amplia salida laboral que tiene la carrera y cómo ésta afecta al rendimiento académico de forma significativa** (los alumnos que trabajan tienden a postergar más la carrera que aquellos que no lo hacen).

A partir de los datos existentes pudimos construir una feature, 'PosteraCarrera', que indica si un alumno cumple los plazos determinados por el plan de estudios de la carrera. Utilizando este nuevo feature, construímos un método predictivo para poder inferir si un alumno es propenso a postergar o no la carrera. Obtuvimos buenos resultados en las predicciones, en los cuales los predictores más fuertes resultaron ser el hecho de usar Internet para el estudio, las horas que trabaja el alumno, si pensó dejar la carrera alguna vez, si mantuvo un mismo grupo de trabajos prácticos durante varias materias y (aunque en menor medida) si tuvo o no alguna beca. Asimismo, los métodos estadísticos utilizados mostraron que los conocimientos previos de programación son un factor irrelevante para predecir nuestro feature, lo cual se condice con lo mencionado anteriormente.

Por último, pudimos análizar la percepción propia de los alumnos. Los que suelen postergar la carrera son autocríticos y, en su mayoría, piensan que les va peor que aquellos para los cuales el feature indica que se encuentran al día con el plan de estudios. **La percepción de los alumnos sobre su propio rendimiento, por más que sea subjetiva, se corresponde con la realidad en gran medida**.

Para la segunda parte del trabajo, en el estudio realizado sobre los datos de los alumnos provisto por el Departamento de Alumnos de la Facultad, logramos identificar factores específicos que relacionan a los alumnos, las materias y la condición de regularidad. Los alumnos tienden a rendir y desaprobar menos las materias más avanzadas de la carrera. **No hay diferencias significativas entre alumnos de Capital Federal y provincia, y tampoco entre los alumnos masculinos y femeninos, salvo en un par de casos puntuales**: los promedios de las materias Álgebra I, Análisis II y Algoritmos y Estructuras de Datos II son mejores para los alumnos de Capital, mientras que en la materia

Análisis II las alumnas tienen a tener un promedio menor a los alumnos de género masculino. Utilizando los datos para analizar la frecuencia de inscripción año a año de las alumnas mujeres, para los años 2000 a 2013 se encontró una tendencia negativa: las mujeres se inscriben cada vez menos a la carrera. Sin embargo, esta tendencia es leve y con información acerca de los años siguientes y las tareas de divulgación de la carrera que se encuentran en proceso, puede llegar a observarse una mejoría.

Comparando el comportamiento de los alumnos en el tiempo y ordén de sus cursadas frente a lo previsto por el plan de estudios, dejamos en descubierto que **el plan de estudios es difícil de cumplir por la mayoría del alumnado**. Materias como Métodos Numéricos y Organización del Computador II, muestran una marcada diferencia entre las fechas en que el plan de estudios estipula que se deberían rendir su examen final y la fecha en la que el promedio de alumnos los rinden. En un trabajo futuro pueden investigarse las causas de éste problema. Como conjetura, la causa puede ser la dificultad de las cursadas y finales, y en el caso de Organización del Computador II, la posibilidad de elegir entre un final o un trabajo práctico como condición de aprobación.

Se observaron que existen grupos de alumnos que terminan de rendir todos los finales pasados varios años más de lo esperado. Esto identifica alumnos que dejan la carrera, posiblemente por cuestiones laborales (como vimos en el análisis de encuestas) y retoman los estudios más adelante. Por último, dentro de la materia Métodos Numéricos hay un grupo de alumnos que decidió rendirla en promedio 5 años luego de lo esperado, posiblemente por una decisión de dejar esa rama de materias para el final de la carrera, en vez de cumplir con lo que estipula el plan de estudios.

