Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática

Probabilidad y estadística en el análisis de datos

Proyecto Speed dating

¿Es posible encontrar el amor en 4 minutos?

Profesores: Andrew Hart
Jocelyn Dunstan E.
Server Martinez A.

Auxiliar: Emir N. Chacra

Integrante: Alejandro Cuevas A.
Fecha: 5 de julio de 2019

Índice de Contenidos

1. Introducción

Desde tiempo inmemoriales la conquista entre parejas ha sido una parte fundamental de la sociedad, permitiendo generar familias y comunidades, ayudando a la supervivencia como especia. Hoy en día la sociedad va cambiando donde el "juego" de conquista entre parejas se va adaptando con las nuevas tecnologías, hoy existen aplicaciones de teléfono que permiten conocer gente de forma casi instantánea lo que cambia los paradigmas de como la gente encuentra pareja.

En el presente informe se mostrará la exploración de una base de datos de citas *express*, que consta en una encuesta hecha a los participantes a través de un experimento de citas rápidas donde solo tienen cuatro minutos para conocer a la otra persona y tomar la decisión si quiere una segunda cita, si ambos coinciden se tiene un *Match* y pueden intercambiar información.

El informe se organiza de la forma siguiente: primero se introduce la base de datos en detalle, explicando tanto su origen como qué información contiene, para luego plantear los objetivos principales de la exploración y pre-procesamiento inicial. Posterior a esto se procese a realizar una exploración de las variables y como esto nos da una intuición del comportamiento de los participantes y para qué y cómo usar los modelos a justar, por último se ajustan modelos basados en árboles para predecir en primer lugar la toma de decisión unilateral si desean una segunda cita y segundo para predecir si existirá un match entre las personas dado distintos factores.

2. Base de datos: Speed dating

Durante el periodo $2002 - 2004$ por *Columbia Business School* [1], se formuló un experimento de citas express, las cuales consistían en que los participantes tienen una "primera cita" de cuatro minutos con cada participante del sexo opuesto, al final de los cuatro minutos se les preguntaba a los participantes si les gustaría tener una segunda cita, en tal caso se tiene un *Match*. También se les pedía que calificaran a sus citas dentro de seis atributos: atractivo, sinceridad, inteligencia, sentido del humor, ambición e intereses compartidos.

Sumado a esto, se les pedía contestar una encuesta a lo largo del experimento, donde se incluían campos tales como: edad, raza, auto-percepción de los seis atributos mencionados, que preferencias tenían ellos frente a esos seis atributos con respecto a posibles parejas, y como ellos creían que otras personas valoraban dichos atributos.

El experimento consto de 21 sesiones, donde en cada una se tenían entre 20 a 40 participantes, teniendo un total de 552 personas distintas participando, en las que se desarrollaron 4189 citas a lo largo de todo el experimento.

Esto lleva a que, en términos prácticos se tengan dos filas por cita, una del punto de vista de cada participante, donde se tiene tanto información de la cita como de la encuesta, formando una matriz de 8378 observaciones con 195 variables.

2.1. Objetivos

Como objetivos principales se tiene la exploración de esta base de datos con el fin de extraer patrones de comportamiento de las personas que realizaron el experimento, frente a esto se plantea responder las siguientes preguntas.

- ¿Qué atributos valoran más las mujeres en los hombres (y viceversa)?
- ¿Existe un sesgo dado por la raza u otro contexto social al elegir pareja?
- ¿Es posible tener una impresión fidedigna de una persona en solo cuatro minutos?
- ¿Es posible predecir un match dado como que se conoce como se perciben los participantes?

2.2. Pre-procesamiento

Debido a la larga extensión de la encuesta, donde mide como valoran los participantes los seis atributos mencionados en múltiples ocasiones a través del experimento, la base de datos cuenta con un número elevado de valores faltantes, por lo que es necesario descartar datos puesto que no hay una relación entre filas que permita hacer interpolación.

Para esto se decide un umbral de 400 valores faltantes, donde si un atributo (columna) tiene más de 400 valores faltantes, esta será descartada. Esto reduce las variables de 195 a 85 donde mediante inspección visual se obtiene que la mayoría de las variables eliminadas corresponden a preguntas redundantes sobre como se ellos creen que otras personas valoran los atributos. Una variable interesante que es eliminada en este paso es como valoran la ambición de su pareja, lo que podría indicar que cuatro minutos no son suficientes para conocer de forma consistente a una persona. En la fig.2.1 se muestran los valores faltantes por variable, donde en rojo se muestra el umbral.

