Universitat Oberta de Catalunya

uoc.edu

Análisis estadístico de los factores de riesgo asociados a patologías de columna lumbar para la población ocupada.

Vanessa Mariela Mena López
Máster en Bioinformática y Bioestadística
Área 2 - Análisis de datos y técnicas de clustering

Daniel Fernández Martínez
Marc Maceira Duch

Junio 2021

Vanessa Mena López

A) Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial 3.0 España de Creative Commons

B) GNU Free Documentation License (GNU FDL)

Copyright © AÑO TU-NOMBRE.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.

A copy of the license is included in the section entitled "GNU Free Documentation License".

C) Copyright

© (el autor/a)

Reservados todos los derechos. Está prohibido la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la impresión, la reprografía, el microfilme, el tratamiento informático o cualquier otro sistema, así como la distribución de ejemplares mediante alquiler y préstamo, sin la autorización escrita del autor o de los límites que autorice la Ley de Propiedad Intelectual.

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis estadístico de los factores de riesgo asociados a patologías de columna lumbar para la población ocupada*.
Nombre del autor:	*Vanessa Mariela Mena López*
Nombre del consultor/a:	*Daniel Fernández Martínez*
Nombre del PRA:	*Marc Maceira Duch*
Fecha de entrega (mm/aaaa):	*06/2021*
Titulación:	*Máster Universitario en Bioinformática y Bioestadística UOC-UB*
Área del Trabajo Final:	*Análisis de datos y técnicas de clustering*
Idioma del trabajo:	*Castellano*
Número de créditos:	*15 créditos*
Palabras clave	*Patologías lumbares, lumbalgia, factores de riego.*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

La región lumbar se ubica en la parte inferior de la columna vertebral, permite el giro, la torsión y la flexión, proporciona fuerza para pararse, caminar y levantar objetos. (1). El presente estudio permite utilizar análisis estadísticos para identificar los factores de riesgo asociados a las patologías lumbares, diagnosticadas en radiografías. Los datos son tomadas de las historias clínicas de 668 pacientes ecuatorianos con empleo. La metodología estadística incluye un análisis exploratorio univariante y bivariante, la aplicación de modelos probabilísticos para detectar variables asociadas significativamente con las patologías y la evaluación del mejor modelo predictivo.

Se evidencia los mejores resultados con el modelo Logit, con una sensibilidad = 84%, especificidad = 77%, área bajo la Curva de ROC = 0.81 y precisión = 79%, con un corte optimo= 0.48.

El perfil de los pacientes que tienen la patología más frecuente (lumbalgia) y en orden de importancia son: presencia de dolor, menor edad, sexo femenino, tienden a exponerse a factores de riesgo psicosocial en su trabajo, menor IMC y menor riesgo de tener factores mecánicos. Mientras que el perfil de los pacientes que tienen las patologías menos frecuentes (mielopatía, radiculopatía, otros trastornos de disco invertebral) son: sin presencia de dolor, mayor edad,

sexo masculino, menor riesgo de tener factores psicosociales, mayor IMC y tienden a exponerse a factores de riesgo mecánicos

Abstract (in English, 250 words or less):

The lumbar region is located in the lower part of the spine, allows twisting, and bending, provides strength to stand, walk and lift objects. (1). The present study makes it possible to use statistical analyzes to identify the risk factors associated with lumbar pathologies, diagnosed on radiographs. The data are taken from the medical records of 668 Ecuadorian employed patients. The statistical methodology includes a univariate and bivariate exploratory analysis, the application of probabilistic models to detect variables significantly associated with the pathologies, and the evaluation of the best predictive model.

The best results are evidenced with the Logit model, with a sensitivity = 84%, specificity = 77%, area under the ROC curve = 0.81 and precision = 79%, with an optimal cut = 0.48.

The profile of patients who have the most frequent pathology (low back pain) and in order of importance are: thenpresence of pain, younger age, female sex, they tend to be exposed to psychosocial risk factors at work, lower BMI, and lower risk of having mechanical factors. While the profile of patients who have the less frequent pathologies (myelopathy, radiculopathy, other disorders of the invertebral disc) are: no presence of pain, older age, male sex, lower risk of having psychosocial factors, higher BMI and tend to be exposed mechanical risk factors.

