Máster en Estadística Aplicada para Data Science con R Software

Comparativa de los resultados obtenidos en un estudio de aceptación de tecnologías vestibles mediante un algoritmo de Gradient Boosting, con los métodos alternativos de Regresión Logística y Análisis de Correspondencia Múltiple

AUTOR: Óscar Hernández López

DIRECTOR: Juan Luis López Garrancho

FECHA: 14 / 11 /2022

ENTIDAD COLABORADORA

RESUMEN

Los dispositivos informáticos vestibles son aquellos que se insertan en el espacio personal del usuario, siendo capaces de proporcionar detección, procesamiento de datos y comunicación. El uso de esta tecnología está en continuo crecimiento en los últimos años, especialmente en el contexto de la Industria 4.0. En este contexto el autor participó en un proyecto, que sirve como base para este estudio, para identificar los factores que influyen en la aceptación y percepción de los usuarios de las tecnologías wearables en entornos industriales. El método de investigación incluyó una encuesta compuesta por 31 preguntas, respondidas por 871 empleados de empresas brasileñas y europeas (Alemania, Bélgica, España, Italia y Turquía) de diversos sectores industriales. En el mismo se aplicó un análisis de datos sobre las respuestas de la encuesta basado en técnicas de Machine Learning, concretamente a través de un algoritmo adaptativo de Gradient Boosting.

En este trabajo de final de Máster el objetivo principal es comparar los resultados obtenidos mediante el estudio original basado en Gradient Boosting con los resultados de dos técnicas alternativas como son MCA (Multiple Correspondence Analysis) y Regresión Logística. Para ello, se aplicaron las dos técnicas mencionadas a un subset de los datos originales y se obtuvo que las mismas variables que eran consideradas más significativas en la intención comportamental de aceptar la tecnología, aparecían como significativas en la regresión logística. De la misma manera se pudo observar que en el análisis MCA, la dimensión que albergaba la variable de aceptación tenía una correlación positiva entre el valor de la variable de aceptación con los valores de las variables determinados como significativas en el estudio previo.

AGRADECIMIENTOS

A mi tutor Juan Luis López, por acompañarme y guiarme durante todo el proceso. Gracias por ayudarme siempre que lo he necesitado e implicarte tanto.

A la doctora Susana Ferrero por sus claras explicaciones y el material didáctico.

A Gislene por el apoyo diario.

Tabla de contenido

2 Introducción 6
2.1 Antecedentes 7
2.2 Dispositivos portátiles (wearable devices) en el sector industrial y de seguridad.. 8
2.3 Modelos y teorías de aceptación de la tecnología 9
2.4 Evolución de las diferentes teorías 9
2.5 Teoría UTAUT2 10
2.6 Hipótesis 11
3 Material y métodos 12
3.1 Encuesta 12
3.2 Cuestionario 13
3.3 Participantes 13
3.4 Metodología de resolución con Rstudio 13
3.5 Fases del estudio 14
3.5.1 La recogida de datos 14
3.5.2 Preprocesamiento de datos 15
3.5.3. Resolución del modelo mediante las diferentes técnicas 15
3.5.4. Evaluación de los resultados 16
3.6 Metodología especifica de cada tipo de análisis 17
3.6.1 Método de resolución: Árboles de decisión, algoritmo XGBoost 17
3.6.2 Método de resolución: Análisis de correspondencia múltiple 24
3.6.3 Método de resolución: Regresión logística 26
4 Resultados y discusión 29
4.1 Resultados del estudio mediante algoritmo de Gradient Boosting 29
4.2 Resultados del estudio mediante Análisis de Correspondencia Múltiple (MCA) 32
4.2.1 Modelo sin variables suplementarias 32
4.2.2 Modelo con variables suplementarias 33
4.3 Resultados del estudio mediante Regresión Logística 36
4.3.1 Resolución de los modelos 36
4.3.2 Pruebas de bondad de ajuste 38
4.3.2 Predicciones del modelo 39
5 Conclusiones 41
6 Referencias/bibliografía 42
7 Anexos 44
7.1 Fichero de R para la resolución del modelo por Análisis de Correspondencia Múltiple (MCA) 44
7.2 Fichero de R para la resolución del modelo por Regresión Logística 48

Ilustración 1 - Modelo de UTAUT2 - Fuente: Adaptado de Venkatesh et al. (2012).....11
Ilustración 2 – Mapa de las personas encuestadas por país - Fuente: Propia.....13
Ilustración 3 – Matriz de confusión. – Fuente: Propia.....30
Ilustración 4 – Importancia de las variables en el modelo de Gradient Boosting. – Fuente: Propia.....30
Ilustración 5 – Valores SHAP en modelo de Gradient Boosting. – Fuente: Propia.....31
Ilustración 6 – Gráficas de dependencia parcial en modelo de Gradient Boosting. – Fuente: Propia 31
Ilustración 7 – Summary del modelo de MCA sin variables suplementaria. – Fuente: Propia 32
Ilustración 8 – Gráfico Biplot del modelo de MCA sin variables suplementaria. – Fuente: Propia.....33
Ilustración 9 – Summary del modelo de MCA con variables suplementaria. – Fuente: Propia 34
Ilustración 10 – Gráfico Biplot del modelo de MCA con variables suplementaria. – Fuente: Propia.....35
Ilustración 11 – Evaluación de la métrica AIC para los diferentes modelos. – Fuente: Propia.....36
Ilustración 12 – Summary del modelo fit.step de Regresión Logística. – Fuente: Propia.....36
Ilustración 13 – Summary del modelo fit.rev de Regresión Logística. – Fuente: Propia 37
Ilustración 14 – Variables más importantes del modelo de Regresión Logística. – Fuente: Propia.....38
Ilustración 15 – Matriz de confusión del modelo fit.rev. – Fuente: Propia 39
Ilustración 16 – Efectos de las variables del modelo de Regresión Logística. – Fuente: Propia.....39

2 Introducción

La motivación para elegir este trabajo basado en un estudio previo en el que había participado fue la voluntad de aplicar técnicas estadísticas alternativas a la que fue utilizada en el estudio y comparar los resultados obtenidos, con el objetivo de, en primer lugar, consolidar los conocimientos adquiridos durante el curso y en segundo lugar, confirmar los resultados obtenidos con anterioridad. Este estudio, es el estudio de referencia llevado a cabo para escribir un artículo publicado “Acceptance and perception of wearable technologies: A survey on Brazilian and European companies” SCHWAMBACH, Gislene Cássia S. et al. (2022), y por tanto, aunque se explican los resultados y los pasos realizados para obtenerlos, no se proporcionan los datos para replicarlo.

El objetivo la investigación previa era determinar los factores que influyen en la percepción y predicen la aceptación de los usuarios respecto al uso de dispositivos portátiles o también llamados vestibles (wearable devices) en entornos industriales. Según Davis, F. D. (1989) y Kalantari, M. (2017) Schall Jr, M. C., Sesek, R. F., & Cavuoto, L. A. (2018), Choi, B., Hwang, S., & Lee, S. (2017), afirman que la privacidad de los datos es un factor importante, las investigaciones académicas siguen mostrando que la no aceptación en el uso de los dispositivos se da por la confianza o falta de la misma en el manejo de los datos recolectados, ya que la seguridad de esta información genera una incomodidad por parte del usuario, especialmente para los trabajadores. La falta de información clara, de uso final, acceso y tratamiento genera una desconfianza que es percibida como perjudicial.

El uso de tecnologías wearables durante el desarrollo de la actividad laboral favorece la transferencia y el análisis de datos para el diagnóstico y el pronóstico, permitiendo al usuario, en este caso, el empleado con un puesto de trabajo con alto potencial de riesgo asociado, tener un mayor control de su entorno laboral y de sus funciones vitales en el momento de la ejecución de este. Sin embargo, (JACOBS et al., 2019) expone que la opinión del empleado, que utilizará el dispositivo al realizar sus tareas diarias debe ser un aspecto considerado, ya que es fundamental en la aplicabilidad de los dispositivos. De esta manera, el trabajo de base tiene como objetivo identificar a través de una investigación de encuesta, cual es la percepción de los responsables y empleados en las empresas brasileñas y europeas de diferentes segmentos con respecto al uso de los dispositivos vestibles en su entorno de trabajo, para el seguimiento de sus funciones durante su horario de trabajo.

2.1 Antecedentes

Los cambios en la forma de vida de los seres humanos se producen en el día a día y, según un estudio de Wen, Zhang, & Lei, (2017), los dispositivos portátiles atraen cada vez más la atención en este estilo de vida. Con un potencial emergente de uso y aplicaciones, estos accesorios han ido ganando cada vez más atención tanto en la industria como en el mundo académico (ZHANG, LUO, NIE, & ZHANG, 2017). Plausibles de diversos usos, los dispositivos inteligentes aplicados a la vigilancia de la salud son los que tienen mayor aceptación y búsqueda entre los usuarios, cada vez más preocupados por la salud. Según los autores de este estudio, en Estados Unidos en el año 2015 se comercializaron 13,2 millones de unidades de dispositivos de monitorización de la salud, promoviendo un incremento anual del 80%, totalizando 3,3 millones de unidades vendidas y unos ingresos de U$1,46 mil millones. Zhang et al. (2017) señaló existen diferentes tipos de dispositivos portátiles wearables, como la pulsera iWatch, Fitbit y Mi Band y que su crecimiento tiene el potencial de crecer un 46,6% en los siguientes cinco años, generando 1.630,3 millones de dólares para el año 2020.

Según Donati (2005), los dispositivos portátiles o vestibles (wearable devices) son una especie de ordenador interpuesto al cuerpo, de forma que no interfiere en la vida diaria del usuario, este dispositivo debe permanecer siempre encendido y accesible con total control por parte de los usuarios, ayudando en las actividades motoras y cognitivas. Los niveles de interacción de las tecnologías vestibles con el cuerpo del usuario pueden ser implantados, vestibles y aún portátiles, haciendo que la interacción con otros dispositivos que almacenan información sea controlada por el usuario (SABINE SEYMOUR, 2003). Estas tecnologías abarcan tanto accesorios como ropa que a través de funciones electrónicas y computacionales generan información, y el usuario puede tener una retroalimentación de los datos procesados (NEUMAN y SAZONOV, 2014). Según Donati (2005) el dispositivo vestible funciona como una segunda piel, siendo necesario descartar de esta clasificación los implantes, las alteraciones genéticas y los sistemas dedicados. La diferencia entre el ordenador portátil y otros dispositivos móviles, como los teléfonos móviles, es el almacenamiento de información, tanto del usuario como del entorno, lo que hace que sus actividades sean más interactivas. Según Mann (2012) los dispositivos wearables hacen que el usuario sea pasivo porque, se centra en el propio ser humano y sus características. Esto ocurre gracias a los sensores

del sistema, que pueden medir la posición del dispositivo, su desplazamiento y las constantes vitales, además de comprobar la presencia de objetos y personas alrededor y las condiciones del entorno, como la temperatura y la luz (DONATI, 2005).

Con un potencial ya desarrollado para medir señales las 24 horas del día durante los 7 días de la semana, los dispositivos portátiles o vestibles pueden monitorizar en sus usuarios y sus condiciones de salud, como el sueño, las calorías quemadas, la frecuencia cardíaca y la distancia recorrida, todos los datos puestos a disposición en tiempo real, promoviendo que a través del análisis de datos se produzca un autocontrol y que se creen estrategias para el cambio de comportamiento (SHIN et al., 2019).

Según Donati (2005) estas señales pueden ser almacenadas y transmitidas independientemente de la petición del usuario, y a partir de ello, según la programación, generar otras acciones, esta disponibilidad e integración del dispositivo vienen, así, a proponer nuevas conexiones, otra forma de sinergia entre el hombre y el ordenador, que potencialmente puede ampliar y proyectar la capacidad del usuario para interactuar y actuar en el espacio.

2.2 Dispositivos portátiles (wearable devices) en el sector industrial y de seguridad

En las últimas décadas, la revolución industrial ha transformado la perspectiva de la gente respecto al uso y los beneficios de los ordenadores y las tecnologías vestibles. Schwab, K., y Davis, N. (2019). Las innovaciones hacen que surjan procesos más complejos, pero con mayor productividad, más eficiencia y fiabilidad en sus procesos, lo que representa nuevas oportunidades y seguridad, las nuevas tecnologías actúan para prevenir el riesgo de accidentes. Según Vignali, G., Bottani, E., Guareschi, N., et al. (2019) la Industria 4.0 permite en el sector de la producción, procesos más rápidos, eficientes y de calidad, permitiendo a los gestores recoger datos de forma más rápida y eficiente con costes reducidos. El aumento de la productividad será considerablemente mayor, aumentando también la competitividad. También según el autor Vignali, G., Bottani, E., Guareschi, N., et al. (2019), y Boston Consulting, (2019), Las tecnologías cambiarán la forma o relación entre productos y clientes, máquina humana o cambiarán la visión que se tiene de la tecnología. El uso de estas herramientas aumentará la seguridad como: el sistema horizontal el Big data y la analítica, los robots autónomos, la simulación, la integración de sistemas horizontales, la computación en la nube, la

fabricación aditiva, la realidad aumentada, estos nuevos fundamentos tecnológicos permite a la industria, o a los individuos optimizar todo un proceso. Ante esto, Choi, Hwang, & Lee (2017) expone que dichos dispositivos abren nuevas oportunidades para ser aplicados en la seguridad y salud laboral. Los investigadores mencionan que estos dispositivos pueden utilizarse para localizar a los empleados en grandes áreas industriales, especialmente con riesgos asociados, y para controlar su estado fisiológico.

2.3 Modelos y teorías de aceptación de la tecnología

Como hemos comentado antes, el objetivo principal es determinar qué factores influyen en la aceptación de los dispositivos vestibles (wearable devices) utilizados para monitorizar la seguridad en los ambientes de trabajo. Para ello se parte de un conocimiento previo basado en la evolución de las diferentes teorías de aceptación existentes actualmente. Los modelos de Teorías de la aceptación de la tecnología han ido evolucionando a lo largo de los años, con la búsqueda del mejor modelo que explique la adopción de la tecnología individual. Las empresas tecnológicas se han dado cuenta de que el conocimiento de los factores que llevan a la gente a aceptar o rechazar una tecnología es vital para cualquier organización o estado que quiera implantar un nuevo desarrollo. A continuación, mencionaré brevemente las diferentes teorías, centrándome en la explicación de una de las teorías más evolucionadas en cuanto a la aceptación de la tecnología: la teoría UTAUT2.

2.4 Evolución de las diferentes teorías

Desde la década de los setenta han surgido diferentes modelos de aceptación de las tecnologías, las más conocidas, TRA (1975) Teoría de la Acción Racional, SCT (1985) Teoría Social Cognitiva, TAM (1985) Modelo de Aceptación de la Tecnología, TAM2 (2000) y TAM3 (2008), UTAUT (2003) Teoría Unificada de la Aceptación y Uso de la Tecnología y finalmente la UTAUT2 entre otras. Muchos de estas teorías se basan en unos constructos básicos como son “utilidad percibida”, “facilidad de uso percibida” y unos modificadores, “edad”, “sexo”, etc, que determinan la Intención Comportamental del individuo. En este texto nos centraremos en la teoría UTAUT2, ya que es una de las teorías más evolucionadas en describir los factores que determinan la intención comportamental y el comportamiento de uso de una tecnología.