Analizamos el orden en cual se rindieron las materias para los alumnos que las rindieron todas. Estos alumnos están más cerca de graduarse que de quedarse libres. Se observó que Análisis II se rinde en fecha para el dos tercios del alumnado, pero el tercio que la posterga, lo hace en promedio siete cuatrimestres. **Vimos que las materias Métodos Numéricos y Organización del Computador II se suelen postergar**. Esto último indica que puede ser útil cambiar el plan de estudios para contemplar este fenómeno. Otro punto a favor para la revisión del plan de estudios es que (sin contar CBC, tesis de licenciatura ni materias optativas), en promedio, los alumnos tardan 5.29 años en rendir los exámenes contra los 4 años que estipula el plan de estudios.

Analizando a los alumnos que quedaron libres, se descubrió que más de la mitad abandona la carrera sin aprobar ningún final y, en segunda instancia, se abandona luego de rendir Análisis II o Álgebra. Esto indica la necesidad de planes de contención, ayuda y motivación para los alumnos en estas etapas de la carrera. Una vez aprobados esos exámenes, la deserción disminuye considerablemente. Encontramos evidencia de que los alumnos que abandonan la carrera tienen peores calificaciones en los exámenes de estas materias iniciales cuando se los compara con los alumnos que lograron graduarse.

Fueron creados clasificadores predictivos para inferir la condición de carrera de cada alumno. Los clasificadores logran buen puntaje F1 (mayor a 0.85). La información recolectada muestra que las clases L y T (libre y terminó), son muy diferenciables, mientras que la clase R (regular) es más complicada de diferenciar. Al analizar las features más importantes, se tienen en cuenta las que tienen información acerca de la fecha de exámen de Algoritmos II, la edad

del alumno y la edad en la cual se inscribió a la carrera. Así también tienen peso la información de las materias Álgebra y Análisis II.

Se creó otro clasificador entrenado con clases L y T, donde las features más importantes resultaron pertenecer a las referidas a Métodos Numéricos y Organización del Computador II, así como también a la cantidad de materias aprobadas por el alumno. Usando este clasificador, se clasificaron los alumnos que se sabe que pertenecen a la clase R. Se logró separar dos grupos, uno más cercano a recibirse y otro a abandonar la carrera. En un trabajo futuro sería posible, usando esta información, determinar estrategias para identificar y ayudar a estos grupos a terminar la carrera y a continuarla, respectivamente.

La tercer y última parte presenta un estudio de las alumnas de la carrera. La poca cantidad de alumnas en la carrera hizo que no tuvieramos una cantidad de datos suficiente para realizar un análisis más profundo y significativo, por lo que el estudio fue mayormente exploratorio. Analizamos factores que dieran indicios de por qué las mujeres eligen o no estudiar Ciencias de la Computación, y factores de su rendimiento académico en la misma.

En una primera etapa, **algunas características que son elegidas por la mayoría de las alumnas para elegir nuestra carrera son el interés por las matemáticas, tanto en la actualidad como en el secundario**. También vimos como gran parte de las encuestadas ya mostraba cierto interés por la computación desde el secundario, lo cual es muy útil de saber para continuar fomentando las tareas de divulgación de la carrera en esta etapa. Por el lado del éxito académico, vimos que **el mayor factor que afecta el rendimiento de las alumnas según su propia percepción, era el hecho de trabajar**. Este reduce el tiempo de estudio y preparación para los examenes o trabajos prácticos que exige la carrera, lo cual coincide con lo visto en el análisis de la primera encuesta.

En un análisis más detallado, separamos a las alumnas en varios grupos de interés para observar diferentes propiedades de los datos. Una separación fue entre las alumnas que trabajaban y las que no. Aunque no existieron diferencias estadísticamente significativas entre la distribución de las calificaciones de los dos grupos, observando los gráficos sí se observaban algunas diferencias. Por lo tanto, no debemos desestimar que el hecho de trabajar pueda influir negativamente en las alumnas que recién se inician en la carrera. Otro detalle es que las alumnas que trabajan suelen sentir que no les alcanza el tiempo tanto como a aquellas que no trabajan actualmente, y esto fue confirmado estadísticamente por los tests realizados.

Observamos como **aquellas alumnas que valoran mucho el promedio de calificaciones que obtienen en las diferentes materias, suelen obtener mejores notas en general que las que indicaron que resignarían el promedio para recibirse más rápido**.