Posterior a esto se decide descartar las filas que contienen al menos un valor faltante, de esta forma se puede utilizar un abanico más amplio de modelos, como se verá luego, el costo de tener menos datos disponibles no resulta ser un inconveniente en términos de desempeño o sobeajuste. Esto resulta en una matriz sin valores faltantes de 6777×85.

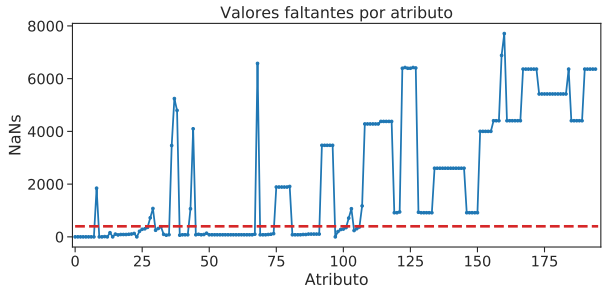

Figura 2.1: Valores faltantes en la base de datos para cada atributo, en rojo se muestra el umbral utilizado para descartar atributos.

Es importante mencionar que en ajuste de modelos solo se utilizaran variables que tengan un sentido, lo cual será detallado más adelante.

3. Metodología y resultados

La metodología se compone en dos grandes partes, la primera consiste en una exploración de variables que pueden dar indicio de un patrón, donde se utilizaran diferentes formas de visualización, luego de esto se plantean modelos basados en árboles para poder predecir tanto un match como la decisión unilateral de querer un match, es decir, cuando una persona quiere una segunda cita, sin tomar en cuenta la decisión del otro.

3.1. Exploración y visualización

La primera pregunta que surge es que tan balanceado están ambos géneros a lo largo del experimento, donde en la fig.3.1 se muestra el histograma de géneros a lo largo de las 21 sesiones, en este se puede ver que existe una proporción similar de hombres y mujeres a lo largo de todo el experimento.

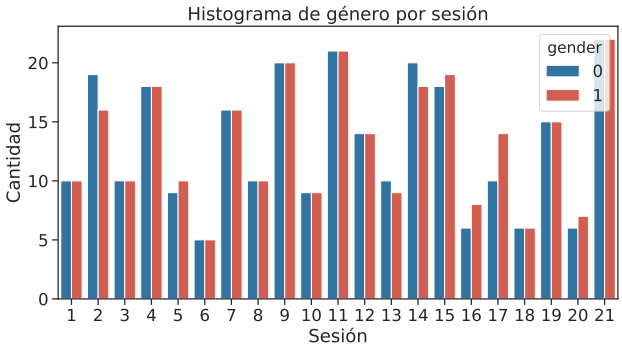

Figura 3.1: Distribución de género según sesiones de citas. Mujer= 0 Hombre= 1.

Con el afán de caracterizar mejor a los participantes, se toma la distribución de edades a través de todo el experimento, donde es de notar que muchas de las personas participaron en más de una sesión por lo que fue necesario primero obtener a las personas únicas antes de poder comparar. Esto se muestra en la fig.3.2, los participantes parecen estar entre el inicio de sus 20 hasta principio de los 30 donde no parece haber mucha diferencia por genero.

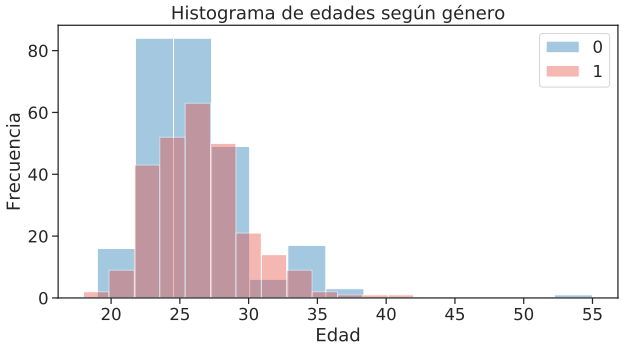

Figura 3.2: Distribución de edad de los participantes según género. Mujer= 0 Hombre= 1.

Una pregunta interesante es la distribución de razas en los participantes, donde tomando nuevamente solo personas únicas es muestra en la fig.3.3 el porcentaje de razas según genero. Existe una preponderancia de gente caucásica, seguido de asíaticos, es de notar que no existe nadie que sea nativo americano.