2.5 Teoría UTAUT2

La aceptación y el uso de la tecnología de la información es un tema que ha recibido la atención de los investigadores, se han realizado varios estudios relativos a las barreras para la adopción de las nuevas tecnologías y el uso extensivo en la SI, muchas teorías se han creado en estudios anteriores para verificar los factores que influyen en la aceptación de los dispositivos tecnológicos. Alkawsi, G. A., Ali, N., Mustafa, A. S., Baashar, Y., Alhussian, H., Alkahtani, A., & Ekanayake, J. (2020). Se determinó que la expectativa de rendimiento, la condición facilitadora, la expectativa de esfuerzo, la influencia social, la motivación hedónica, la confianza, las características individuales y la intención empresarial son factores que influyen en la aceptación. Estos estudios se realizan para buscar una mejora constante, para examinar la intención y la satisfacción de los individuos a la aceptación al uso de la tecnología. Aswani, R., Ilavarasan, P. V., Kar, A. K., & Vijayan, S. (2018).Silva, P. M. D., & Dias, G. A. (2007).

Venkatesh et al. (2012), la adecuación de la UTAUT para analizar el perfil del potencial usuario en la aceptación de la tecnología requirió que se añadieran nuevas relaciones como los constructos añadidos para ampliar la teoría UTAUT. La expectativa de rendimiento se definió como la categoría en la que el uso de una tecnología proporcionará beneficios a los consumidores en la realización de algunas actividades. La expectativa de esfuerzo como percepción de la facilidad asociada al uso de la tecnología por parte de los consumidores. La influencia social como factor muy relevante ya que el consumidor siempre tiene en cuenta la opinión de amigos y familiares y el resto de la sociedad. Las condiciones facilitadoras como el reconocimiento de que el entorno y la infraestructura están ahí para apoyar el uso de la tecnología. La motivación hedónica fue definida como la sensación de bienestar o placer al usar la tecnología, es un factor importante en la determinación. El precio o relación precio-valor percibido, influye directamente en la aceptación de la tecnología, ya que destacan como intercambio, los beneficios percibidos y el coste. El Hábito se definió como la familiaridad o costumbre en el uso de la tecnología o tecnologías similares. En la ilustración 1 se pueden observar los constructos de la teoría UTAUT2, que son derivados de la original UTAUT y donde se mantuvieron las variables modificadoras individuales como la edad, el sexo y la experiencia.

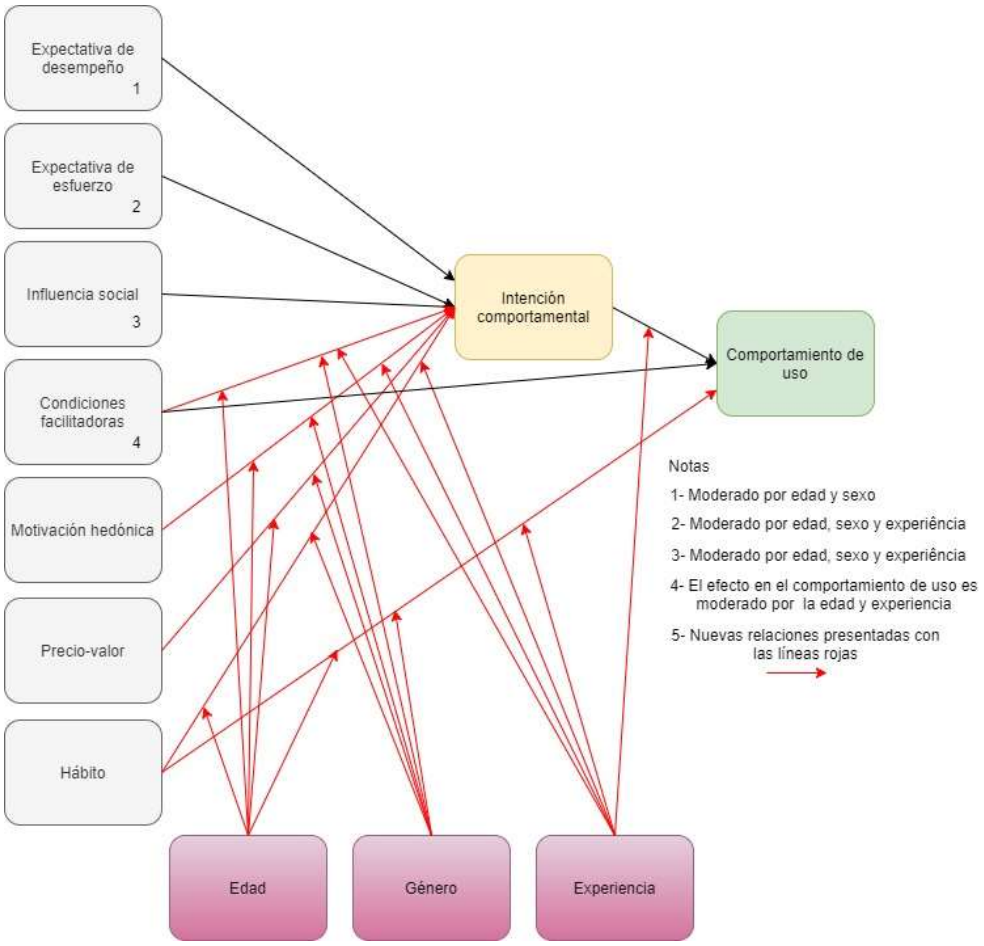

Ilustración 1 - Modelo de UTAUT2 - Fuente: Adaptado de Venkatesh et al. (2012)

2.6 Hipótesis

H01: Los modelos mentales que influyen en la aceptación de los dispositivos wearables durante la jornada laboral están determinados por una combinación de variables que representan las características del usuario, como la edad, el género, experiencia y la distribución geográfica.

H02: La expectativa de rendimiento, la expectativa de esfuerzo, la influencia social, las condiciones facilitadoras y el hábito, constructos postulados por la teoría UTAUT2, influyen en el grado de aceptación de las tecnologías. Estos constructos están representados por las variables definidas en el modelo..

3 Material y métodos

Este capítulo tiene como objetivo describir las técnicas y procedimientos utilizados para la recogida de datos del estudio original, en un intento de construir esta investigación. Los subapartados son de caracterización de la investigación, tipo de investigación, unidad de análisis y observación y técnica de recogida de datos. Se describen los tres estudios realizados sobre los datos recolectados.

La investigación exploratoria se clasifica como aquella que tiene como función principal proporcionar un mayor conocimiento del problema, con vistas a hacerlo más explícito. La investigación puede definirse a partir de los procedimientos técnicos utilizados por el investigador en función de sus objetivos (SANTOS, 2003).

El análisis se hizo con algunas visitas técnicas, a partir de octubre de 2019, en algunas empresas en otras el contacto se hizo por correos electrónico. Debido a tener la característica exploratoria de este estudio, también se realizó una encuesta de datos, entre los empleados y directivos, para un mejor análisis de los factores que influyen para el uso de las tecnologías wearables tanto de la empresa que está siendo el caso de estudio como para otras empresas a las que se les hará parte de la investigación. La investigación hizo una evaluación de los factores que influyen de la aceptación de las tecnologías vestibles con datos de empresas de Brasil y de diversos países de Europa.

3.1 Encuesta

El desarrollo de la encuesta fue personalizado para los directivos y empleados y colaboradores de las empresas participantes, y los criterios definidos fueron personas mayores de 18 años, que trabajaran en las empresas encuestadas sin limitación de género, cargo, nivel de educación y sector. La encuesta constaba de 31 preguntas y se puso a disposición el 22 de abril de 2020 en Brasil y el 26 de julio de 2020 en Alemania, Bélgica, España, Italia y Turquía a través de la plataforma Google Forms. Se remitió el acceso a los responsables de las empresas y los participantes accedieron al enlace para responder. Se insertó una explicación de la tecnología vestible en el encabezado del cuestionario para asegurar, que, a pesar de los diferentes niveles de experiencia previa con las tecnologías vestibles, todos los participantes tuvieran la misma información inicial. Las preguntas incluyen en la escala de Likert, unipolar y bipolar, para identificar las características de los participantes, la experiencia en el uso de las tecnologías, su

percepción de la tecnología y la privacidad en la gestión de los datos recogidos. En la ilustración 2 podemos visualizar los países donde se aplicaron los cuestionarios.

Ilustración 2 – Mapa de las personas encuestadas por país - Fuente: Propia

3.2 Cuestionario

Los cuestionarios aplicados a todos los participantes fueron los mismos, traducidos al idioma de cada país participante, de modo que se pudiera hacer una revisión entre los países europeos. Posteriormente, se reunieron todos los datos recogidos en Europa y en Brasil.

3.3 Participantes

Los participantes en el estudio son personas mayores de 18 años que se encuentran actualmente en el mercado laboral, concretamente son trabajadores de las empresas encuestadas, tanto de Brasil como de diversos países europeos.

3.4 Metodología de resolución con Rstudio

Para el estudio estadístico de los datos obtenidos a través de las encuestas, se utilizó

el programa RStudio 1.3.1073 con el software R en su versión 4.0.3 para 64 bits. Como hemos comentado, dividiremos este capítulo en tres apartados, el primero es el estudio del problema mediante un algoritmo de boosting, el segundo mediante un análisis de correspondencia múltiple (MCA) y por último mediante un análisis de regresión logística.

Además de las bibliotecas más comunes utilizadas, se utilizan las siguientes bibliotecas específicas para la resolución de modelos:

caret: (Classification And REgression Training) La biblioteca caret es un conjunto de funciones que intentan agilizar el proceso de creación de modelos predictivos.

xgboost: proporciona un algoritmo Parallel Tree Boosting (también conocido como GBDT, GBM) que resuelve muchos problemas de ciencia de datos de forma rápida y precisa.

pROC: Herramientas para visualizar, suavizar y comparar las características operativas del receptor (curvas ROC)

ROCR: Es una herramienta flexible para crear curvas de rendimiento 2D con parámetros de corte, combinando libremente dos de más de 25 medidas de rendimiento.

3.5 Fases del estudio

Se enunciarán a continuación 4 las cuatro etapas principales del estudio:

3.5.1 La recogida de datos

Se realizó con la encuesta en 6 países: Brasil, Bélgica, Alemania, Italia, España y Turquía, obteniendo un total de 871 respuestas. El país con más respuestas es Brasil, seguido de Italia y Bélgica. La distribución de las respuestas por región es similar entre Brasil y el resto de los países (Europa).

Tabla 1 - Lista de variables con descripción y tipo de datos - Fuente: Propia

	Nombre	Descripción	Variable dependiente	Notas
1	age	Age	No	Numerical
2	exp_qual	Evaluation of previous experience	No	Numerical
3	mgmt_preoc	Directors' attitude towards safety	No	Yes / No
4	prev_exp_WD	Previous experience Wearable Devices	No	Yes / No
5	part_select	Participated in the choice of the Wearable Devices	No	Yes / No
6	purpose_inf	Was informed of the purpose of use	No	Yes / No
7	accept_WD	Intention to accpet the Wearable Devices	Yes	Yes / No
8	tech_percep	Perception of Wearable Devices	No	Multiple choice
9	privacy_concerns	Concerns for data privacy	No	Multiple choice
10	will_contr	Willingness to contribute / participate	No	Yes / No
11	region	Region	No	Brazil / Europa

Tabla 2 - Relación de las principales variables con los constructos de la teoría UTAUT2. – Fuente: Propia

	Modelo	Constructos & Variables
Variable Dependiente	accept_WD	
1st Variable	purpose_infSim	Performance Expectancy / Facilitating Conditions / Effort Expectancy
2nd Variable	privacy_concernsMuito preocupado	Price - Value
3rd Variable	age	Age
4th Variable	tech_percepPositivo	Performance Expectancy / Effort Expectancy
5th Variable	exp_qual	Experience / Habit

3.5.2 Preprocesamiento de datos

Una de las principales tareas que hay que realizar es preparar los datos para el modelo. Para el modelo de boosting, se renombraron las variables tal y como se presentan con el nombre de la pregunta, se redujeron los nombres de los niveles en las variables, se eliminaron las respuestas incompletas o duplicadas (27 respuestas), se equilibró el conjunto de datos, se codificaron las variables categóricas (One Hot Encoding) probando diferentes procedimientos y se preparó el conjunto de datos en el formato solicitado por el modelo.

Para el estudio mediante Análisis de Correspondencia Múltiple y de Regresión Logística se utilizó un subset de datos original, al cual se codificaron las respuestas para que permitiera una mejor visualización de los resultados.

3.5.3. Resolución del modelo mediante las diferentes técnicas

1. Se hizo un modelo XGBoost, maximizando la tasa de éxito.

2. Estudio mediante Análisis de Correspondencia Múltiple.
3. Estudio mediante Regresión Logística.

3.5.4. Evaluación de los resultados

Machine learning: En el cao del algoritmo de Boosting se obtuvieron las variables más importantes y los valores Shap, lo que nos ayuda a entender cómo se relacionan las variables predictoras con la variable dependiente. Se evalúan la precisión mediante las métricas del modelo y matriz de confusión.

MCA: Se evalúan los gráficos Biplot y inercia, coordenadas, contribución, calidad.

Regresión logística: Realizamos análisis de dominancia para evaluar la importancia de las variables, pruebas de bondad de ajuste, predicciones con el modelo y evaluamos la precisión de este calculando también la matriz de confusión.

3.6 Metodología especifica de cada tipo de análisis

3.6.1 Método de resolución: Árboles de decisión, algoritmo XGBoost

A continuación, se hace una pequeña introducción de árboles de decisión y se describen los pasos para realizar el análisis.

3.6.1.1 Introducción

El análisis predictivo se define como la aplicación de algoritmos para comprender la estructura de los datos existentes y generar reglas de predicción. Estos algoritmos pueden utilizarse en un escenario no supervisado, en el que sólo se dispone de predictores (covariables) en el conjunto de datos, o en problemas supervisados, cuando, además de los predictores, se dispone de una respuesta de interés, encargada de guiar el análisis y tomar decisiones. La resolución de problemas se realiza a través de diferentes fases que se llevan a cabo en cada proceso de Machine Learning. Dos Santos, H. G., do Nascimento, C. F., Izbicki, R., de Oliveira Duarte, Y. A., & Chiavegatto Filho, A. D. P. (2019), Fafalios, S., Charonyktakis, P., & Tsamardinos, I. (2020), Friedman, J. H. (2001).

El algoritmo utilizado puede definirse como un modelo de conjunto (modelos de árboles de decisión combinados) "Boosting" propuesto por Freund y Schapire en (1996). Consisten en ajustar los modelos de árboles de decisión para que cada árbol aprenda del error del modelo anterior. Los árboles se crean de forma secuencial y no paralela (Ramdon Forest), donde cada árbol pretende reducir los errores del árbol anterior. Cada árbol que se cree en la continuación de la secuencia aprenderá de una versión actualizada de los residuos. Los árboles iniciales se denominan aprendices débiles (Weak Learners) cuyo sesgo es grande y su poder predictivo es bajo, pero la unión secuencial de estos aprendices débiles permite crear un modelo final en el que tanto el sesgo como la varianza se reducen, mejorando los árboles individuales y los modelos Bagging.