Índice

Lista de figuras

Lista de tablas

1 Resumen

La región lumbar se ubica en la parte inferior de la columna vertebral, permite el giro, la torsión y la flexión, proporciona fuerza para pararse, caminar y levantar objetos. (1). El presente estudio permite identificar los factores de riesgo asociados a las patologías lumbares, diagnosticadas en radiografías. Los datos son tomadas de las historias clínicas de 668 pacientes ecuatorianos con empleo, la metodología sigue un enfoque cuantitativo, retrospectivo, de corte transversal, estadísticamente se realiza un análisis univariante y bivariante, se identifican las variables asociadas significativamente con las patologías y se evalúa el mejor modelo predictivo.

Se evidencia los mejores resultados con el modelo Logit, con una sensibilidad = 84%, especificidad = 77%, área bajo la Curva de ROC = 0.81 y precisión = 79%. Además, si el modelo estima una probabilidad ≥ 48% (corte óptimo) se afirma que se va a tener 77% de seguridad de tener lumbalgia, mientras que si es lo contrario se tiene un 74% de seguridad de tener las otras patologías

El perfil de los pacientes que tienen la patología más frecuente “lumbalgia” y en orden de importancia son: presencia de dolor, menor edad, sexo femenino, tienden a exponerse a factores de riesgo psicosocial en su trabajo, menor IMC y menor riesgo de tener factores mecánicos. Mientras que el perfil de los pacientes que tienen las patologías menos frecuentes “mielopatía, radiculopatía, otros trastornos de disco invertebral” son: sin presencia de dolor, mayor edad, sexo masculino, menor riesgo de tener factores psicosociales, mayor IMC y tienden a exponerse a factores de riesgo mecánicos.

2 Introducción

2.1 Contexto y justificación del Trabajo

El dolor provocado a nivel de la región lumbar implica una pérdida económica, tanto en terapias médicas, diagnósticos por imágenes (radiografías, tomografías entre otras), hospitalizaciones, intervenciones quirúrgicas, incapacidad y tratamiento. Afecta a la población ocupada en la pérdida de días laborales y disminución de la productividad laboral, entre otros problemas.

“Se estima que el 60-70% de las personas adultas presentan un episodio de dolor lumbar a lo largo de su vida, representa una de las principales causas de limitación física, que puede durar días o pude ser crónico, … es la segunda causa de requerimiento de atención médica en los países desarrollados, la tercera causa de intervención quirúrgica, incapacidad funcional crónica después de las afecciones respiratorias, traumatismos y la quinta en frecuencia de hospitalización” (2)

La Organización Mundial de la Salud (OMS) señala que “El dolor lumbar es la causa principal de vivir con discapacidad durante años en todo el mundo. En 2018, un grupo de trabajo internacional pidió a la OMS que prestara más atención a la carga del dolor lumbar y a la necesidad de evitar soluciones excesivamente médicas.” (3)

“Aunque varias de las patologías lumbares tienen asociación significativa y positiva con dolor lumbar, existe un porcentaje de individuos sintomáticos en los que la radiografía no muestra hallazgos anormales, mientras que pacientes asintomáticos pueden demostrar una amplia gama de alteraciones, por lo que

debe insistirse en que la valoración ocupacional siempre debe hacerse teniendo en cuenta el perfil del cargo y la presencia o no de signos y síntomas". (4)

"Las técnicas de imagen, como la radiografía de columna lumbar, la tomografía computarizada (TC) y la resonancia magnética (RM), junto con una exploración clínica adecuada, determinan la causa del dolor lumbar solamente en un 15 % de los casos. Para una aproximación diagnóstica del dolor lumbar se utiliza una adecuada anamnesis y examen físico...". (5)

La presente investigación surge de la necesidad de identificar los factores de riesgo de las patologías de la columna lumbar, diagnosticadas en radiografías, en la población ecuatoriana ocupada, la región lumbar se ubica en la parte inferior de la columna, permite el giro, la torsión y la flexión, proporciona fuerza para pararse, caminar y levantar objetos, de manera que participa en casi todas las actividades cotidianas. (1)