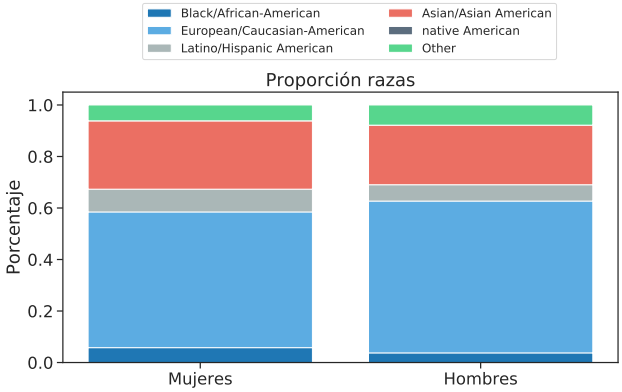

Figura 3.3: Distribución de razas según género.

Dado la estructura que uno podría esperar en cuanto a citas, se calcula el porcentaje de personas que obtuvieron un match, donde en la tabla.3.1 se aprecia que la proporción de matchs es cerca del 20 %, lo interesante aflora al mirar con que frecuencia tanto hombres como mujeres recibían un "no" a la pregunta de la segunda cita, done en la misma tabla se muestra que un hombre es más probable que reciba un "no", indicando en primera instancia que las mujeres son las que en general toman la decisión respecto a si continua la conquista.

Género	Matchs	Rechazo
Mujer	17.45 %	51.43 %
Hombre	17.61 %	61.61 %

Tabla 3.1: Porcentaje de *Matchs* y rechazos según género.

A lo largo de la encuesta le es preguntado a los participantes como son sus preferencias respecto a los seis atributos como también qué piensan ellos que el género opuesto preferencia sobre estos mismos atributos. Recordando que estos atributos son: atractivo, sinceridad, inteligencia, sentido del humor, ambición e intereses compartidos. Se tiene tanto la preferencia de cada persona como la que ellos creen que el genero opuesto tiene.

Esto permite que se puede estudiar la discrepancia entre lo que un genero busca versus lo que el otro *cree*. En la fig.3.4 se muestran los atributos que prefieren las mujeres en un hombre, tanto del punto de vista de las mujeres mismas como lo que los hombres piensan que las mujeres buscan, de este se desprende que hombres piensan que las mujeres valoran más el atractivo sin embargo mujeres parecen preferir la inteligencia y sinceridad en una posible pareja. Viceversa se muestra en la fig.3.5 donde mujeres piensan que hombres valoran mucho el atractivo, que si bien es el atributo que valoran más los hombres en promedio esta sobre estimado por las mujeres. Para ambos casos se puede ver que tanto hombres como mujeres valoran sinceridad e inteligencia mientras que ambos piensan que el otro no valora estos. Esto da indicio a que si un modelo es ajustado usando dentro de los predictores como valoran a su cita en

los seis atributos, se esperaría que el atractivo y la inteligencia sean predictores fuertes. Más adelante corroboraremos si esto se cumple.

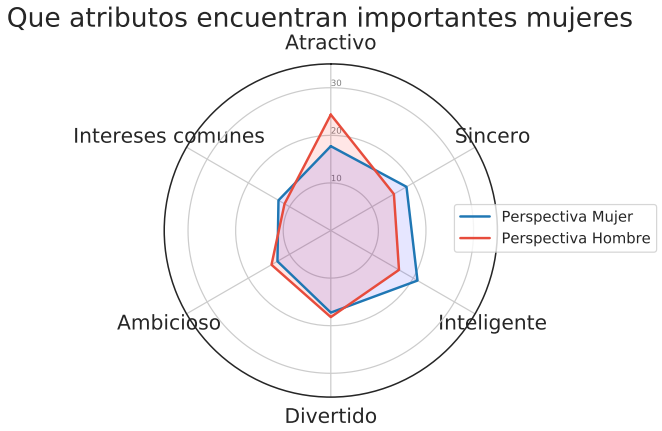

Figura 3.4: Promedio de la encuesta sobre que atributos son importantes para las mujeres, según punto de vista de mujeres y hombres.

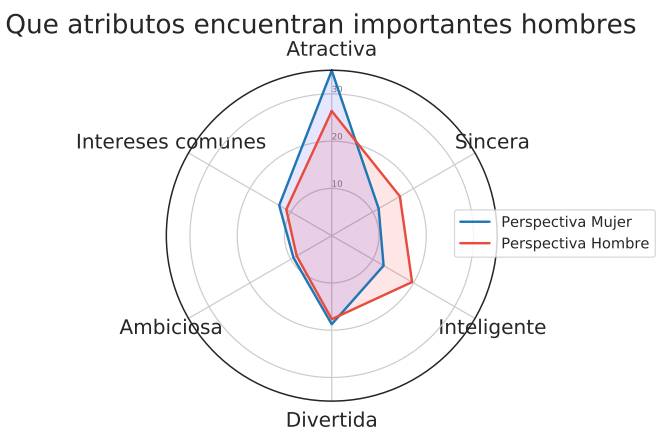

Figura 3.5: Promedio de la sobre que atributos son importantes para las hombres, según punto de vista de hombres y mujeres.