XGBoost es un algoritmo cuya versión inicial surgió en 2014 de la mano de Tianqi Chen, un estudiante de la Universidad de Washington basado en el Stochastic Gradient Boosting. Entre sus características destacan su potencia, la escalabilidad que hace que el aprendizaje sea más rápido y el uso eficiente de la memoria. Incorpora una validación

cruzada y un tratamiento de los valores perdidos con el que aprende si hay una tendencia en ellos.

El procedimiento se genera para entrenar y encontrar los parámetros óptimos que maximizan la precisión:

A. Búsqueda de la profundidad óptima entre 2 y 9 (profundidad máxima) mediante validación cruzada.

B. Buscar la tasa de aprendizaje óptima (eta) que maximice la precisión.

C. Crear el modelo con un máximo de 15.000 iteraciones, deteniendo el modelo cuando después de 50 iteraciones el error no mejora;

D. Las métricas obtenidas durante el conjunto de validación mejoran el punto de corte (cutoff).

Los modelos XGBoost utilizan múltiples parámetros, pero se han reducido a los siguientes:

❏ goal: sirve para indicar el problema al que nos enfrentamos, en este caso binario:logístico para la regresión logística (clasificación binaria).
❏ eta: es la tasa de aprendizaje de nuestro modelo o lo rápido que queremos que aprenda de los datos.
❏ max_depth: es la profundidad máxima del modelo.
❏ nrounds: número máximo de iteraciones en el modelo.
❏ colsample_bytree: es el número de variables que proporcionamos a un árbol. En nuestro caso, todas (colsample_bytree = 1).
❏ min_child_weight: es la suma mínima de instancias necesarias en un nodo secundario. En nuestro caso es 1 (una observación mínima en un nodo).

3.6.1.1.1 Importancia de las variables y de los valores de Shap

Las variables más importantes se obtuvieron mediante métodos de permutación y por valores de Shap.

El método de permutación consiste en medir la importancia de cada predictor en la predicción del modelo y cómo varía si eliminamos esa variable. Aunque es un método habitual para medir las variables más importantes, tiene dos desventajas:

❏ No nos dice cómo se relaciona la variable con la variable dependiente.

❏ No tiene en cuenta las interacciones entre las variables.

Para resolver estos dos problemas, se desarrollaron y utilizaron valores de Shap a partir de 2017 (Lundberg y Lee 2016, 2017) y nos ayudan a interpretar los llamados modelos predictivos de caja negra (XGBoost). Se basa en los valores de Shapley creados en 2012 por Lloyd Shapley utilizando la teoría de juegos.

Los valores de Shap aumentan la significación obtenida por permutación, teniendo en cuenta las interacciones entre variables, teniendo en cuenta subconjuntos de entidades.

Se obtienen valores de Shap de las diferentes combinaciones en diferentes órdenes de variables y cómo afecta esta combinación a la precisión final, siendo una ponderación de las mismas. Además, nos sirve para relacionar la variable dependiente con los predictores.

Uno de los problemas que tiene este método es que cuando tenemos muchas variables las combinaciones aumentan siendo su cálculo muy lento, por eso utilizamos valores aproximados de los valores de Shap.

3.6.1.1.2 Variable dependiente "accept_WD"

Se utiliza como variable dependiente la respuesta a la pregunta: Durante su horario de trabajo, ¿Se siente cómodo utilizando dispositivos wearables para su seguridad personal? Pregunta cerrada con dos opciones y sin valores perdidos.

3.6.1.2 Preprocesamiento de datos

Antes de entrenar un modelo, debemos realizar una serie de transformaciones en los datos para optimizar el rendimiento del algoritmo. Se pueden utilizar varios métodos, pero los pasos a seguir son los siguientes

1. eliminación de variables duplicadas
2. Eliminación de los valores perdidos.
3. Dividir el conjunto de datos.
4. Equilibrar la variable dependiente.
5. Normalización de variables numéricas.
6. Creación de variables binarias.
7. Preparación de los datos para el modelo.

3.6.1.2.1 Eliminación de variables duplicadas

Eliminación de respuestas duplicadas. Uno de los problemas que podemos encontrar son las observaciones cuyas respuestas en todas las variables son idénticas a otras. La incapacidad de eliminar las respuestas repetidas puede hacer que el modelo sea más importante para estas observaciones, reduciendo la precisión del modelo cuando se enfrenta a datos no entrenados.

En nuestro caso tenemos 27 observaciones que coinciden con otras, esto puede deberse a errores en la recogida de datos, fallos en la base de datos, personas que realizan la encuesta varias veces, casualidad etc.

3.6.1.2.2 Valores perdidos

El conjunto de datos no tiene valores perdidos que tengamos que eliminar o editar. El cuestionario se elaboró con preguntas cerradas, sin posibilidad de no responder.

3.6.1.2.3 Dividir el conjunto de datos

Dividimos el conjunto de datos en dos conjuntos a partir de la variable dependiente "accept_WD". Manteniendo la proporción de esa variable dependiente (aunque la distribución de las variables independientes no es similar a la del conjunto de datos original).

❏ Conjunto del llamado train con el 80% de las observaciones con las que entrenaremos el modelo.

❏ Conjunto llamado test con el 20% de las observaciones restantes, que servirá para comprobar la exactitud cuando se enfrente a casos que no haya visto.

El conjunto de datos de entrenamiento tiene 682 observaciones y el conjunto de validación tiene 170 observaciones.

3.6.1.2.4 Equilibrar la variable dependiente

Uno de los problemas que encontramos al analizar la variable dependiente binaria es que las dos categorías estaban desequilibradas, obteniendo más respuestas de la categoría 1 o positiva que de la categoría 0 o negativa. El problema del desequilibrio de clases es importante. Una disparidad en las frecuencias de clase observadas puede tener un impacto negativo significativo en el ajuste del modelo.
El equilibrio de clases se realizó de forma que en el conjunto de entrenamiento existieran las mismas observaciones de clase 1 y 0. Se han realizado dos procedimientos para comprobar cómo mejora el algoritmo:

❏ Muestra ascendente: muestra aleatoria con reemplazo de la clase minoritaria para que tenga el mismo tamaño que la clase mayoritaria.

❏ Muestra descendente: crear subconjuntos aleatorios de todas las clases del conjunto de entrenamiento para que sus frecuencias de clase coincidan con la clase menos prevalente.

Ninguno de los métodos mejoró los resultados del modelo, obteniendo una tasa de aciertos inferior a la obtenida con los datos no equilibrados.

3.6.1.2.5 Normalización de las variables numéricas

El conjunto de datos sólo tiene dos variables numéricas llamadas edad y exp_qual (¿cómo de buena fue su experiencia?). El modelo de árbol de decisión utilizado puede manejar diferentes rangos de las variables numéricas, por lo que no es necesario centrar y escalar las variables. Además, al no normalizar estas variables, su interpretación será más sencilla.

3.6.1.2.6 Creación de variables binarias

Creación de variables binarias. La función dummyVars() se utiliza para crear variables binarias implementadas en la biblioteca caret. Normalmente, el conjunto de datos se divide antes de crear las variables binarias. En nuestro caso, debido a un error en la

última versión de R, reproducido en varios ordenadores, tuvimos que cambiar el código y crear primero las variables binarias ficticias y luego dividir los datos en el conjunto de entrenamiento y el conjunto de validación.

3.6.1.2.7 Preparación de los datos para el modelo

Transformamos los conjuntos de datos en el formato aceptado por el modelo, donde tenemos que indicar las variables independientes por un lado y las variables dependientes por otro. Primero transformamos la variable dependiente en numérica. Como todas las variables son numéricas, utilizamos la función xgb.DMatrix() para dar el formato a los datos que soporta el modelo xgboost.

3.6.1.3 Resolución del modelo

Se resuelve el modelo.

3.6.1.3.1 Entrenamiento del modelo

El proceso de entrenamiento de los modelos XGBoost es el siguiente:

1. Buscamos la profundidad óptima con la función train() de la biblioteca caret. Se ha probado un rango de profundidad mínima entre 2 y 9, obteniendo como profundidad óptima la profundidad máxima = 7. El resto de los parámetros se dejan en sus valores por defecto.

   El modelo indica los diferentes valores de Precisión y Kappa para las distintas profundidades requeridas, dando la profundidad que maximiza la Precisión (tasa de aciertos). Una vez que tenemos la profundidad óptima, buscamos la tasa de aprendizaje.

2. Con la profundidad obtenida, se buscan los parámetros óptimos para maximizar la precisión mediante validación cruzada. El modelo calcula, como en el caso anterior, para la profundidad máxima de 7 la Precisión y Kappa para cada una de las opciones de tasa de aprendizaje, siendo la óptima eta = 0,3.

3. El modelo se entrena con estos parámetros en el conjunto de entrenamiento. Con los valores dados, creamos el modelo con la función xgboost(). El resto de los parámetros se mantienen por defecto.

El modelo con los parámetros indicados redujo la tasa de error hasta que es menor que un valor predeterminado.

4. Realizamos predicciones en el conjunto de pruebas de validación. Una vez creado el modelo, predecimos en el conjunto de pruebas de validación (datos con los que no se ha entrenado). Una característica del modelo es que cuando hacemos predicciones con él, nos da la probabilidad de pertenecer a la clase positiva, pudiendo maximizar el corte. A través de la curva ROC, elegimos el punto de corte de 0,5, es decir, las observaciones a las que el modelo asigna una probabilidad superior a 0,5 se considerarán clase 1 o positivas (SÍ) y las observaciones que el modelo asigna como 0,5 o menos se considerarán clase 0 o negativas (NO).Este punto de corte puede modificarse dependiendo de cual sea el objetivo más importante del estudio.

3.6.1.3.2 Métricas del modelo

Evaluamos la matriz de confusión y las diferentes métricas relevantes para explicar la eficiencia del modelo.

3.6.1.4 Resultados

Se evalúan los resultados de modelo comparado con las hipótesis iniciales.

3.6.1.4.1 Importancia de las variables

En este apartado se revisan que variables fueron las más importantes en el modelo y se compararán con lo comentado en la teoría.

3.6.1.4.2 Valores SHAP

Se realiza también la comparación con los valores SHAP que también da refleja una clasificación de las variables más importantes.

3.6.2 Método de resolución: Análisis de correspondencia múltiple

3.6.2.1 Introducción

El análisis de correspondencias múltiple (MCA) es una extensión del análisis de correspondencia (CA) que se basa en la descomposición de matrices en valores singulares. La metodología la desarrolló Jean-Paul Benzécri, a principios de los años 60, es decir hace más de 50 años, en la Universidad de Rennes (Francia). Nos permite analizar las correlaciones de varias variables dependientes categóricas, por lo cual es muy útil para datos provenientes, por ejemplo, de encuestas sociales. Como tal, también puede considerarse como una generalización del análisis de componentes principales cuando las variables a analizar son categóricas en lugar de cuantitativas. El MCA también puede utilizar variables cuantitativas recodificándolas como "bins" o "categorías".

3.6.2.2 Procesamiento de datos

El fichero usado para el primer análisis tuvo que ser procesado de nuevo, para configurar los valores de manera que fueran auto explicativos al ser representados en los gráficos biplot. Se utilizaron las variables con más importancia en el estudio anterior, así como la variable dependiente, "accept_WD".
Revisamos que los valores de todas las variables no numéricas sean factores y hacemos una pequeña revisión de cada variable.

3.6.2.3 Realización del análisis MCA

Realizamos el análisis MCA bajo dos supuestos:

1) Utilizando las variables cualitativas y excluyendo la variable respuesta. Para ello se utilizan las variables 3 a la 10, siendo la primera y segunda variable. "age" y "qual_exp"

2) Añadiendo las variables cuantitativas suplementarias "age", "qual_exp" y la variable cualitativa adicional "accept_WD" que es la variable respuesta.

3.6.2.4 Resultados. Interpretación

En esta sección se discute los gráficos tanto del modelo sin información suplementaria, como del modelo con las variables suplementarias añadidas. De ellos se extraen conclusiones de que variables se agrupan con comportamiento similar y se observa su distribución en las dimensiones principales.

3.6.3 Método de resolución: Regresión logística

3.6.3.1 Introducción

El modelo de regresión logística se utiliza cuando la variable dependiente es categórica. Una variable categórica es aquella cuyos valores numéricos sólo sirven como etiquetas que distinguen diferentes categorías. Cuando una variable categórica tiene sólo dos resultados mutuamente excluyentes, se utiliza el modelo de regresión logística binaria. Se originó en el siglo XIX, para el estudio del crecimiento de las poblaciones y otros estudios relacionados con reacciones químicas. En los primeros momentos se pensó que el crecimiento podía ser exponencial, pero Alphonse Quetelet (1795-1874) y su alumno Pierre-François Verhulst (1804-1849), hicieron modificaciones a la ecuación exponencial original añadiendo un término de resistencia al crecimiento. La resolución de la ecuación diferencial planteada llevó a lo que el mismo denominó la función logística. Posteriormente en los años 20 del siglo XX la función logística fue redescubierta por Pearl and Reed en un estudio de crecimiento de población. Posteriormente diferentes investigadores desarrollaron logit, Berkson (1892-1982) fue uno de los autores que más luchó para que su uso fuera extendido. En 1973 McFadden, relacionó la logit multinomial con la teoría de elección discreta de la psicología matemática, lo que sirvió de base para el modelo logit y le proporcionó el premio Nobel en 2001.

3.6.3.2 Procesamiento de datos

El fichero usado para el primer análisis tuvo que ser procesado de nuevo, para configurar los valores de manera que fueran auto explicativos al ser representados en los gráficos. Se utilizaron las variables con más importancia en el estudio anterior, así como la variable dependiente, “accept_WD”. El fichero es el mismo que el utilizado para el estudio por Análisis de Correspondencia Múltiple.

Revisamos que los valores de todas las variables no numéricas sean factores y se revisaron los valores del resto de variables.

Se dividen los datos en dos particiones una con el 80% de los datos y la otra con el 20% restante.

3.6.3.3 Resolución del modelo

Los pasos aplicados en la resolución del modelo fueron los siguientes.

3.6.3.3.1 Creación y resolución de diferentes modelos

a) Modelo con un predictor categórico (purpose_inf)
b) Modelo con un predictor continuo (age)
c) Modelo con un segundo predictor continuo (exp_qual)
d) Modelo con dos predictores categóricos y dos continuos (age+exp_qual+purpose_inf+tech_percep)
e) Modelo con dos predictores categóricos y dos continuos (age*exp_qual*purpose_inf*tech_percep)
f) Modelo con todas las variables.
g) Modelo con todas las significativas del punto f.
f) Modelo seleccionado por el método Stepwise.

3.6.3.3.2 Evaluación de los diferentes modelos

Se utiliza el parámetro AIC para ver qué modelo es más adecuado, valorando también la complejidad de este.

3.6.3.3.3 Evaluación de las variables más importantes del modelo elegido

En este caso se evaluarán las variables más importantes de dos modelos mediante la función VarImp y análisis de dominancia.

3.6.3.4 Resultados

Revisaremos las pruebas de bondad de ajuste, prueba de Hommer-Lemeshow, las predicciones del modelo y el resto de los gráficos y tablas.

3.6.3.4.1 Pruebas de bondad de ajuste

Se evalúan para los modelos fit.step, que es el modelo que se ha elegido con el procedimiento paso a paso, y fit.rev, que es el modelo revisado que usa las variables más importantes de todas, pero sin interacción.