Por otro lado, observamos que no existían diferencias estadísticamente significativas en cuanto rendimiento académico entre aquellas alumnas que provienen de escuelas técnicas y con otras orientaciones. A pesar de esto, en los gráficos observamos que para las materias Algoritmos y Estructura de Datos II y III, las alumnas que no provienen de escuelas técnicas tienen modas más marcadas que las que estudiaron en una escuela de este tipo. El factor que distingue a estos dos grandes grupos es el hecho de saber programar antes de ingresar a la carrera, característica que suelen presentar las alumnas provenientes de escuelas técnicas. Anteriormente vimos como el hecho de saber programar no era un factor

importante a la hora de poder estudiar en la carrera ni tampoco graduarse en la misma. Por esto último podemos concluir que **la orientación de la escuela secundaria de la que provengan las alumnas puede ayudar pero no influye determinantemente en su éxito académico**. A su vez, los gráficos y tests sugieren que a las alumnas provenientes de escuelas técnicas tienen menos dificultad para superar el CBC, según su propia percepción. Esto se suma a lo dicho anteriormente, es probable las alumnas de escuela técnica presenten una mayor facilidad para desarrollarse sobre el comienzo de la carrera.

También separamos a las alumnas entre las provenientes de secundarias públicas y privadas, para observar si existía alguna diferencia en cuanto a rendimiento académico en alguno de los dos grupos.A pesar de que en los gráficos se podían observar diferencias en las modas de la distribución de las calificaciones de ambos grupos, los tests estadísticos afirmaron que no existen diferencias significativas en cuanto el rendimiento académico para las diferentes materias y CBC. Esto es un indicio de que, al menos para el caso de alumnos de género femenino, se puede estudiar en la carrera sin importar de que clase de secundaria se provenga.

Creemos que todos estos resultados son de gran valor e importancia a la hora de comprender las causas de fenómenos como la escasa presencia femenina, deserción y éxito académico de los alumnos en nuestra carrera. Esperamos que estos resultados puedan aportar información valiosa para que cada vez más alumnos puedan elegir nuestra área de estudios para formarse como profesionales, se cumplan sus expectativas y se sientan contenidos. Así como también, lograr que los actuales alumnos de nuestra carrera puedan graduarse en la misma con menos complicaciones.

Referencias

[1] Sylvia Beyer, Kristina Rynes, Julie Perrault, Kelly Hay, and Susan Haller. Gender differences in computer science students. In *ACM SIGCSE Bulletin*, volume 35, pages 49–53. ACM, 2003.

[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[3] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Lori Carter. Why students with an apparent aptitude for computer science don't choose to major in computer science. *ACM SIGCSE Bulletin*, 38(1):27–31, 2006.

[6] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[7] Cámara de Empresas de Software y Servicios Informáticos de la República Argentina. Reporte semestral del sector de software y servicios informáticos de la república argentina. Primer Semestre 2013.

[8] Diego Fernandez Slezak, Pablo G Turjanski, Damián Montaldo, and Esteban E Mocskos. Hands-on experience in hpc with secondary school students. *Education, IEEE Transactions on*, 53(1):128–135, 2010.

[9] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156, 1996.

[10] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[11] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[12] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[13] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[14] Marvin L Minsky and Seymour A Papert. *Perceptrons - Expanded Edition: An Introduction to Computational Geometry*. MIT press Boston, MA:, 1987.

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[16] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[17] Fundación Sadosky. Estudio de la fundación sadosky sobre las mujeres en informática. http://www.fundacionsadosky.org.ar/publicaciones-2#mujeres.

[18] Ellen Spertus. Why are there so few female computer scientists? 1991.

[19] David H Wolpert and William G Macready. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1):67–82, 1997.