Otro punto importante que esta incluido en la encuesta es como cada persona se auto-valora en estos atributos, en este caso solo serían cinco pues intereses comunes no es algo auto-valorable. Con esto y la valoración que entrega cada participante sobre su pareja es posible estudiar la discrepancia entre como alguien se auto-percibe versus como lo perciben los demás. Para esto es de notar que en la pregunta de valorar a tu pareja, un porcentaje pequeño de gente fue capaz de asignar una valor a la ambición de la otra persona, posiblemente dado que es un concepto un poco más abstracto que los otros, y que en solo cuatro minutos no es posible conocer de forma profunda a una persona, es por esto que no se muestra la diferencia en este campo.

En la fig.3.6 se muestra la distribución de la diferencia entre $(valoracion_{propia} - valoracion_{demas})$ separados por sexo. Valores negativos en la distribución quiere decir una sub-valoración del atributo, es decir que la auto-percepción es más baja de lo que en general opina el sexo opuesto. Con esto podemos ver que en general la auto-percepción es relativamente acertada, con la mediana cerca de 0, en cuanto a diferencias de cada género, podemos ver que la distribución de atributos en mujeres tiende a tener colas pesadas en valores positivos, lo que quiere decir que se consideran con un atributo más alto de lo que los hombres piensan de ellas. También se aprecia que la media de la distribución para los hombres es un poco más baja que para las mujeres sugiriendo una sub-valoración, esto es más notorio en el caso del atractivo.

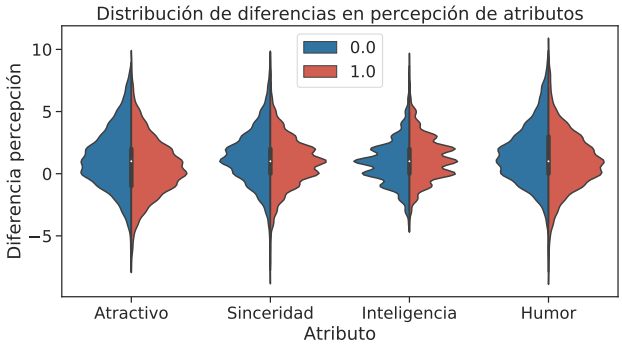

Figura 3.6: Distribución de la diferencia entre la auto-valoración de atributos frente a como los demás valoran. La diferencia es obtenida como (auto-percepción − percepción-pareja) por lo que un valor negativo en la distribución es sub-autovaloración mientras que positivos son sobre-autovaloración. Mujer= 0 Hombre= 1.

3.2. Modelos

Una vez que se tiene una idea general de como distribuyen las variables, se decide intentar predecir tres escenarios: (i) utilizando solo como valoran a su pareja, si tienen la misma raza, correlación de intereses y genero, poder predecir si esa persona desea seguir con una segunda cita, (ii) dado las mismas variables predecir cuanto le gusta a la persona su pareja, en una escala de 0 a 10, por último (iii) utilizando la valoración mutua que tiene una pareja (es decir como ella lo valora a él y él a ella en los 6 atributos) junto con intereses correlacionados, y si se tiene la misma raza, poder predecir finalmente un match. Desde ahora se referirá como atributos a los 4 atributos: Atractivo, sinceridad, inteligencia y humor. Ambición se descartó por falta de respuestas en la encuesta y los intereses mutuos se obtienen mediante la correlación entre respuestas de preferencias de pasatiempos por cada uno.

Es de notar que fueron usadas pocas variables respecto al número inicial de la base de datos, esto se debe a que las demás variables o eran redundancias como volver a incluir preferencia de atributos, o eran variables que no deberían tienen relación de la persona sobre un match, un ejemplo de esto es la variable que dice cuan feliz o satisfecho se está con el experimento total, o si logró establecer una conexión luego de la segunda cita. Es de mencionar que 17 variables correspondiente a gustos independientes de cada personas respecto a hobbies y actividades quedan condensados en la correlación entre intereses.

Para cada experimento se ajustaron tres modelos: Descision Tree, Random Forest y XGBOOST. Donde el primer y tercer experimento corresponden a modelos de clasificación y el segundo de regresión. Todos los modelos fueron entrenados con los mismos datos realizando una partición de entrenamiento y test de 2/3 y 1/3 respectivamente.