3.6.3.4.2 Prueba de Hommer Lemeshow

La realizamos sólo al modelo fit.rev, a partir de este momento sólo revisamos los valores para este modelo. El motivo es que es el modelo más entendible de los estudiados y el objetivo del estudio es revisar si las variables más significativas en este modelos son también las que se habían encontrado en la resolución por el algoritmo de boosting.

3.6.3.4.3 Predicciones del modelo

Se realizan las predicciones y se comprueba la cantidad de valores acertados. También se estudia la matriz de confusión y el valor ROC.

3.6.3.4.4 Resultados: Gráficos, tabla y report del modelo

Finalmente se imprime el report del modelo y se pueden revisar las conclusiones del mismo, junto con la tabla del modelo y los gráficos de las diferentes variables. En las conclusiones omitiremos algunas de las representaciones gráficas que si están en los modelos de R para no alargar excesivamente el trabajo. En el archivo de R también se han realizado comentarios y se ha estructurado con apartados para acompañar el flujo de la resolución del modelo.

4 Resultados y discusión

4.1 Resultados del estudio mediante algoritmo de Gradient Boosting

En este estudio como fue hecho anteriormente y se han explicado los pasos del estudio, no se va a entrar en los pormenores de cada paso.

Una vez aplicados todos los pasos obtenemos la siguiente matriz de confusión, donde las principales métricas son:

Exactitud (Accuracy): puede definirse como la exactitud del modelo, es decir, el número de predicciones que el modelo ha realizado correctamente.

Una exactitud o acierto de 0,9. Indica que el modelo tiene un éxito del 90,0% para clasificar los eventos. A partir de esta métrica, podemos calcular la tasa de error:

El modelo se equivoca en sus predicciones en el 10,0% de los casos (tasa de error).

Precisión: la precisión indica cuántas identificaciones positivas fueron correctas, es decir, los verdaderos positivos entre el total de predicciones positivas. En este caso fue del 90,6%.

Especificidad: indica los verdaderos negativos entre los verdaderos negativos y los falsos positivos. En este caso fue del 54,84%.

La especificidad es claramente la métrica inferior del modelo. El hecho de que haya un gran desequilibrio entre las respuestas positivas y las negativas significa que el modelo predice mejor las respuestas positivas que las negativas.
Con más datos, también se podría mejorar el rendimiento del modelo a la hora de predecir valores negativos.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  17   3
         1  14 136

               Accuracy : 0.9
                 95% CI : (0.8447, 0.9407)
    No Information Rate : 0.8176
    P-Value [Acc > NIR] : 0.00219

                  Kappa : 0.611

 Mcnemar's Test P-Value : 0.01529

            Sensitivity : 0.9784
            Specificity : 0.5484
         Pos Pred Value : 0.9067
         Neg Pred Value : 0.8500
             Prevalence : 0.8176
         Detection Rate : 0.8000
   Detection Prevalence : 0.8824
      Balanced Accuracy : 0.7634

       'Positive' Class : 1
```

Ilustración 3 – Matriz de confusión. – Fuente: Propia

Los resultados mostraban que las variables más importantes eran las siguientes:

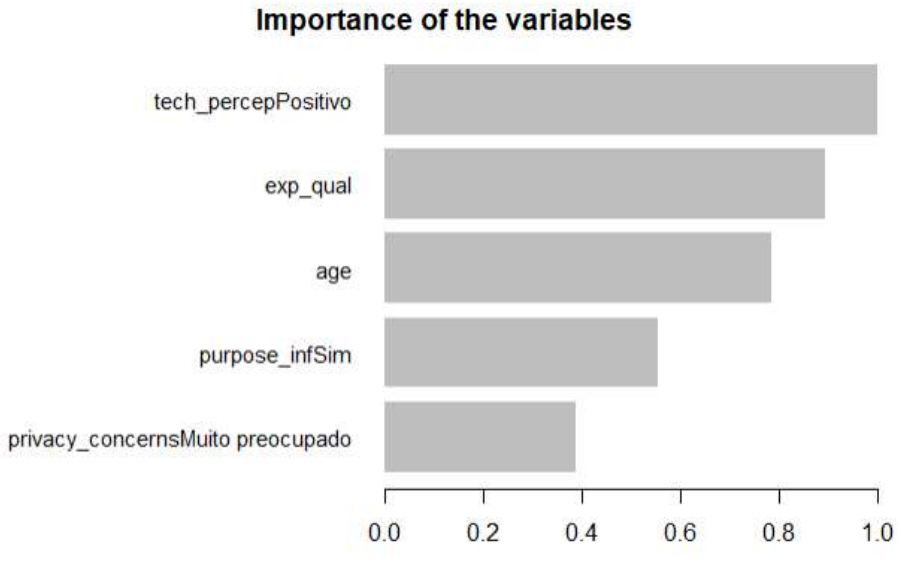

Ilustración 4 – Importancia de las variables en el modelo de Gradient Boosting. – Fuente: Propia

Esto confirma que la variable tech_percepPositivo asociada al constructo de "expectativa de desempeño" es importante para la aceptación. De la misma manera exp_qual, que está relacionado con el modificador Experiencia o con el Hábito, también

se considera importante. La variable age como modificador de los constructos principales también es una de las cinco variables más importantes.

Purpose_infSim que se relaciona con la "expectativa de esfuerzo" también es considerada importante. Realmente el hecho de que se informe a los usuarios del uso que se dará a la tecnología y a los datos colectados, influye en la actitud comportamental. Por último, privacy_concernsMuito preocupado, que se relaciona también con el constructo precio-valor y se considera importante en la intención comportamental.

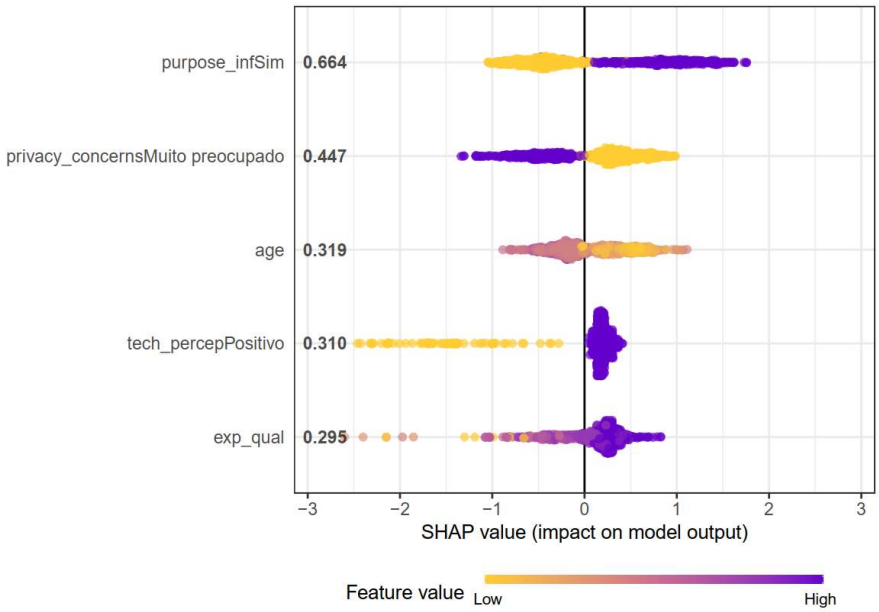

Ilustración 5 – Valores SHAP en modelo de Gradient Boosting. – Fuente: Propia

Como se puede ver los valores SHAP también indican que las variables más importantes son estas, y los gráficos de correlación parcial muestran las tendencias para cada variable.

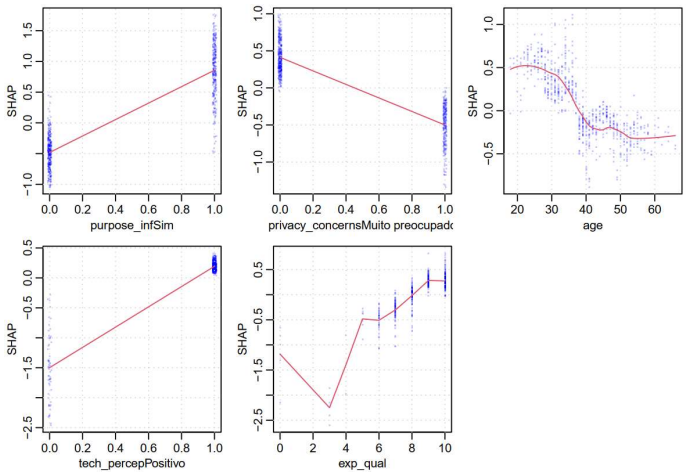

Ilustración 6 – Gráficas de dependencia parcial en modelo de Gradient Boosting. – Fuente: Propia

4.2 Resultados del estudio mediante Análisis de Correspondencia Múltiple (MCA)

4.2.1 Modelo sin variables suplementarias

Como se comentó anteriormente existe un fuerte desbalanceo entre las respuestas positivas y negativas.

Cargamos los datos y ejecutamos el modelo.
Los resultados que se obtienen de ejecutar el modelo son los siguientes.

```
Call:
MCA(X = dados[, 3:10], graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9
Variance               0.242   0.198   0.142   0.124   0.113   0.102   0.085   0.065   0.054
% of var.             21.508  17.557  12.605  10.996  10.047   9.083   7.593   5.770   4.842
Cumulative % of var.  21.508  39.065  51.670  62.666  72.713  81.795  89.389  95.158 100.000

Individuals (the 10 first)
                      Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1                   |  0.085  0.003  0.008 | -0.509  0.151  0.280 |  0.368  0.110  0.146 |
2                   |  0.462  0.101  0.154 | -0.020  0.000  0.000 |  0.516  0.216  0.192 |
3                   |  1.014  0.487  0.578 | -0.180  0.019  0.018 |  0.194  0.031  0.021 |
4                   |  0.233  0.026  0.048 | -0.146  0.012  0.019 |  0.155  0.020  0.021 |
5                   | -0.241  0.028  0.036 | -0.197  0.023  0.024 |  0.997  0.805  0.617 |
6                   |  0.191  0.017  0.059 | -0.527  0.161  0.448 | -0.302  0.074  0.147 |
7                   | -0.460  0.100  0.202 | -0.094  0.005  0.008 |  0.447  0.162  0.191 |
8                   |  0.141  0.009  0.013 |  0.036  0.001  0.001 |  0.807  0.528  0.419 |
9                   | -0.471  0.105  0.163 | -0.323  0.061  0.077 |  0.636  0.328  0.298 |
10                  |  0.898  0.383  0.332 |  0.377  0.083  0.059 |  0.356  0.102  0.052 |

Categories (the 10 first)
                        Dim.1     ctr    cos2  v.test     Dim.2     ctr    cos2  v.test     Dim.3     ctr    cos2  v.test
purpose_infSim        |  -0.728  12.274  0.430 -19.347 |  -0.213   1.282  0.037  -5.650 |   0.514  10.443  0.215  13.662 |
purpose_infNao        |   0.591   9.952  0.430  19.347 |   0.172   1.040  0.037   5.650 |  -0.417   8.468  0.215 -13.662 |
tech_percepPositivo   |  -0.147   1.010  0.216 -13.694 |  -0.001   0.000  0.000  -0.122 |  -0.013   0.013  0.002  -1.197 |
tech_percepNegativo   |   1.470  10.125  0.216  13.694 |   0.013   0.001  0.000   0.122 |   0.129   0.132  0.002   1.197 |
Nao estou preocupado  |  -0.399   3.780  0.136 -10.862 |   0.726  15.362  0.450  19.783 |  -0.357   5.160  0.108  -9.715 |
Muito preocupado      |   0.294   1.743  0.055   6.940 |  -0.994  24.397  0.632 -23.455 |  -0.077   0.203  0.004  -1.814 |
Irrelevante           |   0.461   1.637  0.037   5.693 |   0.359   1.219  0.023   4.439 |   1.301  22.265  0.297  16.072 |
will_contrNao         |   0.731   6.528  0.166  12.000 |   0.788   9.285  0.192  12.930 |   0.788  12.956  0.193  12.941 |
will_contrSim         |  -0.226   2.022  0.166 -12.000 |  -0.244   2.876  0.192 -12.930 |  -0.244   4.013  0.193 -12.941 |
Brasil                |   0.248   1.573  0.061   7.259 |  -0.804  20.361  0.640 -23.593 |   0.225   2.217  0.050   6.597 |

Categorical variables (eta2)
                      Dim.1 Dim.2 Dim.3
purpose_inf         | 0.430 0.037 0.215 |
tech_percep         | 0.216 0.000 0.002 |
privacy_concerns    | 0.139 0.648 0.313 |
will_contr          | 0.166 0.192 0.193 |
region              | 0.061 0.640 0.050 |
prev_exp_WD         | 0.168 0.042 0.244 |
mgmt_preoc          | 0.281 0.009 0.058 |
part_select         | 0.476 0.014 0.061 |
```

Ilustración 7 – Summary del modelo de MCA sin variables suplementaria. – Fuente: Propia

Está claro que muchas de las variables son importantes ya que el valor absoluto de V.test de todas ellas es >2.
El aporte de cada variable a cada dimensión se puede ver claramente en el valor de eta2.

Como se puede apreciar las variables que mas influyen en la dimensión 1 son Purpose_inf, tech_percep, privacy_concerns, will_contrib.

En el gráfico biplot, ilustración 8, podemos ver como los valores positivos de las variables se distribuyen en el lado izquierdo del gráfico.

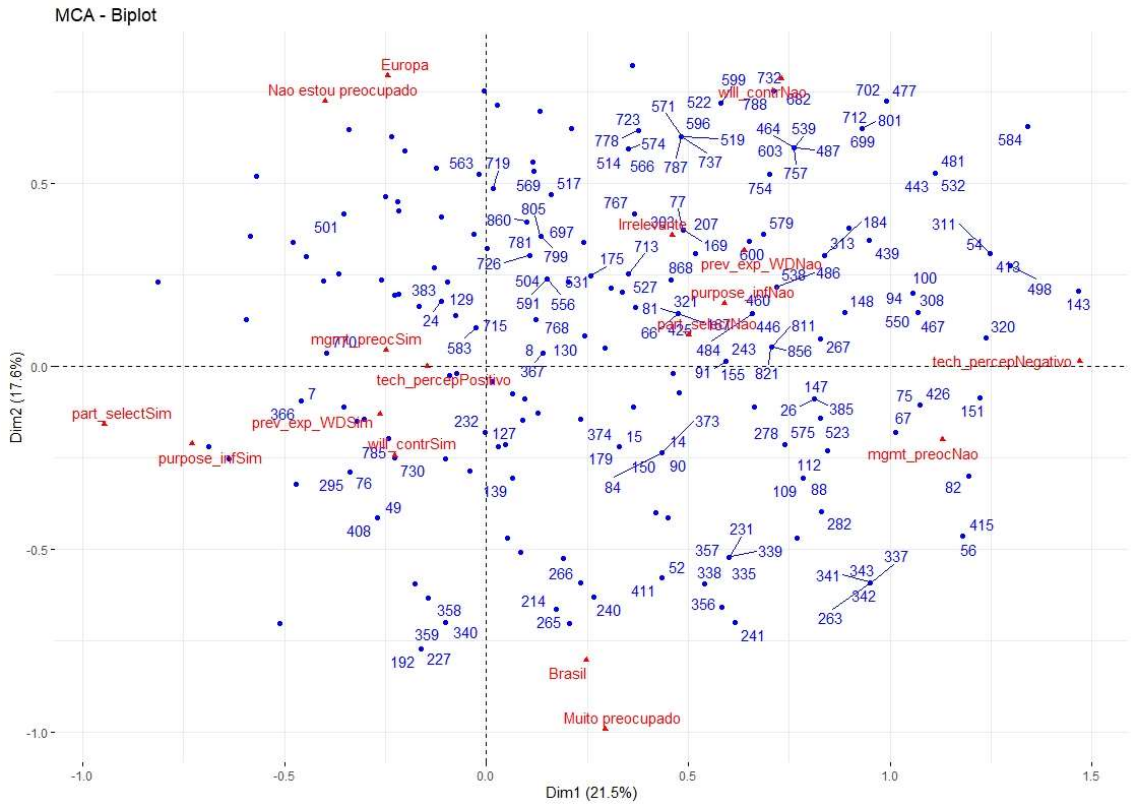

Ilustración 8 – Gráfico Biplot del modelo de MCA sin variables suplementaria. – Fuente: Propia

Todas los valores de las variables que influyen en la aceptación de manera positiva se distribuyen en la parte izquierda del gráfico. La dimensión 1 tiene una clara correlación con la aceptación, siendo los valores que influyen negativamente en la aceptación los que quedan a la derecha del gráfico.

En la dimensión 2 se puede ver que se ha representado la distribución de la región quedando abajo a la derecha del eje las respuestas de Brasil y los valores de la variable privacy_concernsMuito preocupado. Se deduce claramente en los datos que la mayoría de las personas que respondieron como muy preocupadas provenían de Brasil y tiene una leve influencia negativa.

4.2.2 Modelo con variables suplementarias

En este punto añadimos las variables cuantitativas age y exp_qual y la variable respuesta accept_WD. Con ello situamos la variable respuesta en el modelo para ver cual es la dimensión que la representa y como se podía esperar, en la dimensión 1 es donde tiene casi la totalidad de su importancia representada. Con el valor negativo

situado a la derecha del gráfico y el valor positivo a la izquierda. Tal como se anticipaba por la distribución de las otras variables cualitativas.