Ubicación de la columna lumbar

Figura 1. Región lumbar.
Fuente: Manual MSD (1)

Con la aplicación de modelos predictivos adecuados, el presente estudio permite identificar los factores que pueden elevar el riesgo de que un paciente presente patologías lumbares, basada en una exploración física, hábitos, antecedentes personales y factores de riesgo en el trabajo. Identificar los factores de riesgo permitirá dar recomendaciones al paciente y evitar los factores que son modificables o minimizar su impacto, prevenir futuras anomalías lumbares y posibles dolor lumbar que afectan la salud y la perdida de jornadas de trabajo, orientar las conductas terapéuticas que debe seguir el paciente y prevenir complicaciones mayores, lo cual contribuye a tener una base científica para prevenir enfermedades y promover una vida saludable

2.2 Objetivos del Trabajo

2.2.1 Objetivo general:

- Evaluar los factores de riesgo de patologías lumbares, en la población ocupada, diagnosticadas en radiografías.

2.2.2 Objetivos específicos:

- Analizar posibles valores faltantes y atípicos en la base de datos y las variables que tienen mayor incidencia en las patologías lumbares más frecuentes, basadas en exploración física, hábitos, antecedentes personales y factores de riesgo en el trabajo.

- Identificar los factores que pueden elevar el riesgo de que un paciente presente patologías lumbares, basadas en una exploración física, hábitos, antecedentes personales y factores de riesgo en el trabajo, diagnosticados en radiografías.

- Construir un perfil del paciente con patologías lumbares más frecuentes, diagnosticadas en radiografías, basadas en exploración física, hábitos, antecedentes personales y factores de riesgo en el trabajo

- Validar las técnicas estudiados, que permitan aceptar los resultados obtenidos

2.3 Enfoque y método seguido

El enfoque a emplearse en la presente investigación es el enfoque cuantitativo, retrospectivo, de corte transversal, pues los datos son tomadas de las historias clínicas de diferentes pacientes en su primera atención, entre el año 2018 a 2020, que permite a través de técnicas y herramientas estadísticas obtener descripciones, inferencias y predicciones con precisión.

Técnicas de obtención de datos:

Los datos que permiten encaminar la investigación, se obtiene de fuentes secundarias, mediante historias clínicas, de diferentes pacientes con patologías lumbares diagnosticadas en radiografías en su primera atención, provenientes de la población ocupada con seguro privado de salud de la empresa ecuatoriana "Biodimed" entre el año 2018 al 2020

Técnicas estadísticas de análisis de datos:

Para dar cumplimiento a los objetivos se iniciará con un análisis descriptivo univariante y bivariante de las frecuencias de patologías lumbares, según exámenes físicos, hábitos, factores de riesgo en el trabajo y antecedentes personales, se evaluará con pruebas de hipótesis adecuadas las variables que están relacionadas significativamente con las patologías de la columna lumbar.

Con la aplicación de modelos predictivos adecuados, se prevee evaluar el mejor modelo que permita identificar los factores que pueden elevar el riesgo de que un paciente presente patologías lumbares y de paso a construir un perfil del paciente

Para identificar los factores de riesgo en las patologías lumbares, se basará en la exploración física, hábitos, factores de riesgo en el trabajo y antecedentes personales, como sub-dimensiones de estos factores se tiene:

Patologías de la columna lumbar, codificadas según el CIE10 como: M51.06, M51.16, M51.26, M51.36, M51.86, M51.9, M53.26, M54.16 y M54.5, patologías de la región lumbar, cabe indicar que no incluye la región cervical, dorsal, lumbo-sacra y sacra – coccígea. (Ver *Tabla 1*)

Exploración física vertebral realizada por el especialista: curvatura, flexión, dolor y lasegue.

Hábitos: consumo de alcohol, sustancias psicotrópicas, cigarrillo y práctica de actividad física (deporte).

Factores y peligros de riesgo que está expuesto el paciente en el trabajo: accidentes mayores, biológicos, ergonómicos, físicos, mecánicos, químicos y psicosociales.

Antecedentes personales: edad, peso, talla, sexo, cargo, índice de masa corporal (IMC).