Predicción de decisión: para el primer experimento se quiere intentar predecir si, dado la información que tiene la persona de su pareja, va a decidir querer una segunda cita o no, esto es independiente de qué decide la otra persona. Para esto se usan 7 predictores, los primeros 4 es como califica a su cita en los cuatro atributos junto a esto se suma correlación intereses comunes, una variable binaria que indica que ambos pertenecen a la misma raza y por último el género. El género se incluye puesto que al utilizar árboles se espera que si existe diferencia en la forma de tomar decisión para hombres y mujeres,

una rama del árbol permite que se modele esta diferente y así evitar ajustar dos modelos diferentes por genero.

Regresión de gusto: usando los mismos atributos que en caso anterior, se busca predecir cuanto le gusta a un participante su cita, en una escala de 0 a 10.

Predicción de match: Usando ahora la calificación de ambas partes respecto a los *atributos* y descartando el uso de género pues ahora se quiere clasificar una cita en vez de la opinión de una persona, se tienen 10 variables para predecir un match, que es cuando ambas personas dicen sí al primer experimento.

Luego, para cada experimento cada modelo (clasificador o regresor según el caso) fue entrenado bajo 100 inicializaciones distintas, a fin de tener más robustez en el resultado, en el caso de Random Forest y XGBOOST ambos fueron entrenados en cada caso con $n = 50$ árboles. Como métrica de comparación se utilizo el *accuracy* promedio en los casos de clasificación y el error absoluto medio promedio (MAE) para el caso de regresión. En la tabla.3.2 se muestra el resultado de los 100 intentos, para cada modelo, para cada experimento.

Experimento	Modelo	Rendimiento
Clasificación sobre decisión	Decision Tree	$0.655 \pm 3.063 \cdot 10^{-03}$
	Random Forest	$0.705 \pm 4.048 \cdot 10^{-03}$
	XGBOOST	$\mathbf{0.735 \pm 1.443 \cdot 10^{-15}}$
Regresión sobre gusto	Decision Tree	$1.192 \pm 6.509 \cdot 10^{-03}$
	Random Forest	$0.913 \pm 3.905 \cdot 10^{-03}$
	XGBOOST	$\mathbf{0.845 \pm 1.221 \cdot 10^{-15}}$
Clasificación sobre Match	Decision Tree	$0.768 \pm 4.216 \cdot 10^{-03}$
	Random Forest	$0.843 \pm 2.864 \cdot 10^{-03}$
	XGBOOST	$\mathbf{0.853 \pm 9.992 \cdot 10^{-16}}$

Tabla 3.2: Métricas de rendimiento en el conjunto de test para los 3 experimentos con 100 intentos por cada modelo, por cada experimento. Se muestra la media de la métrica de rendimiento y la desviación estándar para los 100 instancias. Se usa Accuracy promedio para los problemas de clasificación y MAE promedio para el de regresión. En negrita se muestra el mejor modelo por experimento

Respecto al primer experimento, se puede ver que el mejor modelo es XGBOOST que alcanza un accuracy promedio del 73 % y una desviación estándar pequeña, indicando que el modelo consistentemente es capaz de predecir si una persona decidirá preguntar por una segunda cita.

Para la regresión sobre cuanto le gusta a una persona su cita, podemos ver que nuevamente XGBOOST es el que tiene mejor desempeño con un error absoluto medio de cercano a 0.8 en una escala de 0 a 10, es interesante estudiar como se relaciona cuanto le gusta a una persona con la decisión de preguntar por la segunda cita. Es de esperar que exista al menos una correlación positiva, lo cual empíricamente no se encuentra ya que, (i) no existe correlación entre cuanto le gusta una persona con la decisión de preguntar con la segunda cita, (ii) la adición de esta variable como predictor no mejora el rendimiento

de los modelos ni en el segundo ni tercer experimento.

Finalmente en el tercer experimento, en donde se quiere predecir un match, XGBOOST termina por ser el mejor modelo, siendo capaz de predecir un match con un 85 % de accuracy para el conjunto de test y con una desviación estándar varios ordenes de mágnitud menores al de Random Forest.

3.3. Ranking variables

Por último, dado que los modelos son capaces de consistentemente predecir tanto una decisión unilateral como un match, se busca en cada caso comparar el ranking de atributos intrinseco dado por los modelos basados en árboles. Para esto, por cada experimento se toma el último modelo entrenado y se evalúa el ranking de características, dado en los tres modelos por la perdida promedio por variable de la función objetivo del árbol, índice gini en el caso de clasificación y error cuadrático medio en el de regresión.