```
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9
Variance               0.242   0.198   0.142   0.124   0.113   0.102   0.085   0.065   0.054
% of var.             21.508  17.557  12.605  10.996  10.047   9.083   7.593   5.770   4.842
Cumulative % of var.  21.508  39.065  51.670  62.666  72.713  81.795  89.389  95.158 100.000

Individuals (the 10 first)
                      Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1                   |  0.085  0.003  0.008 | -0.509  0.151  0.280 |  0.368  0.110  0.146 |
2                   |  0.462  0.101  0.154 | -0.020  0.000  0.000 |  0.516  0.216  0.192 |
3                   |  1.014  0.487  0.578 | -0.180  0.019  0.018 |  0.194  0.031  0.021 |
4                   |  0.233  0.026  0.048 | -0.146  0.012  0.019 |  0.155  0.020  0.021 |
5                   | -0.241  0.028  0.036 | -0.197  0.023  0.024 |  0.997  0.805  0.617 |
6                   |  0.191  0.017  0.059 | -0.527  0.161  0.448 | -0.302  0.074  0.147 |
7                   | -0.460  0.100  0.202 | -0.094  0.005  0.008 |  0.447  0.162  0.191 |
8                   |  0.141  0.009  0.013 |  0.036  0.001  0.001 |  0.807  0.528  0.419 |
9                   | -0.471  0.105  0.163 | -0.323  0.061  0.077 |  0.636  0.328  0.298 |
10                  |  0.898  0.383  0.332 |  0.377  0.083  0.059 |  0.356  0.102  0.052 |

Categories (the 10 first)
                        Dim.1     ctr    cos2  v.test     Dim.2     ctr    cos2  v.test     Dim.3     ctr    cos2  v.test
purpose_infSim        | -0.728  12.274  0.430 -19.347 |  -0.213   1.282  0.037  -5.650 |   0.514  10.443  0.215  13.662 |
purpose_infNao        |  0.591   9.952  0.430  19.347 |   0.172   1.040  0.037   5.650 |  -0.417   8.468  0.215 -13.662 |
tech_percepPositivo   | -0.147   1.010  0.216 -13.694 |  -0.001   0.000  0.000  -0.122 |  -0.013   0.013  0.002  -1.197 |
tech_percepNegativo   |  1.470  10.125  0.216  13.694 |   0.013   0.001  0.000   0.122 |   0.129   0.132  0.002   1.197 |
Nao estou preocupado  | -0.399   3.780  0.136 -10.862 |   0.726  15.362  0.450  19.783 |  -0.357   5.160  0.108  -9.715 |
Muito preocupado      |  0.294   1.743  0.055   6.940 |  -0.994  24.397  0.632 -23.455 |  -0.077   0.203  0.004  -1.814 |
Irrelevante           |  0.461   1.637  0.037   5.693 |   0.359   1.219  0.023   4.439 |   1.301  22.265  0.297  16.072 |
will_contrNao         |  0.731   6.528  0.166  12.000 |   0.788   9.285  0.192  12.930 |   0.788  12.956  0.193  12.941 |
will_contrSim         | -0.226   2.022  0.166 -12.000 |  -0.244   2.876  0.192 -12.930 |  -0.244   4.013  0.193 -12.941 |
Brasil                |  0.248   1.573  0.061   7.259 |  -0.804  20.361  0.640 -23.593 |   0.225   2.217  0.050   6.597 |

Categorical variables (eta2)
                      Dim.1 Dim.2 Dim.3
purpose_inf         | 0.430 0.037 0.215 |
tech_percep         | 0.216 0.000 0.002 |
privacy_concerns    | 0.139 0.648 0.313 |
will_contr          | 0.166 0.192 0.193 |
region              | 0.061 0.640 0.050 |
prev_exp_WD         | 0.168 0.042 0.244 |
mgmt_preoc          | 0.281 0.009 0.058 |
part_select         | 0.476 0.014 0.061 |

Supplementary categories
                        Dim.1    cos2  v.test     Dim.2    cos2  v.test     Dim.3    cos2  v.test
acceptWDNao           |  0.774  0.132  10.703 |   0.136  0.004   1.876 |   0.013  0.000   0.184 |
acceptWDSim           | -0.170  0.132 -10.703 |  -0.030  0.004  -1.876 |  -0.003  0.000  -0.184 |

Supplementary categorical variables (eta2)
                      Dim.1 Dim.2 Dim.3
accept_WD           | 0.132 0.004 0.000 |

Supplementary continuous variables
                       Dim.1    Dim.2    Dim.3
age                  | -0.157 |  0.192 |  0.030 |
exp_qual             | -0.145 | -0.220 | -0.100 |
```

Ilustración 9 – Summary del modelo de MCA con variables suplementaria. – Fuente: Propia

Las variables cuantitativas age y exp_qual se ven representadas casi por igual en la dimensión 1 y la 2.

Como conclusiones podríamos decir observando la ilustraión 9, que se confirma la correlación entre la aceptación de la tecnología y los valores positivos de las variables, tech_percep, purpose_inf, part_selec, will_contrib, mgmt_preoc, prev_exp_WD. Siendo las más importantes purpose_inf, tech_percep, will_contrib y los valores de la variable privacy_concerns.

Se puede observar por tanto en estos resultados la misma tendencia que se obtuvo en el estudio original de referencia y que se quería comprobar. El método MCA aunque no permite hacer predicciones, indica gráficamente y numéricamente estas correlaciones y permite obtener conclusiones muy válidas. En la ilustración10 presentada a

continuación se aprecia la situación de la variable dependiente, en color verde, respecto al resto de las variables y los ejes.

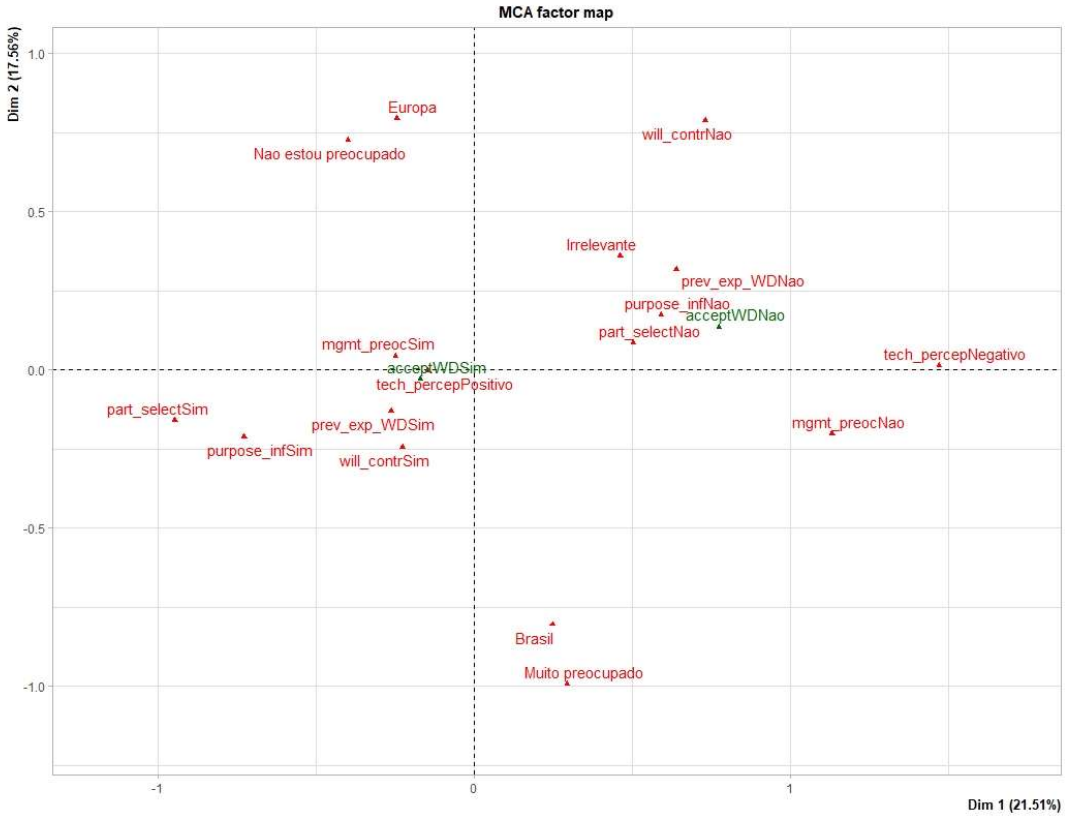

Ilustración 10 – Gráfico Biplot del modelo de MCA con variables suplementaria. – Fuente: Propia

La representación gráfica biplot con las variables suplementarias, permite analizar las tendencias de agrupación respecto a la variable respuesta.

4.3 Resultados del estudio mediante Regresión Logística

Mediante el método de Regresión Logística comprobaremos que modelo y que variables son las más importantes en el modelo elegido. Una vez se cargaron los datos y se visualizaron, se ejecutaron diferentes modelos como se describió en el apartado de Material y Métodos.

4.3.1 Resolución de los modelos

La evaluación de los diferentes modelos mediante la métrica AIC nos proporciona los siguientes resultados:

	df	AIC
fit.12345678910	12	532.8635
fit.1234	5	555.2417
fit.1234int	16	561.7787
fit.age	2	660.3317
fit.rev	8	530.8637
fit.purp_inf	2	606.1808
fit.qual_exp	2	643.5887
fit.step	21	514.4971

Ilustración 11 – Evaluación de la métrica AIC para los diferentes modelos. – Fuente: Propia

El modelo fit.step es el que tiene la métrica más óptima (mínima) , pero a cambio la complejidad del mismo es más elevada.

```
Call:
glm(formula = accept_WD ~ age + exp_qual + purpose_inf + tech_percep +
    privacy_concerns + will_contr + region + prev_exp_WD + mgmt_preoc +
    part_select + age:tech_percep + age:mgmt_preoc + exp_qual:purpose_inf +
    exp_qual:prev_exp_WD + exp_qual:part_select + purpose_inf:tech_percep +
    purpose_inf:prev_exp_WD + tech_percep:privacy_concerns +
    will_contr:region + will_contr:mgmt_preoc + region:prev_exp_WD,
    family = binomial, data = traindados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9897   0.1496   0.3083   0.5464   2.5373

Coefficients:
                                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                                              -2.39652    1.91118  -1.254 0.209860
age                                                       0.03101    0.02627   1.180 0.237856
exp_qual                                                  0.19582    0.18304   1.070 0.284697
purpose_infpurpose_infNao                                 2.19514    1.53290   1.432 0.152140
tech_perceptech_percepNegativo                           14.08876  768.64560   0.018 0.985376
privacy_concernsMuito preocupado                         -0.85176    0.32739  -2.602 0.009277 **
privacy_concernsIrrelevante                               0.97414    0.50724   1.920 0.054797 .
will_contrwill_contrSim                                   2.02981    0.55983   3.626 0.000288 ***
regionEuropa                                              0.18758    0.54307   0.345 0.729785
prev_exp_WDprev_exp_WDSim                                -3.73301    1.44969  -2.575 0.010023 *
mgmt_preocmgmt_preocSim                                   3.61191    1.22773   2.942 0.003262 **
part_selectpart_selectSim                                 2.32112    1.53298   1.514 0.129995
age:tech_perceptech_percepNegativo                       -0.05157    0.03679  -1.402 0.161041
age:mgmt_preocmgmt_preocSim                              -0.06203    0.02913  -2.129 0.033233 *
exp_qual:purpose_infpurpose_infNao                       -0.32023    0.17691  -1.810 0.070278 .
exp_qual:prev_exp_WDprev_exp_WDSim                        0.64932    0.15600   4.162 3.15e-05 ***
exp_qual:part_selectpart_selectSim                       -0.30209    0.17768  -1.700 0.089091 .
purpose_infpurpose_infNao:tech_perceptech_percepNegativo -15.49873  768.64363  -0.020 0.983913
purpose_infpurpose_infNao:prev_exp_WDprev_exp_WDSim      -1.25186    0.59017  -2.121 0.033905 *
tech_perceptech_percepNegativo:privacy_concernsMuito preocupado  2.04687  1.20554  1.698 0.089530 .
tech_perceptech_percepNegativo:privacy_concernsIrrelevante  0.38840  1.36676  0.284 0.776273
will_contrwill_contrSim:regionEuropa                      0.79026    0.55357   1.428 0.153413
will_contrwill_contrSim:mgmt_preocmgmt_preocSim          -1.63169    0.60187  -2.711 0.006708 **
regionEuropa:prev_exp_WDprev_exp_WDSim                   -1.00473    0.54886  -1.831 0.067162 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 659.16  on 697  degrees of freedom
Residual deviance: 461.99  on 674  degrees of freedom
AIC: 509.99

Number of Fisher Scoring iterations: 15
```

Ilustración 12 – Summary del modelo fit.step de Regresión Logística. – Fuente: Propia

Realmente tienen sentido algunas interacciones apuntadas como significativas, pero en general no es un modelo claro. Como la intención del estudio era valorar si las variables que explican la intención comportamental eran las mismas que en el estudio inicial, el modelo fit. rev será el que analizará en detalle.