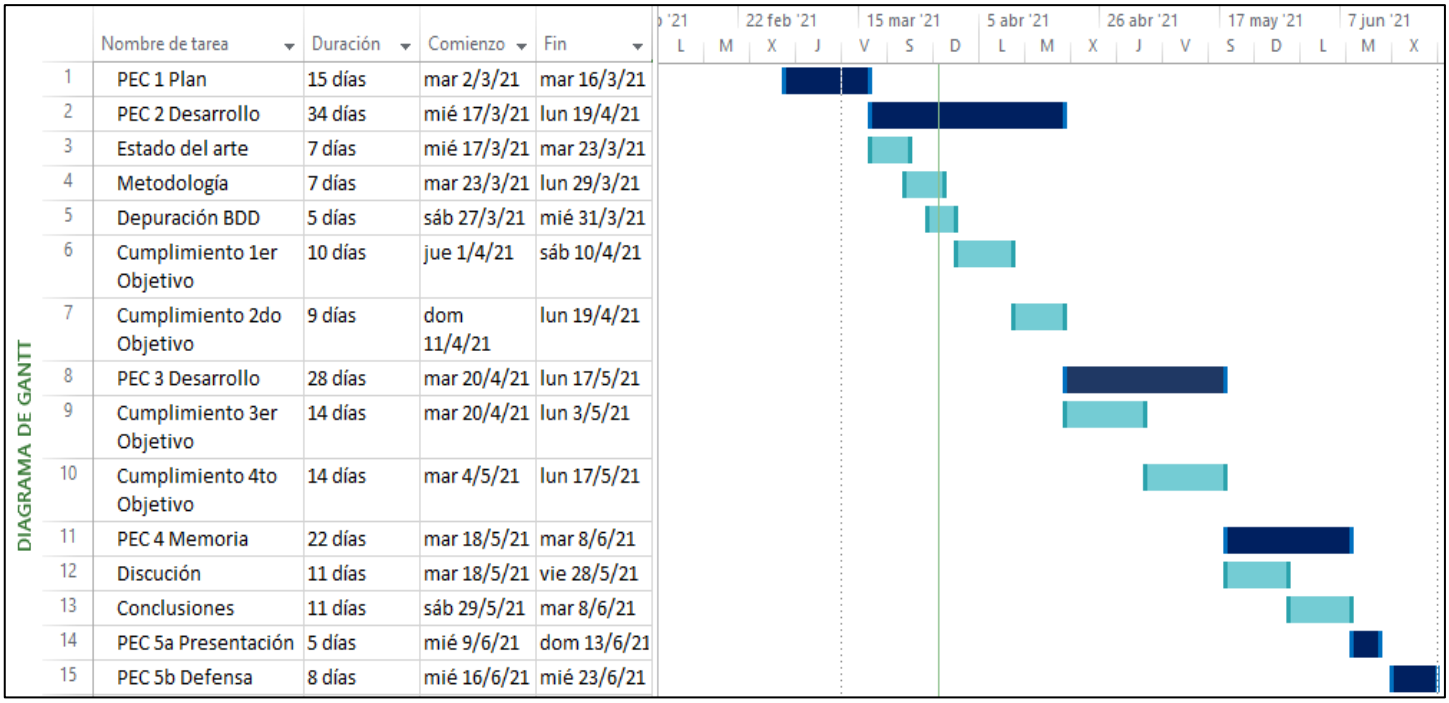
2.4 Planificación del Trabajo

2.4.1 Tareas

A continuación, se enumeran las tares llevadas a cabo para demostrar los objetivos y concluir con la investigación

- Búsqueda de información para construir el estado del arte
- Tratamiento de base de datos
- Identificar el tipo de variables: cuantitativas o cualitativas
- Evaluar las pruebas de hipótesis que permitan identificar las variables relacionadas con las patologías
- Identificar el mejor modelo predictivo que permita dar respuesta a los objetivos
- Análisis de datos
- Obtención de resultados
- Elaboración de discusión
- Elaboración de la conclusión