Clasificación de decisión: Se muestra en las fig.3.7 fig.3.8 y fig.3.9 el ranking de variables para Decision Tree (DT), Random Forest (RF) y XGBOOST (XGB) respectivamente. En estos se puede ver que en DT y RF los atributos más importantes parecen ser en primer lugar los intereses comunes y en segundo el atractivo, mientras que en XGB es altamente preponderante el atractivo y lejano aparece el humor.

Es interesante notar que el género no resulta ser una variable decisiva en la decisión, lo cual índica que la forma en que se toman las decisiones no varia mucho entre hombres y mujeres.

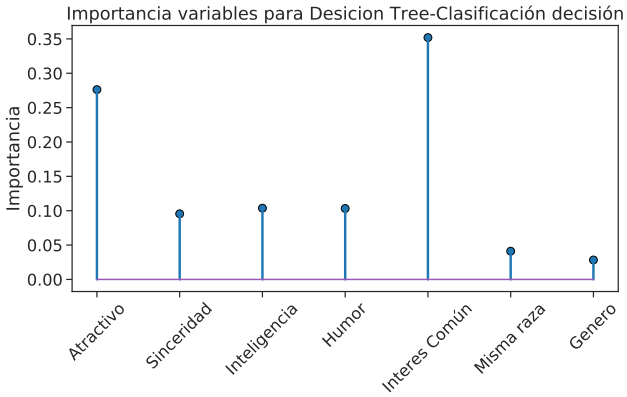

Figura 3.7: Ranking atributos para Decision Tree en clasificación de decisión.

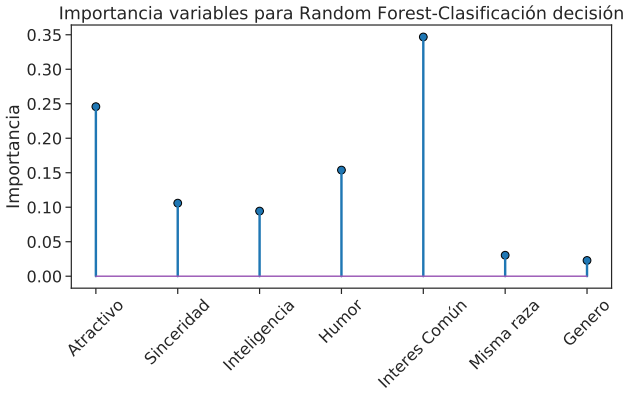

Figura 3.8: Ranking atributos para Random Forest en clasificación de decisión.

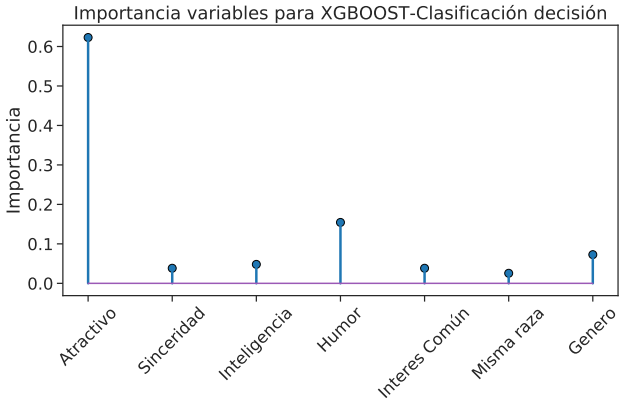

Figura 3.9: Ranking atributos para XGBOOST en clasificación de decisión.

Regresión Gusto: Siguiendo el mismo análisis, en las fig.3.10 fig.3.11 y fig.3.12 se muestra el ranking de atributos para la regresión sobre cuanto le gusta su pareja, en este podemos ver que en los tres modelos el humor muestra ser el atributo más importante. Lamentablemente el gusto no parece estar relacionado con hacer match, por lo que no se puede concluir mucho.

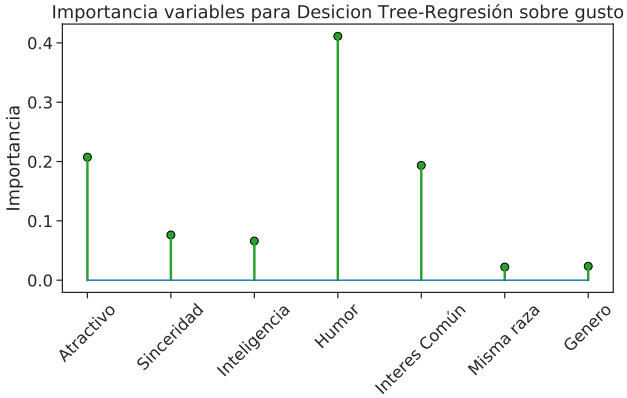

Figura 3.10: Ranking atributos para Decision Tree en regresión gusto.