```
Call:
glm(formula = accept_WD ~ age + exp_qual + purpose_inf + tech_percep +
    privacy_concerns + will_contr, family = binomial, data = traindados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7643   0.2309   0.3664   0.5771   1.8977

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      1.53037    0.82684   1.851  0.06419 .
age                             -0.02254    0.01141  -1.976  0.04816 *
exp_qual                         0.18188    0.06442   2.823  0.00475 **
purpose_infpurpose_infNao       -1.44944    0.27207  -5.328 9.96e-08 ***
tech_perceptech_percepNegativo  -1.59398    0.31624  -5.040 4.64e-07 ***
privacy_concernsMuito preocupado -0.70557    0.25578  -2.759  0.00581 **
privacy_concernsIrrelevante      0.18847    0.36780   0.512  0.60835
will_contrwill_contrSim          1.20896    0.24194   4.997 5.82e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 659.16  on 697  degrees of freedom
Residual deviance: 514.86  on 690  degrees of freedom
AIC: 530.86

Number of Fisher Scoring iterations: 5
```

Ilustración 13 – Summary del modelo fit.rev de Regresión Logística. – Fuente: Propia

Apreciamos que las variables que resultan significativas son: Tech_percepNegativo, purpose_infNao, will_contribSim, privacy_concernsMuito Preocupado, exp_qual y en último lugar la variable age.

Teniendo en cuenta que en este caso la clase de referencia es accept_WDNao, como se puede ver, los signos de los coeficientes son los correctos y coinciden con lo visto en el análisis MCA, la edad rebaja la aceptación a medida que se incrementa y tiene un coeficiente negativo. Exp_qual tiene coeficiente positivo ya que valores mayores tienden a ser indicativo de alta probabilidad de aceptación. Purpose_infNao tiene un coeficiente negativo, de la misma manera que el valor positivo tendría un coeficiente positivo. En el caso de que no se haya informado de la finalidad del dispositivo y de los datos colectados, la aceptación tiende a ser menor. Es lo mismo que ocurre con tech_percepNegativo rebaja la probabilidad de aceptar ya que tiene un coeficiente negativo. Privacy_concernsMuito preocupado, tiene coeficiente negativo, ya que como vimos las personas que están muy preocupadas por la privacidad de los datos tendían a ser más reacias a aceptar la tecnología. En último lugar, will_contribSim, tiene un

coeficiente positivo, lo que indicaría que las personas que declaran que contribuirían a la definición de que dispositivo usar y como implementarlo, son mas proclives a aceptar la tecnología.

Al comparar los resultados obtenidos mediante el algoritmo de Gradient Boosting y la Regresión Logística, vemos que las variables importantes son casi las mismas, aunque el orden de importancia es un poco diferente en los dos casos, ver ilustraciones 4 y 14. Además, en el estudio de regresión logística aparece la variable will_contrib, que no era significativa, o al menos no estaba en el top 5 del estudio de Gradient Boosting.

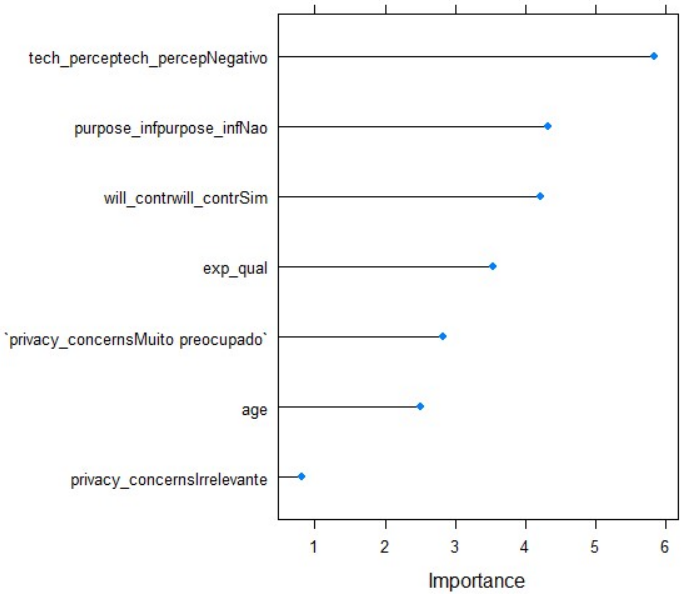

Ilustración 14 – Variables más importantes del modelo de Regresión Logística. – Fuente: Propia

Otro tema a tener en cuenta por los diferentes valores de las variables es que en el estudio original la variable de referencia fue accept_WDSim , en la regresión logística fue accept_WDNao.

4.3.2 Pruebas de bondad de ajuste

Aunque en el modelo fit.step conseguíamos un Tjur's R2 de 0.299 en el modelo fit.rev se obtiene un índice de 0.234, lo cual puede parecer no muy elevado pero para estudios sociológicos puede ser significativo.

La prueba global de ajuste de Hosmer-Lemeshow dio un valor menor que 0.05 por lo que implica que el modelo no representaría bien los datos.

4.3.2 Predicciones del modelo

Una realizadas las predicciones sobre el conjunto el 77,65% de los valores predichos fueron correctos y la matriz de confusión fue la siguiente:

```
Confusion Matrix and Statistics

          Reference
Prediction    acceptWDNao acceptWDSim
  acceptWDNao          10           9
  acceptWDSim          21         133

               Accuracy : 0.8266
                 95% CI : (0.7618, 0.8798)
    No Information Rate : 0.8208
    P-Value [Acc > NIR] : 0.46894

                  Kappa : 0.3054

 Mcnemar's Test P-Value : 0.04461

            Sensitivity : 0.3226
            Specificity : 0.9366
         Pos Pred Value : 0.5263
         Neg Pred Value : 0.8636
             Prevalence : 0.1792
         Detection Rate : 0.0578
   Detection Prevalence : 0.1098
      Balanced Accuracy : 0.6296

       'Positive' Class : acceptWDNao
```

Ilustración 15 – Matriz de confusión del modelo fit.rev. – Fuente: Propia

Como se puede ver en la ilustración 15 los valores de sensitividad y especificidad están invertidos porque la clase positiva se ha tomado como la opuesta al modelo de Gradient Boosting. Los valores son similares, aunque la reducción en número de variables y el grupo de datos desbalanceado hace que las predicciones de los datos negativos sean peores que en el modelo de Gradient Boosting. El valor ROC es de 80.82%, superior al 80%.

Por último, vemos los efectos de las diferentes variables en la ilustración 16 comparados con los gráficos de dependencia parcial que se veían en el modelo de Gradient Boosting, ilustración 6.

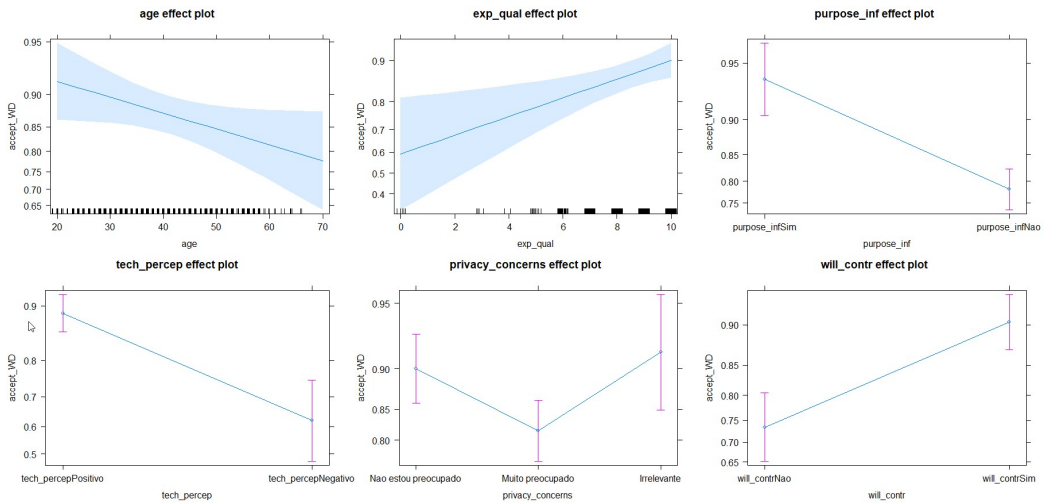

Ilustración 16 – Efectos de las variables del modelo de Regresión Logística. – Fuente: Propia

Hay que tener en cuenta que en el caso del modelo de Gradient Boosting son valores Shap y que algunas de las variables son el valor opuesto a la variable presentada en el primer estudio, por tanto, será una imagen invertida de la obtenida en el modelo de Regresión Logística.

5 Conclusiones

El objetivo del estudio era comparar si los resultados obtenidos al estudiar unos datos mediante un modelo de Machine Learning, concretamente con un algoritmo de Gradient Boosting podían ser comparables a los resultados obtenidos por técnicas estadísticas clásicas.

Creo que se han cumplido los objetivos de este estudio que era comprobar si las variables designadas como más significativas en la predicción de la intención comportamental por el algoritmo de Gradient Boosting, son las mismas que las que se han obtenido mediante la resolución con un modelo de Regresión Logística, apoyado con un Análisis de Correspondencia Múltiple.

Se ha podido determinar que existe correlación significativa entre ciertas variables recogidas en el estudio, que representan los constructos de la teoría UTAUT2 y la intención comportamental de aceptar la tecnología. Aunque en mi opinión sería necesario un estudio más extenso para consolidar los resultados apuntados.

Personalmente estoy muy satisfecho de haber elegido este estudio, porque me ha permitido consolidar las técnicas anteriormente descritas y revisar desde otro punto de vista un estudio previamente realizado.

6 Referencias/bibliografía

ALKAWSI, Gamal Abdulnaser et al. A hybrid SEM-neural network method for identifying acceptance factors of the smart meters in Malaysia: Challenges perspective. **Alexandria** Engineering Journal, 2020.

ASWANI, Reema et al. Adoption of public WiFi using UTAUT2: An exploration in an emerging economy. **Procedia computer science**, v. 132, p. 297-306, 2018.

Boston Consulting Group, "**Embracing Industry 4.0 and Rediscovering Growth**", 2019. [Online]. Available: https://www.bcg.com/itit/capabilities/operations/embracing-industry-4.0- rediscovering-growth.aspx.

CHOI, Byungjoo; HWANG, Sungjoo; LEE, SangHyun. What drives construction workers' acceptance of wearable technologies in the workplace?: Indoor localization and wearable health devices for occupational safety and health. **Automation in Construction**, v. 84, p. 31-41, 2017

DAVIS, Fred D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. **MIS quarterly**, p. 319-340, 1989.

DOS SANTOS, Hellen Geremias et al. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. **Cad. Saúde Pública,** v. 35, n. 7, p. e00050818, 2019.

DONATI, Luisa Angelica Paraguai et al. **O computador como veste-interface:(re) configurando os espaços de atuação**. 2005.

FAFALIOS, Stefanos; CHARONYKTAKIS, Pavlos; TSAMARDINOS, Ioannis. **Gradient Boosting Trees**. 2020.

FRIEDMAN, Jerome H. 1999 Reitz Lecture**. The Annals of Statistics**, v. 29, n. 5, p. 1189-1232, 2001.

JACOBS, Jesse V. et al. Employee acceptance of wearable technology in the workplace. **Applied ergonomics**, v. 78, p. 148-156, 2019.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. 2017. p. 4765-4774.

MANN, S. Computação Wearable. **Encyclopedia of Interação Humano-Computador. Aarhus, Dinamarca**: A Fundação Interaction-Design. org, 2012.

KALANTARI, Mahdokht. Consumers' adoption of wearable technologies: literature review, synthesis, and future research agenda. **International Journal of Technology Marketing**, v. 12, n.3, p. 274-307, 2017.

SAZONOV, Edward (Ed.). **Wearable Sensors: Fundamentals, implementation and applications.** Elsevier, 2014.

SILVA, Patrícia Maria da; DIAS, Guilherme Ataíde. **Teorias sobre Aceitação de Tecnologia: por que os usuários aceitam ou rejeitam as tecnologias de informação**?. 2007

Schwambach, G. C. S., et al. (2022). **Acceptance and perception of wearable technologies:** Asurvey on Brazilian and European companies. Tec. Soc., 68: 101840.

SHIN, G., JARRAHI, M. H., FEI, Y., KARAMI, A., GAFINOWITZ, N., BYUN, A., & Lu, X. (2019). Wearable activity trackers, accuracy, adoption, acceptance and health impact: A systematic literature review. **Journal of Biomedical Informatics**, 93(October 2018),

SCHALL JR, Mark C.; SESEK, Richard F.; CAVUOTO, Lora A. Barriers to the adoption of wearable sensors in the workplace: A survey of occupational safety and health professionals. **Human factors**, v. 60, n. 3, p. 351-362, 2018.

SEYMOUR, Sabine. Fashionable technology: **The intersection of design, fashion, science, and technology**. New York: Springer, 2008.

SCHWAB, Klaus; DAVIS, Nicholas. **Aplicando a quarta revolução industrial**. Edipro, 2019.

VIGNALI, Giuseppe et al. Development of a 4.0 industry application for increasing occupational safety**:** guidelines for a correct approach. In: **2019 IEEE International Conference on Engineering,** Technology and Innovation (ICE/ITMC). IEEE, 2019. p. 1-6.

VENKATESH, Viswanath; THONG, James YL; XU, Xin. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. **MIS quarterly**, p. 157-178, 2012.

WEN, Dong; ZHANG, Xingting; LEI, Jianbo. Consumers' perceived attitudes to wearable devices in health monitoring in China: A survey study. **Computer methods and programs in biomedicine**, v. 140, p. 131-137, 2017.

Zhang, M., Luo, M., Nie, R., & Zhang, Y. (2017). Technical attributes, health attribute, consumer attributes and their roles in adoption intention of healthcare wearable technology. **International Journal of Medical Informatics**, 108(September), 97–109.

7 Anexos

7.1 Fichero de R para la resolución del modelo por Análisis de Correspondencia Múltiple (MCA)

```
# 1 Preparamos los datos
## ----------------------------------------
## ----------------------------------------

# 1.1 Cargamos los datos
## ----------------------------------------

library(FactoMineR)
library(gplots)
library(factoextra)
library(readr)
library(CAinterprTools)
library(ca)
library(ggpubr)

dados <- read_delim("C:/Users/u971259/Desktop/Oscar/Seguridad/GS/Proyecto/Resultado da pesquisa/Estudio estadistico/20201121/DadosEnviados20201026_Master3.csv", delim = ";", escape_double = FALSE, col_types = cols(purpose_inf = col_factor(levels = c("purpose_infSim","purpose_infNao")), tech_percep = col_factor(levels = c("tech_percepPositivo", "tech_percepNegativo")), privacy_concerns = col_factor(levels = c("Nao estou preocupado", "Muito preocupado", "Irrelevante")), will_contr = col_character()), trim_ws = TRUE)

dados$age <- as.integer(dados$age)
dados$exp_qual <- as.integer(dados$exp_qual)
dados$purpose_inf <- as.factor(dados$purpose_inf)
dados$tech_percep <- as.factor(dados$tech_percep)
dados$privacy_concerns<-as.factor(dados$privacy_concerns)
dados$will_contr <- as.factor(dados$will_contr)
dados$region <- as.factor(dados$region)
dados$prev_exp_WD <- as.factor(dados$prev_exp_WD)
dados$mgmt_preoc <- as.factor(dados$mgmt_preoc)
dados$part_select <- as.factor(dados$part_select)
dados$accept_WD <- as.factor(dados$accept_WD)