2.4.2 Calendario



	Nombre de tarea	Duración	Comienzo	Fin
1	PEC 1 Plan	15 días	mar 2/3/21	mar 16/3/21
2	PEC 2 Desarrollo	34 días	mié 17/3/21	lun 19/4/21
3	Estado del arte	7 días	mié 17/3/21	mar 23/3/21
4	Metodología	7 días	mar 23/3/21	lun 29/3/21
5	Depuración BDD	5 días	sáb 27/3/21	mié 31/3/21
6	Cumplimiento 1er Objetivo	10 días	jue 1/4/21	sáb 10/4/21
7	Cumplimiento 2do Objetivo	9 días	dom 11/4/21	lun 19/4/21
8	PEC 3 Desarrollo	28 días	mar 20/4/21	lun 17/5/21
9	Cumplimiento 3er Objetivo	14 días	mar 20/4/21	lun 3/5/21
10	Cumplimiento 4to Objetivo	14 días	mar 4/5/21	lun 17/5/21
11	PEC 4 Memoria	22 días	mar 18/5/21	mar 8/6/21
12	Discución	11 días	mar 18/5/21	vie 28/5/21
13	Conclusiones	11 días	sáb 29/5/21	mar 8/6/21
14	PEC 5a Presentación	5 días	mié 9/6/21	dom 13/6/21
15	PEC 5b Defensa	8 días	mié 16/6/21	mié 23/6/21

2.4.3 Análisis de riesgos

Los factores que pueden repercutir negativamente en el seguimiento del plan de trabajo y en la consecución del proyecto son: el tiempo que se dispone para llevar a cabo la investigación, el alcance de la investigación ya que abarcar varias patologías lumbares, lo cual implica un amplio trabajo y tiempo, y no disponer de

la base de datos con los atributos que necesito y a tiempo como se comprometió la empresa que me facilitará, son los factores de riego más importantes

2.5 Breve sumario de contribuciones y productos obtenidos

Al final del proyecto se espera obtener los siguientes documentos que serán evaluados en las PECs:

2.5.1. Plan de trabajo

2.5.2. Memoria

2.5.3. Presentación virtual

2.5.4. Autoevaluación del proyecto

Todos estos productos obtenidos permitirán determinar los factores de riesgo asociado a patologías de columna lumbar

2.6 Estructura del proyecto

El proyecto de investigación está estructurado en las siguientes etapas:

- En primer lugar, la planificación y organización de la investigación que consiste en establecer con claridad y precisión lo que se quiere investigar, se señala la importancia y justificación de la investigación, identificación de objetivos y se realiza una breve explicación de la planificación necesaria para llevar a cabo la investigación.
- En segundo lugar, se incluye el estado del arte, donde se muestra el estado actual de conocimiento en el campo de investigación, los aspectos y resultados relevantes sobre el tema tratado
- En tercer lugar, se incluye la metodología, donde se detalla el método que es el camino sistemático a seguir para dar cumplimiento a los objetivos
- En cuarto lugar, se incluye los resultados obtenidos que se complementa con la discusión donde se pretende justificar e interpretar los resultados obtenidos
- Por último, se incluye las conclusiones donde se resumen los resultados relevantes obtenidos

3 Estado del Arte

A continuación, se presenta una fundamentación teórica de los factores de riesgo asociado a patologías radiográficas de columna lumbar, basado en una revisión de la literatura, investigaciones y proyectos similares relacionados

“La columna lumbar se ubica en la parte inferior de la columna, se conecta a la parte superior de la espalda llamada columna torácica por arriba y a la pelvis a través del sacro por debajo”(1). La columna lumbar se compone de 5 vértebras L1-L5, apiladas verticalmente con un disco intervertebral entre cada vértebra, articulaciones facetarías que permiten doblar la parte baja de la espalda en todas las direcciones, nervios que pasan por la parte de la columna lumbar, ligamentos y tendones.