6. Anexo 1: Encuesta General

Versión Digital:

Haciendo click aquí se puede acceder a la versión digital de la encuesta

Versión Impresa:

Sección Personal:

1. Género
2. ¿Trabajás?
3. ¿Desde que año trabajás?
4. ¿Cuántas horas trabajás por semana?
5. ¿Cursaste menos materias por trabajar?
6. ¿Hacés algun deporte o tocas algún instrumento?

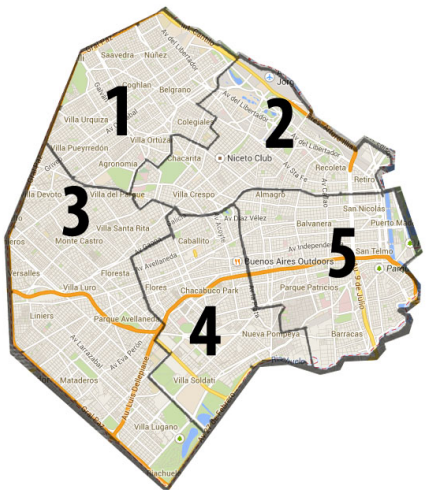
Seccion Facultad:

1. Año de ingreso a la facultad (No considerar CBC)
2. ¿Por qué elegiste la carrera?
   - Porque alguien de tu familia estudia lo mismo
   - Porque te la recomendo un amigo
   - Porque asististe a alguna charla del DC
   - Porque te interesa la investigación
   - Ya conocias de que se trataba la carrera de Cs de la Computación
   - Otra
3. ¿Cuántos cuatrimestres te llevó completar el CBC?
4. ¿Sos egresado?
5. ¿Sos/fuiste docente? (ay2, ay1, JTP, profesor)
6. ¿Sabías programar antes de empezar la carrera?
7. ¿Tenés un grupo de TP con el que hayas compartido varias materias? (más de 5 materias)
8. ¿Estás respetando/respetaste el plan de estudio?
9. ¿Cursaste materias que se solapaban los horarios o correlativas entre sí en un mismo cuatrimestre?
10. ¿Alguna vez pensaste en dejar la carrera?

11. ¿Adeudás el final de Análisis II(C)?
12. ¿Hacés la mayoría de los ejercicios de las prácticas? (En general)
13. ¿Solés hacer preguntas durante la clase?
14. ¿Solés hacer preguntas durante un examen?
15. ¿Tenés o tuviste becas?
16. ¿Hacés cosas relacionadas a la computación en tu tiempo libre?
17. ¿Tenés algún libro de la bibliografía de alguna materia? (comprado o fotocopiado, no digitalizado)
18. ¿Qué materia del CBC te costó más?
19. ¿Cuánto influye Internet en tu éxito académico? (Califique con valor 1-5)
20. ¿Cómo sentís que te está yendo en la carrera? (Califique con valor 1-5)
21. ¿Preferís el perfil de industria o de investigación?
22. ¿Cuántas cursadas (sin final) de materias te falltan para terminar la carrera? (Sin contar optativas)
23. ¿Cuántos finales adeudás?
24. ¿Cuántas veces recursaste en total?

Seccion Transporte:

1. ¿Cuánto tardás en ir a la facultad? (actualmente)
2. ¿Qué usás para venir a la facu?
3. ¿Te mudaste por motivo de las cursadas?
4. ¿En que zona vivís actualmente? (ver mapa para CABA)



Sección Opcional:

1. LU (opcional)

7. Anexo 2: Encuesta Alumnos de género femenino

Versión Digital:

Haciendo click aquí se puede acceder a la versión digital de la encuesta

Versión Impresa:

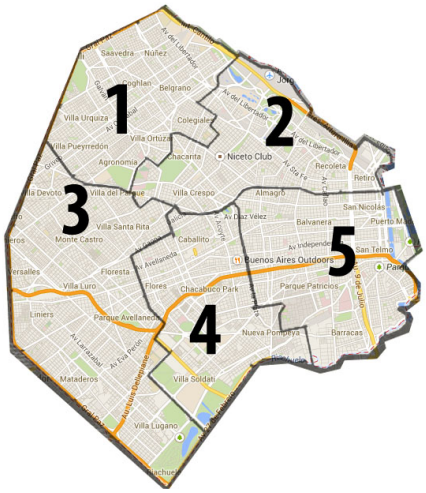
Sección 1:

1. ¿Tu escuela era privada o pública?
2. En tu opinión ¿Cómo te iba en matemática en la secundaria? (Califique con valor 1-10)
3. ¿Estudiaste en una escuela técnica?
4. ¿Te interesan las cuestiones matemáticas de los temas que ves en la carrera?
5. ¿Tenías temas de programación en la secundaria?
6. ¿Mostrabas interés en computación antes de entrar a la carrera?
7. ¿Sabías inglés antes de entrar a la carrera?
8. ¿Como calificarías tu nivel de inglés actual? (Califique con valor 1-10)
9. ¿Trabajás actualmente?
10. ¿Cuántas horas por semana? (Si no trabajás responder 0)
11. ¿Estás contenta con tu rendimiento academico? (Califique con valor 1-10)
12. ¿Sentís que rendis mejor estudiando sola o con un grupo de estudio?
13. ¿Cuántas mujeres hay en tu grupo de estudios / camada de cursadas?
14. ¿Buscás integrarte con mujeres para los TPs siempre que podés o te da igual?
15. En función de tus expectativas académicas (materias que querés aprobar por año, promedio que querés mantener, etc), ¿cuánto sentís que te alcanza el tiempo que le dedicás a la facu? (Califique con valor 1-10)
16. (Si respondiste 6 o menos a la pregunta anterior) ¿Qué factores hacen que no te alcance el tiempo?
17. ¿Te interesa el promedio o lo dejarías de lado si podés meter más materias en menos tiempo?
18. ¿Los conocimientos del trabajo te sirvieron para que alguna cursada se te haga más simple?

19. ¿Qué calificación obtuviste en el final de:
    - Álgebra I
    - Algo II
    - Algo III
20. ¿Cuánto te costó superar el CBC? (Califique con valor 1-10)
21. ¿Te gusta lo que ves en la carrera? (Califique con valor 1-10)

Sección 2:

1. ¿A que edad tuviste acceso a tu primera PC? (aprox)
2. ¿En que zona vivís? (ver mapa para CABA)



3. ¿Cuánto influyó la amplia salida laboral que tiene la carrera en tu decisión de estudiar en el DC? (Califique con valor 1-10)
4. ¿Trabajabas de algo relacionado con sistemas antes de decidir estudiar en el DC?
5. ¿Algun familiar tuyo estudió/trabajó de algo relacionado con sistemas?
6. ¿Asististe a alguna charla del DC antes de elegir la carrera?
7. ¿Pusiste en duda estudiar sistemas por saber que iba a haber mayoría masculina en las clases?
8. ¿Tuviste algun ayudante/profesor de género femenino que te haya inspirado/te haya dado más seguridad para continuar esforzandote en la carrera?
9. ¿Estudiaste otra carrera antes de empezar en el DC?
10. ¿La otra carrera tenía que ver con Sistemas?

11. Cuanto influyo la charla en tu decisión de elegir la carrera? (Califique con valor 0-10. Si no asististe, elegir 0)

Sección Opcional:

1. (Opcional) ¿Cuál es tu promedio (aproximado) en la carrera?
2. (Opcional) ¿Por qué elegiste la carrera?