# dados$age<- cut(dados$age, breaks = c(0, 9, 19, 29, 39, 49,59,69, Inf),
#     labels = c("0-9", "10-19", "20-29", "30-39","40-49","50-59","60-69","70-100"),
#     include.lowest = TRUE)
#
# dados$exp_qual<- cut(dados$exp_qual, breaks = c(0, 5, 7, Inf),
#                labels = c("Muy Baja", "Media", "Alta"),
#                include.lowest = TRUE)

str(dados)

summary(dados)


# 1.2 describimos los valores de aceptación
## ----------------------------------------
plot(dados[,11], main=colnames(dados)[11],
    ylab = "Count", col="steelblue", las = 2)

# 2 Realizamos el analisis de correspondencia.
## ----------------------------------------
## ----------------------------------------
dados.mca<-MCA(dados[, 3:10], graph=FALSE)

summary(dados.mca)

# 3 Interpretamos los resultados.
## ----------------------------------------
## ----------------------------------------

# Eigenvalues
#                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9
# Variance               0.242   0.198   0.142   0.124   0.113   0.102   0.085   0.065   0.054
# % of var.             21.508  17.557  12.605  10.996  10.047   9.083   7.593   5.770   4.842
# Cumulative % of var.  21.508  39.065  51.670  62.666  72.713  81.795  89.389  95.158 100.000
#
# Individuals (the 10 first)
#                       Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
#   1                |  0.085  0.003  0.008 | -0.509  0.151  0.280 |  0.368  0.110  0.146 |
#   2                |  0.462  0.101  0.154 | -0.020  0.000  0.000 |  0.516  0.216  0.192 |
#   3                |  1.014  0.487  0.578 | -0.180  0.019  0.018 |  0.194  0.031  0.021 |
#   4                |  0.233  0.026  0.048 | -0.146  0.012  0.019 |  0.155  0.020  0.021 |
#   5                | -0.241  0.028  0.036 | -0.197  0.023  0.024 |  0.997  0.805  0.617 |
#   6                |  0.191  0.017  0.059 | -0.527  0.161  0.448 | -0.302  0.074  0.147 |
#   7                | -0.460  0.100  0.202 | -0.094  0.005  0.008 |  0.447  0.162  0.191 |
```

```
#   8                 |  0.141  0.009  0.013 |  0.036  0.001  0.001 |  0.807  0.528  0.419 |
#   9                 | -0.471  0.105  0.163 | -0.323  0.061  0.077 |  0.636  0.328  0.298 |
#   10                |  0.898  0.383  0.332 |  0.377  0.083  0.059 |  0.356  0.102  0.052 |
#
#   Categories (the 10 first)
#                        Dim.1    ctr    cos2  v.test     Dim.2     ctr    cos2  v.test     Dim.3     ctr    cos2  v.test
#   purpose_infSim       | -0.728  12.274   0.430 -19.347 |  -0.213   1.282   0.037  -5.650 |   0.514  10.443   0.215  13.662 |
#   purpose_infNao       |  0.591   9.952   0.430  19.347 |   0.172   1.040   0.037   5.650 |  -0.417   8.468   0.215 -13.662 |
#   tech_percepPositivo  |  -0.147   1.010   0.216 -13.694 |  -0.001   0.000   0.000  -0.122 |  -0.013   0.013   0.002  -1.197 |
#   tech_percepNegativo  |   1.470  10.125   0.216  13.694 |   0.013   0.001   0.000   0.122 |   0.129   0.132   0.002   1.197 |
#   Nao estou preocupado |  -0.399   3.780   0.136 -10.862 |   0.726  15.362   0.450  19.783 |  -0.357   5.160   0.108  -9.715 |
#   Muito preocupado     |   0.294   1.743   0.055   6.940 |  -0.994  24.397   0.632 -23.455 |  -0.077   0.203   0.004  -1.814 |
#   Irrelevante          |   0.461   1.637   0.037   5.693 |   0.359   1.219   0.023   4.439 |   1.301  22.265   0.297  16.072 |
#   will_contrNao        |   0.731   6.528   0.166  12.000 |   0.788   9.285   0.192  12.930 |   0.788  12.956   0.193  12.941 |
#   will_contrSim        |  -0.226   2.022   0.166 -12.000 |  -0.244   2.876   0.192 -12.930 |  -0.244   4.013   0.193 -12.941 |
#   Brasil               |   0.248   1.573   0.061   7.259 |  -0.804  20.361   0.640 -23.593 |   0.225   2.217   0.050   6.597 |
#
#   Categorical variables (eta2)
#                         Dim.1 Dim.2 Dim.3
#   purpose_inf          | 0.430 0.037 0.215 |
#   tech_percep          | 0.216 0.000 0.002 |
#   privacy_concerns     | 0.139 0.648 0.313 |
#   will_contr           | 0.166 0.192 0.193 |
#   region               | 0.061 0.640 0.050 |
#   prev_exp_WD          | 0.168 0.042 0.244 |
#   mgmt_preoc           | 0.281 0.009 0.058 |
#   part_select          | 0.476 0.014 0.061 |

# Está claro que mucas de las variables son importantes ya que el valor absoluto de V.test es >2.
# El aporte de cada variable a cada dimensión se puede ver claramente en eta2.

# 3.1 Revisamos las dimensiones y su aportación.
## ----------------------------------------

fviz_screeplot(dados.mca, addlabels = TRUE)+geom_hline(yintercept=12.5, linetype=2, color="red")

plot(dados.mca, repel=TRUE)


fviz_mca_biplot(dados.mca, repel = TRUE, ggtheme = theme_minimal())

# Podemos ver la disposición de las variables donde a la derecha en la dimension 1 quedan las variables que predicen una negativa
# en la aceptación y a la izquierda las que predicen aceptación positiva.

# Contribución
## ----------------------------------------
fviz_contrib(dados.mca, choice = "var", axes = 1, top = 10)

fviz_contrib(dados.mca, choice = "var", axes = 2, top = 10)

# Calidad
## ----------------------------------------
fviz_cos2(dados.mca, choice = "var", axes = 1, top = 10)

fviz_cos2(dados.mca, choice = "var", axes = 2, top = 10)

# Vemos el mapa de factores con los indivuos escondidos para tener más claridad en las variables.
plot(dados.mca,invisible="ind")

# 4 Añadimos información adicional.
## ----------------------------------------
## ----------------------------------------

dados.mca2<-MCA(dados, ind.sup = NULL,quanti.sup = 1:2, quali.sup = 11, graph=FALSE)

summary(dados.mca2)

# Eigenvalues
#                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9
# Variance               0.242   0.198   0.142   0.124   0.113   0.102   0.085   0.065   0.054
# % of var.             21.508  17.557  12.605  10.996  10.047   9.083   7.593   5.770   4.842
# Cumulative % of var.  21.508  39.065  51.670  62.666  72.713  81.795  89.389  95.158 100.000
#
# Individuals (the 10 first)
#                         Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
#   1                   |  0.085  0.003  0.008 | -0.509  0.151  0.280 |  0.368  0.110  0.146 |
#   2                   |  0.462  0.101  0.154 | -0.020  0.000  0.000 |  0.516  0.216  0.192 |
#   3                   |  1.014  0.487  0.578 | -0.180  0.019  0.018 |  0.194  0.031  0.021 |
#   4                   |  0.233  0.026  0.048 | -0.146  0.012  0.019 |  0.155  0.020  0.021 |
#   5                   | -0.241  0.028  0.036 | -0.197  0.023  0.024 |  0.997  0.805  0.617 |
#   6                   |  0.191  0.017  0.059 | -0.527  0.161  0.448 | -0.302  0.074  0.147 |
#   7                   | -0.460  0.100  0.202 | -0.094  0.005  0.008 |  0.447  0.162  0.191 |
#   8                   |  0.141  0.009  0.013 |  0.036  0.001  0.001 |  0.807  0.528  0.419 |
#   9                   | -0.471  0.105  0.163 | -0.323  0.061  0.077 |  0.636  0.328  0.298 |
#   10                  |  0.898  0.383  0.332 |  0.377  0.083  0.059 |  0.356  0.102  0.052 |
#
#   Categories (the 10 first)
```

```
#                        Dim.1    ctr   cos2  v.test    Dim.2    ctr   cos2  v.test    Dim.3    ctr   cos2
# purpose_infSim       | -0.728  12.274  0.430 -19.347 |  -0.213   1.282  0.037  -5.650 |   0.514  10.443  0.215
# purpose_infNao       |  0.591   9.952  0.430  19.347 |   0.172   1.040  0.037   5.650 |  -0.417   8.468  0.215
# tech_percepPositivo  | -0.147   1.010  0.216 -13.694 |  -0.001   0.000  0.000  -0.122 |  -0.013   0.013  0.002
# tech_percepNegativo  |  1.470  10.125  0.216  13.694 |   0.013   0.001  0.000   0.122 |   0.129   0.132  0.002
# Nao estou preocupado | -0.399   3.780  0.136 -10.862 |   0.726  15.362  0.450  19.783 |  -0.357   5.160  0.108
# Muito preocupado     |  0.294   1.743  0.055   6.940 |  -0.994  24.397  0.632 -23.455 |  -0.077   0.203  0.004
# Irrelevante          |  0.461   1.637  0.037   5.693 |   0.359   1.219  0.023   4.439 |   1.301  22.265  0.297
# will_contrNao        |  0.731   6.528  0.166  12.000 |   0.788   9.285  0.192  12.930 |   0.788  12.956  0.193
# will_contrSim        | -0.226   2.022  0.166 -12.000 |  -0.244   2.876  0.192 -12.930 |  -0.244   4.013  0.193
# Brasil               |  0.248   1.573  0.061   7.259 |  -0.804  20.361  0.640 -23.593 |   0.225   2.217  0.050
#                        v.test
#   purpose_infSim       13.662 |
#   purpose_infNao      -13.662 |
#   tech_percepPositivo  -1.197 |
#   tech_percepNegativo   1.197 |
#   Nao estou preocupado -9.715 |
#   Muito preocupado     -1.814 |
#   Irrelevante          16.072 |
#   will_contrNao        12.941 |
#   will_contrSim       -12.941 |
#   Brasil               6.597 |
#
#   Categorical variables (eta2)
#                         Dim.1 Dim.2 Dim.3
#   purpose_inf          | 0.430 0.037 0.215 |
#   tech_percep          | 0.216 0.000 0.002 |
#   privacy_concerns     | 0.139 0.648 0.313 |
#   will_contr           | 0.166 0.192 0.193 |
#   region               | 0.061 0.640 0.050 |
#   prev_exp_WD          | 0.168 0.042 0.244 |
#   mgmt_preoc           | 0.281 0.009 0.058 |
#   part_select          | 0.476 0.014 0.061 |
#
#   Supplementary categories
#                         Dim.1    cos2  v.test     Dim.2    cos2  v.test     Dim.3    cos2  v.test
#   acceptWDNao          |  0.774   0.132  10.703 |   0.136   0.004   1.876 |   0.013   0.000   0.184 |
#   acceptWDSim          | -0.170   0.132 -10.703 |  -0.030   0.004  -1.876 |  -0.003   0.000  -0.184 |
#
#   Supplementary categorical variables (eta2)
#                         Dim.1 Dim.2 Dim.3
# accept_WD              | 0.132 0.004 0.000 |
#
#   Supplementary continuous variables
#                         Dim.1    Dim.2    Dim.3
#   age                  | -0.157 |  0.192 |  0.030 |
#   exp_qual             | -0.145 | -0.220 | -0.100 |


# Aunque las variables cuantitivas parecen no influir demasiado, lo que es muy claro es que la respuesta a aceptación
# está muy correlacionada con la dimensión 1 y por tanto con las variables que se situan a ambos lados de la dimensión

# la dimensión dos, separa claramente los valores de Europa y Brasil , y como se comprueba en los datos la preocupación por la
# privacidad de los datos está altamente correlacionada con la región.

# 5 Interpetamos los resultados
## ----------------------------------------
## ----------------------------------------

fviz_screeplot(dados.mca2, addlabels = TRUE)+geom_hline(yintercept=12.5, linetype=2, color="red")

plot(dados.mca2, repel=TRUE)

fviz_mca_biplot(dados.mca2, repel = TRUE, ggtheme = theme_minimal())

# Podemos ver la disposición de las variables donde a la derecha en la dimension 1 quedan las variables que predicen una negativa
# en la aceptación y a la izquierda las que predicen aceptación positiva.

# Contribución
## ----------------------------------------
fviz_contrib(dados.mca2, choice = "var", axes = 1, top = 10)

fviz_contrib(dados.mca2, choice = "var", axes = 2, top = 10)


# Calidad
## ----------------------------------------
fviz_cos2(dados.mca2, choice = "var", axes = 1, top = 10)

fviz_cos2(dados.mca2, choice = "var", axes = 2, top = 10)

# Vemos el mapa de factores con los indivuos escondidos para tener más claridad en las variables.
plot(dados.mca2,invisible="ind")
```

```
fviz_mca_ind(dados.mca2, label = "none", habillage = "accept_WD", palette = c("#00AFBB", "#E7B800"), addEllipses = TRUE, ellipse.type
= "confidence", ggtheme = theme_minimal())

plotellipses(dados.mca2,keepvar=c("purpose_inf","tech_percep","privacy_concerns","will_contr"))

plotellipses(dados.mca2,keepvar=c("region","prev_exp_WD","mgmt_preoc","part_select"))

plotellipses(dados.mca2,keepvar=c("accept_WD"))


## # Para obtener las predicciones
dados.mca2$quali.sup
dados.mca2$quanti
## dados.mca2$ind.sup
# No tenemos individuos suplementarios

# 6 Realizamos análisis cluster jerárquico.
## ----------------------------------------
## ----------------------------------------

dados.hcpc<-HCPC(dados.mca2, cluster.CA = "rows")

# Realizamos el análisis cluster y vemos las agrupaciones creadas según el nivel de corte.

dados.hcpc$desc.axes

# Individuos representativos de cada grupo (filas)
## ----------------------------------------------
dados.hcpc$desc.ind$para


# Miramos la especificidad de los individuos (que tan lejos está el individuo de otros grupos)
## ----------------------------------------------
dados.hcpc$desc.ind$dist

# En este caso la información no es muy clara respecto a que agrupa en cada cluster, pero se ha realizado para tener el ejercicio completo.
```

7.2 Fichero de R para la resolución del modelo por Regresión Logística

```
# 1 Preparamos los datos
## ----------------------------------------
## ----------------------------------------

# 1.1 Cargamos los datos
## ----------------------------------------

library(PASWR)
library(dplyr)
library(tidyverse)
library(summarytools)


dados <- read_delim("C:/Users/u971259/Desktop/Oscar/Seguridad/GS/Proyecto/Resultado da pesquisa/Estudio estadistico/20201121/DadosEnviados20201026_Master3.csv", delim = ";", escape_double = FALSE, col_types = cols(purpose_inf = col_factor(levels = c("purpose_infSim","purpose_infNao")), tech_percep = col_factor(levels = c("tech_percepPositivo", "tech_percepNegativo")), privacy_concerns = col_factor(levels = c("Nao estou preocupado", "Muito preocupado", "Irrelevante")), will_contr = col_character()), trim_ws = TRUE)

dados$age <- as.integer(dados$age)
dados$exp_qual <- as.integer(dados$exp_qual)
dados$purpose_inf <- as.factor(dados$purpose_inf)
dados$tech_percep <- as.factor(dados$tech_percep)
dados$privacy_concerns<-as.factor(dados$privacy_concerns)
dados$will_contr <- as.factor(dados$will_contr)
dados$region <- as.factor(dados$region)
dados$prev_exp_WD <- as.factor(dados$prev_exp_WD)
dados$mgmt_preoc <- as.factor(dados$mgmt_preoc)
dados$part_select <- as.factor(dados$part_select)
dados$accept_WD <- as.factor(dados$accept_WD)

summary(dados)
str(dados)