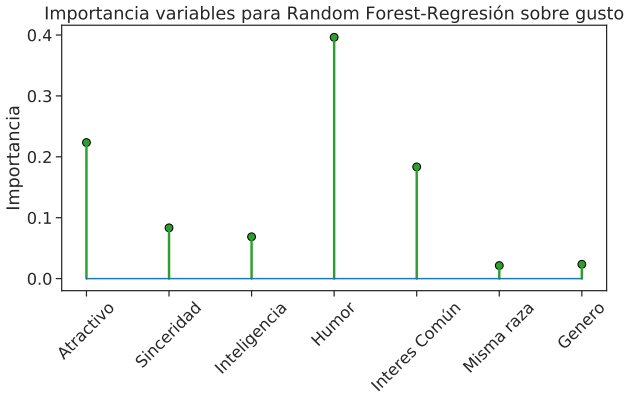

Figura 3.11: Ranking atributos para Random Forest en regresión gusto.

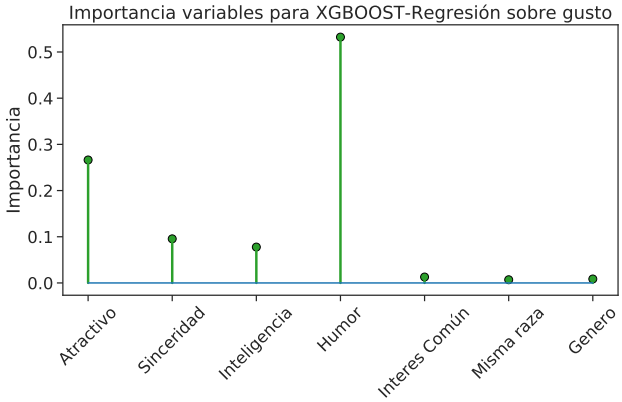

Figura 3.12: Ranking atributos para XGBOOST en regresión gusto.

Clasificación Match: Finalmente se tienen los ranking de atributos el clasificación de match en las fig.3.13 fig.3.14 y fig.3.15. Para el caso de DT y RF se ve que los intereses comunes juegan un rol importante en la clasificación quedando en segundo lugar atractivo y humor. Sin embargo en el caso de XGBOOST el atractivo y humor quedan en primer lugar. En los tres casos es de notar que como existen dos variables asociadas a los atributos (pues se esta viendo la percepción del uno del otro en la cita) y se puede ver que mismos atributos tienen ranking similares para todos los modelos, lo cual muestra consistencia.

El resultado de este último ranking es llamativo al contrastarlo con la valoración de atributos que se busca en la otra persona mostrado en las fig.3.4 y fig.3.5 donde la inteligencia y sinceridad mostraban ser los atributos predominantes en ambos casos. Esto contrasta directamente con lo obtenido en los modelos, donde el sentido del humor y el atractivo muestran ser los factores más predominantes a la hora de tomar la decisión. Una explicación de esto puede ser que, como la citas solo toman cuatro minutos, este tiempo no es suficiente para evaluar la inteligencia y sinceridad del otro, y la decisión debe ser tomada frente a atributos más directos como el atractivo o el sentido del humor al empezar la conversación liviana. Una conclusión puede ser que, si se cuenta con poco tiempo, es mejor usar el sentido del humor que tratar de entablar una conversación profunda!.

Otro punto interesante a notar es que tanto RF como XGB tuvieron desempeños similares a la hora de predecir un match, sin embargo estos usaron variables diferentes para realizar la predicción.

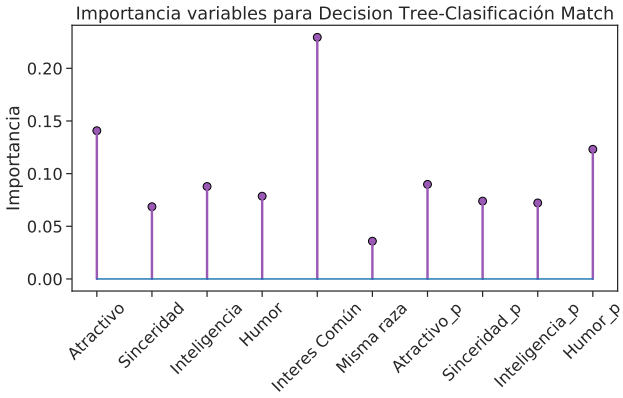

Figura 3.13: Ranking atributos para Decision Tree en clasificación match.