8. Anexo 3: Features Encuesta Mixta

- **PublicaPrivada**: Fecha en que el encuestado completó la encuesta
- **Género**: Género del encuestado (masculino o femenino).
- **Trabaja**: Indica si el encuestado trabaja actualmente o no
- **AnioDesdeTrabaja**: Indica desde que año trabaja el encuestado.
- **HorasTrabaja**: Indica la carga horaria del trabajo actual del encuestado, tener alguno.
- **CursoMenosMateriasPorTrabajar**: Indica si el alumno encuestado tuvo que disminuir (con respecto a lo que indica el plan de estudios) la cantidad de materias a cursar en algún momento de la carrera, por razones laborales.
- **DeporteInstrumento**: Indica si el alumno encuestado realiza algún deporte o toca algún instrumento musical.
- **PorQueEligioCarrera**: Contiene las razones por las cuales el alumno eligió la carrera. Las opciones posibles se encuentran detalladas en la pregunta 2 de la Sección Facultad de la encuesta mixta, detallada en el Anexo 1
- **CuatrimestresCbc**: Indica cuántos cuatrimestres le llevó al alumno completar el Ciclo Básico Común de la Universidad de Buenos Aires.
- **MateriaCbcCostoMas**: Indica la materia del CBC que al alumno le costó superar más.
- **InfluenciaInternet**: Valor numérico entre 1 y 5 que indica la percepción del alumno sobre cuanto influye internet en su éxito académico.
- **ComoSentisQueTeEstaYendo**: Valor numérico entre 1 y 5 que refleja como el alumno encuestado siente que le va en la carrera, según su propia percepción.
- **IndustriaInvestigacion**: Indica si el alumno se inclina por el perfil de industria o por el de investigación.
- **CursadasSinFinalFaltan**: Indica cuántas materias sin final adeuda el alumno para recibirse.
- **CantFinalesAdeuda**: Indica la cantidad de finales de materias que adeuda el encuestado (contando las materias aún no cursadas).
- **CantVecesRecurso**: Indica cuántas veces recursó alguna materia el alumno encuestado.
- **CuantoTarda**: Indica el tiempo que tarda el encuestado en llegar a la facultad desde su casa o puesto laboral.
- **QueUsa**: Indica qué medios de transporte utiliza el alumno para llegar a la facultad

- **SeMudoPorCursar**: Indica si el encuestado debió mudarse para poder continuar sus estudios en la facultad.
- **ZonaVive**: Indica la zona donde vive el alumno.
- **LU**: El número de libreta universitaria del alumno (opcional).
- **Graduado**: Indica si el encuestado es graduado de la Facultad.
- **EsFueAyudante**: Indica si el encuestado es o fue ayudante de alguna materia de la facultad.
- **SabiaProgramarAntes**: Indica si el alumno poseía conocimientos de programación antes de ingresar a la facultad
- **TieneGrupoTpFiel**: Indica si el encuestado tiene un grupo de trabajos prácticos con el que haya compartido más de 5 materias.
- **SolapaMaterias**: Indica si el encuestado cursó alguna vez materias cuyos horarios se solapaban.
- **RespetaPlan**: Indica si el encuestado respeta el plan de estudios de la carrera.
- **PensoDejarCarrera**: Indica si el encuestado pensó en dejar la carrera en algún momento.
- **HaceLasPracticas**: Indica si el encuestado realiza la mayoría de los ejercicios de las guías prácticas de las materias de la carrera.
- **DebeAnalisis**: Indica si el encuestado adeuda el final de Análisis II(C), materia de primer año de la carrera.
- **PreguntaEnClase**: Indica si el encuestado suele consultar dudas en clase.
- **PreguntaEnExamen**: Indica si el encuestado suele consultar dudas durante un exámen.
- **TuvoBeca**: Indica si el encuestado tiene o tuvo alguna vez, becas.
- **ComputacionEnTiempoLibre**: Indica si el encuestado realiza tareas que tienen que ver con programación en su tiempo libre.
- **TieneLibro**: Indica si el encuestado tiene algun libro impreso (en formato no pdf) de alguna materia de la carrera.