# 1) Explorarmos los datos


dados %>% group_by(accept_WD,part_select,mgmt_preoc,prev_exp_WD,region, will_contr,privacy_concerns,tech_percep,purpose_inf) %>% summarize(mean = mean(age,na.rm = TRUE),sd=sd(age,na.rm = TRUE),n=n())

dados %>% group_by(accept_WD,purpose_inf) %>% summarize(mean = mean(age,na.rm = TRUE),sd=sd(age,na.rm = TRUE),n=n())

ggplot(dados,aes(x=as.factor(accept_WD),y=age,fill=as.factor(tech_percep)))+geom_boxplot()

ggplot(dados,aes(x=as.factor(accept_WD),y=age,fill=as.factor(purpose_inf)))+geom_boxplot()


dados %>% group_by(accept_WD,part_select,mgmt_preoc,prev_exp_WD,region, will_contr,privacy_concerns,tech_percep,purpose_inf) %>% summarize(mean = mean(exp_qual,na.rm = TRUE),sd=sd(exp_qual,na.rm = TRUE),n=n())

dados %>% group_by(accept_WD,purpose_inf) %>% summarize(mean = mean(exp_qual,na.rm = TRUE),sd=sd(exp_qual,na.rm = TRUE),n=n())

ggplot(dados,aes(x=as.factor(accept_WD),y=exp_qual,fill=as.factor(tech_percep)))+geom_boxplot()

ggplot(dados,aes(x=as.factor(accept_WD),y=exp_qual,fill=as.factor(purpose_inf)))+geom_boxplot()


ctable(dados$purpose_inf,as.factor(dados$accept_WD))

ctable(dados$tech_percep,as.factor(dados$accept_WD))

ctable(dados$privacy_concerns,as.factor(dados$accept_WD))

ctable(dados$will_contr,as.factor(dados$accept_WD))

ctable(dados$region,as.factor(dados$accept_WD))

ctable(dados$prev_exp_WD,as.factor(dados$accept_WD))

ctable(dados$mgmt_preoc,as.factor(dados$accept_WD))

ctable(dados$part_select,as.factor(dados$accept_WD))


# 2) Dividimos los datos en dos

library(caret)
index <- createDataPartition(y = dados$accept_WD, p = 0.80, list = FALSE)
```

```
traindados <- dados[index,]
testdados <- dados[-index,]

dim(traindados)
dim(testdados)


# 3)crear un modelo

# 3.1)crear un modelo con el predictor categórico

fit.purp_inf <- glm(accept_WD ~ purpose_inf, data=traindados, family=binomial)
summary(fit.purp_inf)
coef(fit.purp_inf)
# <!-- odds ratio-->
exp(coef(fit.purp_inf))
# <!-- intervalo confianza-->
exp(confint(fit.purp_inf))

library(effects)
# <!-- probabilidades segun categorias-->
allEffects(fit.purp_inf)
# <!-- IConfianza para probabilidades segun categorias-->
summary(allEffects(fit.purp_inf))
# <!-- Gráficos de las probabilidades segun categorias-->
plot(allEffects(fit.purp_inf))

# El valor de la variable student es significativo como se parecia la ejecutar la función Summary.Se aprecia luego al imprimir los efectos.
# El OR es un 46% mayor para los estudiantes.


# 3.2)crear un modelo con el primer predictor continuo

traindados %>% group_by(accept_WD) %>% summarize(mean=mean(age),sd = sd(age))
boxplot(age ~ accept_WD, data = traindados)

fit.age <- glm(accept_WD ~ age, data=traindados, family=binomial)
summary(fit.age)
coef(fit.age)
# <!-- odds ratio-->
exp(coef(fit.age))
# <!-- intervalo confianza-->
exp(confint(fit.age))

# library(effects)
# <!-- probabilidades segun categorias-->
allEffects(fit.age)
# <!-- IConfianza para probabilidades segun categorias-->
summary(allEffects(fit.age))
# <!-- Gráficos de las probabilidades segun categorias-->
plot(allEffects(fit.age))

# El valor de la variable balance también es significativo como se aprecia la ejecutar la función Summary. Aunque el valor es pequeño
porque la variable va
# desde 18 hasta 66. Es un valor decreciente con el aumento de la variable


# 3.3)crear un modelo con el segundo predictor continuo

traindados %>% group_by(accept_WD) %>% summarize(mean=mean(exp_qual),sd = sd(exp_qual))
boxplot(exp_qual ~ accept_WD, data = traindados)

fit.qual_exp <- glm(accept_WD ~ exp_qual, data=traindados, family=binomial)
summary(fit.qual_exp)
coef(fit.qual_exp)
# <!-- odds ratio-->
exp(coef(fit.qual_exp))
# <!-- intervalo confianza-->
exp(confint(fit.qual_exp))

# library(effects)
# <!-- probabilidades segun categorias-->
allEffects(fit.qual_exp)
# <!-- IConfianza para probabilidades segun categorias-->
summary(allEffects(fit.qual_exp))
# <!-- Gráficos de las probabilidades segun categorias-->
plot(allEffects(fit.qual_exp))

# El valor de la variable exp_qual es significativo  y en teoría decrece el OR en 32% por cada unidad de aumento del exp_qual.


# 3.4)crear un modelo con  cuatro de las variables, dos cualitativas y dos cuantitativas.


fit.1234 <- glm(accept_WD ~ age+exp_qual+purpose_inf+tech_percep, data=traindados, family=binomial)
summary(fit.1234)
coef(fit.1234)
```

```
# <!-- odds ratio-->
exp(coef(fit.1234))
# <!-- intervalo confianza-->
exp(confint(fit.1234))

# library(effects)
# <!-- probabilidades segun categorias-->
allEffects(fit.1234)
# <!-- IConfianza para probabilidades segun categorias-->
summary(allEffects(fit.1234))
# <!-- Gráficos de las probabilidades segun categorias-->
plot(allEffects(fit.1234))

# Las cuatro variables son significativas en el modelo.


# 3.5)crear un modelo con 4 predictores con interaccion

fit.1234int <- glm(accept_WD ~ age*exp_qual*purpose_inf*tech_percep, data=traindados, family=binomial)
summary(fit.1234int )
coef(fit.1234int )
# <!-- odds ratio-->
exp(coef(fit.1234int ))
# <!-- intervalo confianza-->
exp(confint(fit.1234int ))

# library(effects)
# <!-- probabilidades segun categorias-->
allEffects(fit.1234int )
# <!-- IConfianza para probabilidades segun categorias-->
summary(allEffects(fit.1234int ))
# <!-- Gráficos de las probabilidades segun categorias-->
plot(allEffects(fit.1234int ))

library(sjPlot)
plot_model(fit.1234int, type="int")

# Las interacciones cuadruples como era de esperar no son fáciles de interpretar y además no dan ningún valor significativo.


# 3.6)crear un modelo con todas las variables.

fit.12345678910 <- glm(accept_WD ~ age + exp_qual + purpose_inf + tech_percep + privacy_concerns + will_contr + region + prev_exp_WD + mgmt_preoc + part_select, data=traindados, family=binomial)
summary(fit.12345678910)
coef(fit.12345678910)
summary(fit.12345678910)$coef

# Los resultados son muy interesantes porque muestran que las variables significativas son . Age,exp_qual,purpose_inf,tech_percep,privacy_concerns,will_contr y part_select aunque esta última solo marginalmente


# <!-- odds ratio-->
exp(coef(fit.12345678910))
# <!-- intervalo confianza-->
exp(confint(fit.12345678910))

# library(effects)
# <!-- probabilidades segun categorias-->
allEffects(fit.12345678910)
# <!-- IConfianza para probabilidades segun categorias-->
summary(allEffects(fit.12345678910))
# <!-- Gráficos de las probabilidades segun categorias-->
plot(allEffects(fit.12345678910))

# 3.6.2)crear un modelo con las variables significativas del modelo anterior

fit.rev <- glm(accept_WD ~ age + exp_qual + purpose_inf + tech_percep + privacy_concerns + will_contr, data=traindados, family=binomial)
summary(fit.rev)
coef(fit.rev)
summary(fit.rev)$coef

# Los resultados son muy interesantes porque muestran que las variables significativas son . Age,exp_qual,purpose_inf,tech_percep,privacy_concerns,will_contr y part_select aunque esta última solo marginalmente


# <!-- odds ratio-->
exp(coef(fit.rev))
# <!-- intervalo confianza-->
exp(confint(fit.rev))

# library(effects)
# <!-- probabilidades segun categorias-->
allEffects(fit.rev)
# <!-- IConfianza para probabilidades segun categorias-->
summary(allEffects(fit.rev))
```

```
# <!-- Gráficos de las probabilidades segun categorias-->
plot(allEffects(fit.rev))


# 3.7) stepwise

# <!-- Stepwise -->

fit.all <- glm(accept_WD ~ .^2,data= traindados,family = binomial)
(fit.step <- step(fit.all, trace=0))
summary(fit.step)

# 3.8 Evaluación de los diferentes modelos.

AIC(fit.12345678910,fit.1234,fit.1234int,fit.age,fit.rev,fit.purp_inf, fit.qual_exp,fit.step)

# El indice AIC me dice que el modelo más adecuado es fit.step aunque es un modelo extremadamente complejo y optamos por fit.rev que es el modelo en el que se usan solo las variables
# que salieron significativas en el modelo completo sin interacción.


# 3.9 Evaluación de las variables más importantes en el caso de los dos modelos más representativos

# 3.9.1 Evaluación de las variables más importantes mediante VArImp

set.seed(1979)

fit.caret <- train(accept_WD ~ age + exp_qual + purpose_inf + tech_percep +
                     privacy_concerns + will_contr + region + prev_exp_WD + mgmt_preoc +
                     age:will_contr + age:mgmt_preoc + exp_qual:tech_percep +
                     exp_qual:will_contr + exp_qual:prev_exp_WD + purpose_inf:prev_exp_WD +
                     tech_percep:region + tech_percep:prev_exp_WD,data = traindados, method="glm", na.action=na.omit)


(imp <- varImp(fit.caret, scale=FALSE))
plot(imp)


fit.caret2 <- train(accept_WD ~ age + exp_qual + purpose_inf + tech_percep + privacy_concerns + will_contr,data = traindados, method="glm", na.action=na.omit)

(imp2 <- varImp(fit.caret2, scale=FALSE))
plot(imp2)


# 3.9.2 Análisis de dominancia.

# <!-- Analisis de dominancia para evaluar importancia predictores-->

# 3.9.2.1 Análisis de dominancia modelo con menor AIC.

library(dominanceanalysis)
res <- dominanceAnalysis(fit.step)
getFits(res,"r2.m")

plot(res,which.graph="conditional",fit.function="r2.m")

plot(res,which.graph="general",fit.function="r2.m")


# 3.9.2.1 Análisis de dominancia modelo preferido.

res2 <- dominanceAnalysis(fit.rev)
getFits(res2,"r2.m")

plot(res2,which.graph="conditional",fit.function="r2.m")

plot(res2,which.graph="general",fit.function="r2.m")



# 4  pruebas bondad de ajuste ----------------------------------------------------------

#4.1 coeficiente pseudo-R2

#4.1.1 coeficiente pseudo-R2 modelo fit.step
library(performance)
r2(fit.step)
model_performance(fit.step)

#4.1.2 coeficiente pseudo-R2 modelo fit.rev

# library(performance)
r2(fit.rev)
model_performance(fit.rev)

# A partir de este momento utilizamos solo el modelo lógico, aunque tenga menos explicación
```

ya que la intención era ver si las variables imporantes eran las obtenidas en el primer estudio.

#4.2 Prueba global de ajuste de Hosmer-Lemeshow

```
library(ResourceSelection)
data<-na.omit(fit.rev$data[,c("accept_WD","age","exp_qual","purpose_inf","tech_percep", "privacy_concerns", "will_contr")])
hoslem.test(x= data$accept_WD, y=fitted(fit.rev))
```

#H=0 el modelo ajust a los datos , si pvalue>0,05 no podemos rechazar y #por tanto se ajusta (NO ES EL CASO)

#aunque las variables son significativas , el modelo no es adecuado para explicar los datos, aunque en modelos de encuestas
y teorías sociológicas los valores de correlación no suelen ser muy elevados.

Predicciones del modelo ----------

#4.3 Predicciones en regresión logistica
<!-- Como vamos a preveer valores -->

```
  # <!-- predicciones en escala logit -->
  head(predict(fit.rev,type="link"))

# <!-- predicciones de probibilidad del evento codificado como 1 -->
  head(prob <- predict(fit.rev,type="response"))

# <!-- ponemos el corte y clasificamos -->
  pred.classes <- ifelse(prob > 0.5,"acceptWDSim","acceptWDNao")
head(pred.classes)

# <!-- Evaluamos precisión del modelo -->

  mean(pred.classes == testdados$accept_WD)

  # Un 75% de los valores predichos fueron acertados

# <!-- Calculamos matriz de confusión -->
library(caret)
pdata <- predict(fit.rev, newdata = testdados, type="response")
pdataclases <- ifelse(pdata > 0.5,"acceptWDSim","acceptWDNao")
confusionMatrix(data = as.factor(pdataclases),reference = as.factor(testdados$accept_WD),positive = "acceptWDSim")
# <!-- Aunque la sensitividad es alta , la especifidad es muy baja. el ratio de verdaderos negativos es muy bajo.


# <!-- calculamos el valor Roc (Receiver Operating Characteristic) que deberia ser mayor que 80% -->

  performance_roc(fit.rev)

  # Es mayor que el 80%


# 4.4 Mostrar resultados
# <!-- representatciones -->
#
#4.4.1 Gráfico del modelo
#   <!-- gráfico -->

  ggplot(traindados,aes(x=age,y=
exp_qual,fill=accept_WD))+geom_point(position=position_jitter(height=0.05),aes(y=exp_qual,color=accept_WD))+stat_smooth(method="
glm",method.args=list(family="binomial"),se=TRUE)

# Se pueden hacer otros muchos gráfios para visualizar las variables de interes...esta es solo un ejemplo.



#4.4.2 tabla del modelo
# <!-- tabla -->

  library(texreg)
texreg(fit.rev,single.row=TRUE)

#4.4.3 report del modelo
# <!-- report -->

  library(report)
report(fit.rev)
summary(report(fit.rev))

# El modelo aunque tiene coeficientes significativos no pasa el test the hemswor y tendria problemas para estimar, el R2 también es muy
bajo.
# Sería interesante explorar si otras variables que no estén en el modelo pueden añadir explicación.Sobre todo las interacciónes, pero
como el objetivo era
# hacer un estudio comparativo para ver si las mismas variables eran importantes. El objetivo ha sido realizado.
```