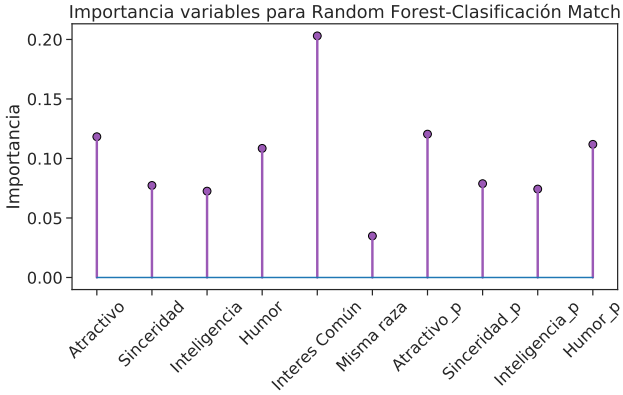

Figura 3.14: Ranking atributos para Random Forest en clasificación match.

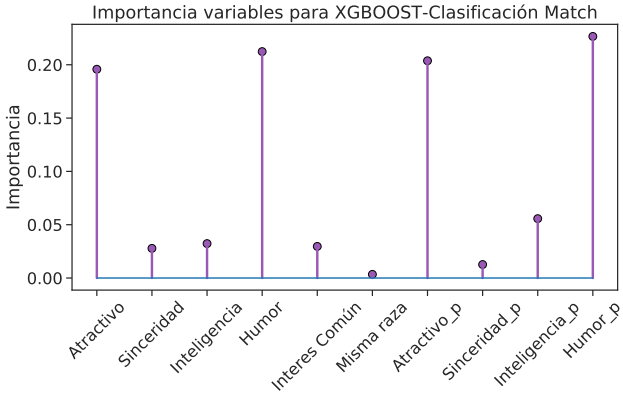

Figura 3.15: Ranking atributos para XGBOOST en clasificación match.

4. Conclusión

Se exploró de forma satisfactoria la base de datos de citas express, pudiendo comparar el comportamiento de hombres y mujeres a la hora de elegir pareja cuando se tiene una restricción de cuatro minutos para entablar la conversación. Todo esto frente a valoraciones y percepciones de cinco atributos: atractivo, sinceridad, ambición e inteligencia. Fue posible entrenar modelos predictivos de forma exitosa y estimar si una pareja hará un match.

Al comparar las diferencias de que parece preferir un genero frente a lo que el otro piensa que prefiere, se puede ver que ambos sexos piensan que el otro valora de sobremanera el atractivo físico, donde en ambos casos se llega a un sobre estimación de la importancia de este atributo. Para mujeres parece ser más importante la inteligencia y sinceridad, mientras que para los hombres el atractivo e inteligencia ocupan los primeros lugares.

Sorprendentemente, la auto-percepción de los participantes frente a estos cinco atributos es en general acertada, donde la mediana de la diferencia entre lo auto-percibido y lo que los demás perciben se centra en 0. A pesar de esto existen diferencias entre géneros donde mujeres tienen a sobre-estimar sus atributos mientras que hombres tienen a permanecer centrales, con una tendencia muy leve a sub-estimar su propio atractivo.

Respecto a los modelos ajustados, se logro predecir tanto si una persona decide unilateralmente preguntar por una segunda cita, como cuanto le gusta la otra persona como finalmente si se logra un match. Para esto los métodos basados en árboles resultados ser efectivos en rendimiento y al mismo tiempo entregar un ranking sobre que atributo impacta más en la decisión.

Ni el sexo ni la raza jugaron un rol importante en la predicción, lo que da a pensar que no existen muchas diferencias en la forma de tomar la decisión para hombres y mujeres, y que el ser de la misma etnia bajo un contexto universitario tampoco afecta.

Interesante es como cuando es preguntado la importancia de cada atributo, la inteligencia y sinceridad en el caso de las mujeres, y atractivo e inteligencia en el caso de los hombres parecen ser los factores decisivos, sin embargo en el caso de ser citas express de cuatro minutos, el humor y el atractivo terminan siendo los factores importantes, sin importar el género. Una explicación es que el corto tiempo que se tiene para hablar con la persona no permite evaluar conceptos más complejos como la inteligencia y sinceridad.

Referencias

[1] R. Fisman, S. S. Iyengar, E. Kamenica, and I. Simonson, “Gender differences in mate selection: Evidence from a speed dating experiment,” The Quarterly Journal of Economics, vol. 121, no. 2, pp. 673–697, 2006.