9. Anexo 4: Features Encuesta Mujeres

- **PublicaPrivada**: Indica si la escuela secundaria a la que asistió la alumna encuestada era pública o privada.
- **MatemSecundaria**: Valor numérico entre 1 y 10 indicando como percibe la encuestada que le iba en matemática en la secundaria.
- **EscuelaTecnica**: Indica si la escuela secundaria de la encuestada era técnica.
- **MatemEnCarrera**: Indica si a la encuestada le interesan las cuestiones matemáticas que se ven en la carrera.
- **ProgramacionSecu**: Indica si la encuestada tuvo temas de programación en la secundaria.
- **InteresComputacion**: Indica si la encuestada mostraba interés en computación antes de ingresar a la carrera.
- **InglesAntesDeEntrar**: Indica si la encuestada sabía inglés antes de ingresar a la carrera.
- **InglesActual**: Valor numérico entre 1 y 10 indicando el nivel actual de inglés de la encuestada según su percepción.
- **Trabaja**: Indica si la alumna encuestada trabaja actualmente.
- **HorasTrabaja**: Indica la carga horaria del puesto laboral de la encuestada, en caso de trabajar.
- **ContentaRendimiento**: Valor numérico entre 1 y 10 que indica cuan contenta está la alumna con su propio rendimiento académico.
- **SolaOGrupoDeEstudio**: Indica si la encuestada prefiere estudiar sola o en grupo.
- **GrupoDeEstudio**: Indica la cantidad de mujeres que hay en el grupo de estudio de la alumna encuestada.
- **IntegrarConMujeres**: Indica si la alumna busca integrarse con mujeres a la hora de conformar un grupo de TP siempre que sea posible, o si le es indistinto.
- **TiempoAlcanza**: Valor numérico entre 1 y 10 representando cuánto le alcanza el tiempo que la encuestada le dedica a la facultad segun sus expectativas académicas.
- **FactoresTiempo**: Indica los factores que hacen que el tiempo no le alcance el tiempo a la encuestada.
- **DejariaPromedio**: Indica si la encuestada dejaría de lado el promedio para obtener el título más rápidamente, o si prefiere cuidar el promedio.
- **ConocimientosTrabajo**: Indica si los conocimientos del trabajo le sirvieron a la encuestada para que la cursada de alguna materia le resulte más fácil.

- **Algebra1**: Indica la calificación obtenida en el final de Álgbera I.
- **Algo2**: Indica la calificación obtenida en el final de Algoritmos y Estructura de Datos II.
- **Algo3**: Indica la calificación obtenida en el final de Algoritmos y Estructura de Datos III.
- **CostoCBC**: Valor numérico entre 1 y 10 representando cuánto le costó a la encuestada superar el Ciclo Básico Común de la Universidad de Buenos Aires.
- **GustaCarrera**: Valor numérico entre 1 y 10 que representa cuanto le gusta la carrera a la encuestada.
- **PrimeraPC**: Edad a la que la encuestada obtuvo su propia PC.
- **ZonaVive**: Indica la zona donde vive la encuestada.
- **InfluyeSalidaLaboral**: Valor numérico entre 1 y 10 representanto cuánto influyó la salida laboral en la decisión de elegir la carrera, para la encuestada.
- **TrabajabaSistemas**: Indica si la encuestada trabajaba en algún trabajo relacionado con Sistemas antes de ingresar a la carrera.
- **FamiliarSistemas**: Indica si algún familiar de la encuestada estudió/trabajó en algo relacionado con sistemas.
- **CharlaDC**: Indica si la encuestada asistió a alguna charla del Departamento de Computación de la Universidad de Buenos Aires.
- **MayoriaMasculina**: Indica si la encuestada puso en duda estudiar la carrera por saber de antemano que iba a haber mayoría masculina en las clases.
- **AyudanteInspiro**: Indica si la encuestada tuvo algún ayudante/profesor de género femenino que la haya inspirado/dado más seguridad para continuar esforzándose en la carrera.
- **OtraCarrera**: Indica si la encuestada estudió otra carrera antes de elegir la actual.
- **OtraCarreraDeSistemas**: Indica si la otra carrera elegida por la alumna era del área de sistemas.
- **InfluyoCharla**: Valor numérico entre 0 y 10 que representa cuanto influyó la charla del Departamento de Computación en la decisión de la encuestada de elegir la carrera. En caso de no haber asistido, se califica con 0.
- **Promedio**: El promedio de notas de final de la encuestada.
- **PorqueCarrera**: Contiene las razones por las cuales la encuestada eligió la carrera.