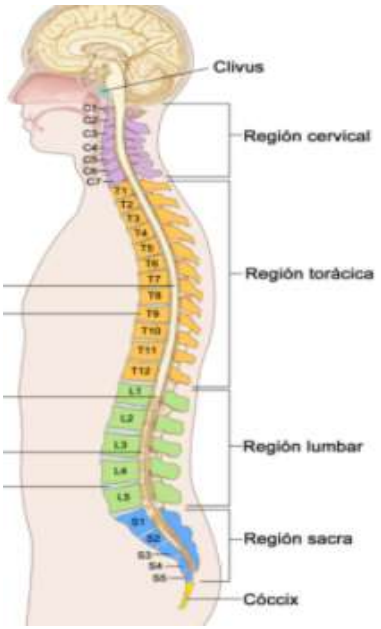


Figura 2. Anatomía de la columna vertebral
Fuente: Instituto Nacional del Cáncer. (6)

Figura 3. Anatomía de la columna lumbar
Fuente: Revista portales médicos (7)

El examen de columna lumbar pretende identificar las lesiones preexistentes que pudieran generar incapacidad o ausentismo laboral debido a la generación de dolor lumbar, sin embargo, es común encontrar personas con alguna anomalía radiográfica que no necesariamente son sintomáticos, citado por (4)

El dolor lumbar es la causa principal de vivir con discapacidad durante años en todo el mundo (3). El 75% a 80% de la población en general ha sufrido de dolor lumbar en alguna ocasión. En la mayoría de los casos, el dolor de espalda agudo es auto limitado y benigno, sin causa identificada en el 95%. En tales pacientes, la causa es una lesión muscular o ligamentosa, citado por (4)

“Los trastornos músculo esqueléticos son unos de los problemas más importantes de salud en el trabajo en países industrializados y en vías de desarrollo que afectan la calidad de vida de muchas personas; un gran número de trabajadores de distintos sectores sufren de molestias, que constituyen la enfermedad relacionada con el trabajo más común en Europa y suponen más del 45% de todas las enfermedades profesionales… En el mundo laboral existen múltiples motivos que conllevan a un deterioro a nivel general de la salud. Tal es el caso que la Organización Internacional de Trabajo (OIT) cada año reportan alrededor de 160 millones de casos nuevos de enfermedades profesionales no mortales, que causan enormes costos para los trabajadores y sus familias, así como para los países…Las lesiones músculo esqueléticas, de acuerdo con estadísticas proporcionadas actualmente constituyen la principal fuente de ausentismo laboral; de los datos extraídos en la revista de Riesgo del Trabajo del Ecuador (2013) el lumbago ocupó el 36% y, el síndrome del túnel carpo 40% ” (8)

Según la investigación titulada. *Patologías de columna lumbar diagnosticadas por radiografía convencional,* realizada en el Perú, se obtuvo los siguientes resultados, las patologías más frecuentemente halladas fueron la espondiloartrosis y las espondilosis, con predominio del sexo femenino. (9)

Dada la importancia de las enfermedades lumbares, hace necesario evaluar los factores de riego en pacientes con empleo, que son las “personas de 15 años y más que, durante la semana de referencia, se dedicaban a alguna actividad para producir bienes o prestar servicios a cambio de remuneración o beneficios” (10)

Patologías de la región lumbar

Tabla 1. Enfermedad de la región lumbar, según los datos obtenidos y CIE10 (código internacional de enfermedades)

CÓDIGO CIE10	NOMBRE	CÓDIGO CIE10	NOMBRE (específico)
M51.0	Trastornos de disco intervertebral dorsal, dorsolumbar y lumbosacro con mielopatía	M51.06	Trastornos de disco intervertebral con mielopatía, región lumbar
M51.1	Trastornos de disco intervertebral dorsal, dorsolumbar y lumbosacro con radiculopatía	M51.16	Trastornos de disco intervertebral con radiculopatía, región lumbar
M51.2	Otros tipos de desplazamiento de disco intervertebral dorsal, dorsolumbar y lumbosacro	M51.26	Otros desplazamientos de disco intervertebral, región lumbar
M51.3	Otros tipos de degeneración de disco intervertebral dorsal, dorsolumbar y lumbosacro	M51.36	Otros tipos de degeneración de disco intervertebral, región lumbar
M51.8	Otros tipos de trastorno de disco intervertebral dorsal, dorsolumbar y lumbosacro	M51.86	Otros trastornos de disco intervertebral, región lumbar
M51.9	Trastorno no especificado de disco intervertebral dorsal, dorsolumbar y lumbosacro		Según el diagnóstico ampliado se especifica hernia discal lumbar
M53	Otras dorsopatías y las no especificadas, no clasificadas bajo otro concepto	M53.2X6	Inestabilidades vertebrales, región lumbar
M54.1	Radiculopatía	M54. 16	Radiculopatía, región lumbar
M54.5	Dolor en la parte inferior de la espalda Lumbago no especificado		Según el diagnóstico ampliado se especifica lumbalgia