10. Anexo 5: Árbol de Correlatividades de la carrera de Ciencias de la Computación

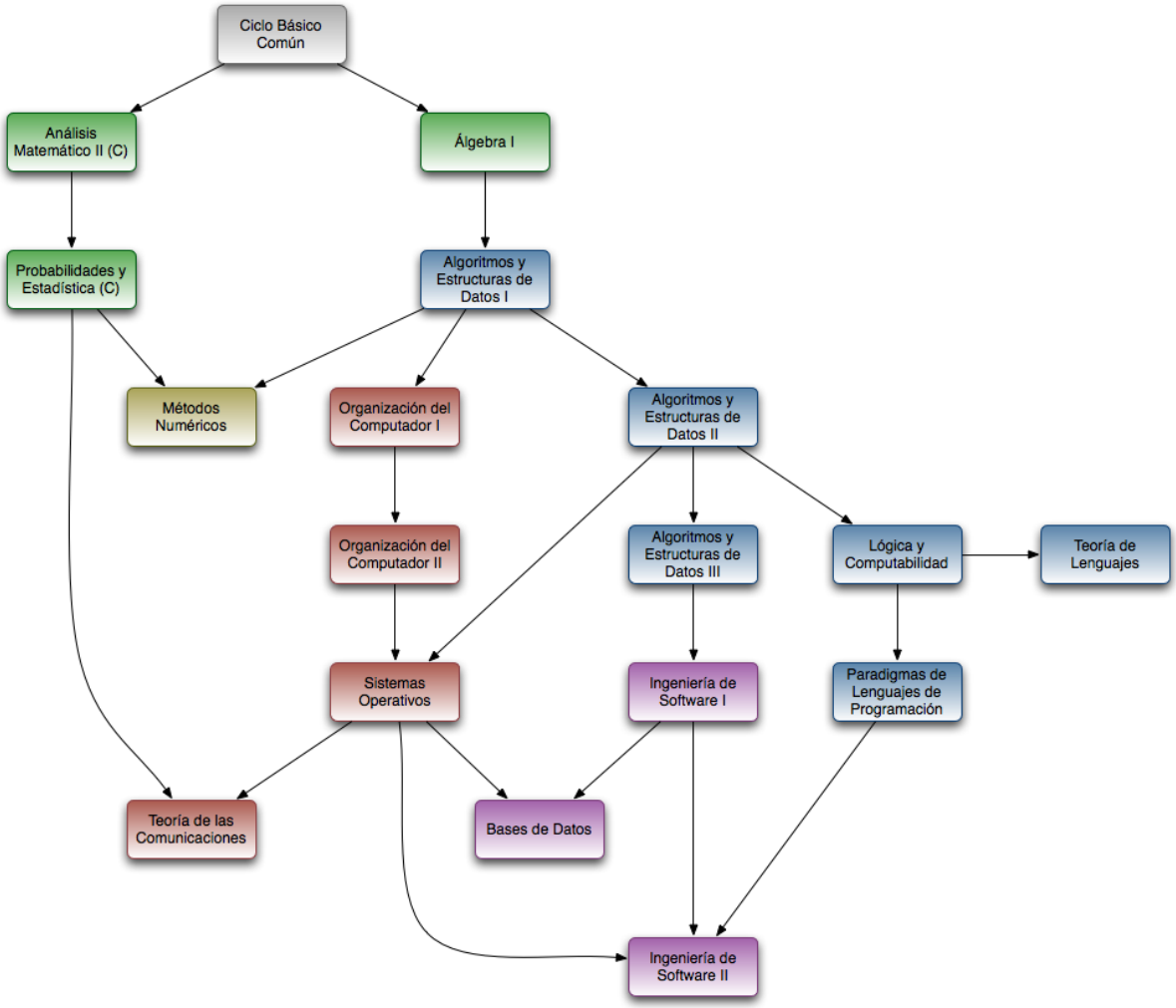


Figura 57: Árbol de correlatividades de la carrera de Ciencias de la Computación de la Universidad de Buenos Aires. Fuente: http://www.cubawiki.com.ar/