Elaborador por: El autor

Fuente: Tuotromedico.com (2021). (11)

Las filas no coloreadas no se han incluido en el análisis por tener una mínima frecuencia de ocurrencia

Diagnóstico

Para determinar una patología a nivel lumbar, es necesario tomar en cuenta los pasos del examen médico que según Guarderas son: la anamnesis, exploración física y los exámenes complementarios.

En la anamnesis se van a proceder a tomar información como: la edad, peso, talla, índice de masa corporal (IMC), sexo, cargo, antecedentes personales, familiares y sociales, cabe señalar que los antecedentes familiares no se estudian en la presente investigación

Índice de masa corporal (IMC): Es el cociente de dividir la masa en kilogramos para la estatura en metros al cuadrado. Según la OMS el IMC se categoriza de la siguiente manera:

Tabla 2. Índice de masa corporal

IMC	Estado
Por debajo de 18.5	Bajo peso
18,5–24,9	Peso normal
25.0–29.9	Pre-obesidad o Sobrepeso
30.0–34.9	Obesidad clase I
35,0–39,9	Obesidad clase II
Por encima de 40	Obesidad clase III, o Mórbida, o Extrema

Fuente: OMS (2020), (12)

También se procede a tomar información sobre los hábitos por ejemplo si el paciente consume o no alcohol, sustancias psicotrópicas, cigarrillo y si el paciente realiza o no actividad física.

Factores de riesgo que está expuesto el paciente en el trabajo: accidentes mayores, biológicos, ergonómicos, físico, mecánico, químico y psicosociales.

La Exploración física vertebral realizada por el especialista va a consistir en la inspección y palpación poniendo énfasis en la curvatura vertebral, flexión – extensión de la columna vertebral, dolor y lasegue.

Signo de Lasegue: con el paciente en decúbito dorsal, se realiza la maniobra de flexión de la cadera y elevación del miembro inferior con la rodilla extendida. Provoca dolor en la región glútea que se irradia por la cara posterior del miembro inferior explorado. El signo es positivo cuando el dolor aparece antes de que el miembro inferior llegue a los 40º de flexión. (13)

Figura 4. Signo de Lasegue
Fuente: Premian Madrid. (14)

Es necesario para apoyar al diagnóstico la realización de exámenes complementarios, dentro de las patologías lumbares son recomendables las radiografías.

Radiografías

"La radiografía simple sigue siendo una de las técnicas más usadas para realizar el diagnóstico, a pesar de la introducción de nuevas tecnologías, tales como la tomografía computarizada (TC) y la resonancia magnética (RM), han modificado sus indicaciones en la práctica médica diaria. Actualmente está bien establecido que la radiografía simple en ausencia de traumatismo tiene un valor limitado, ya que los cambios degenerativos son muy comunes y varias patologías más relevantes, de tipo tumoral o infecciosa, pueden pasar desapercibidas". (9)

El artículo "Signos radiológicos más frecuentes…", menciona, que la radiografía posee un buen rendimiento diagnóstico en lesiones traumáticas y enfermedades degenerativas y tumorales. Desafortunadamente, el rendimiento es bajo en la detección de enfermedades inflamatorias y discopatías, en donde los exámenes que dan una mejor información son el TC y la RM (4)

En el estudio titulado "Radiografía de columna lumbosacra…", se obtuvo los siguientes resultados, los hallazgos patológicos de la radiografía fue de 69,8 % y las radiografías restantes (30,2 %) fueron leídas como normales. (5)

Como el estudio inicial de imágenes se realiza la radiografía convencional, que, a diferencia de la tomografía o resonancia magnética, es más accesible por el paciente por los costos que implica, sin embargo, es necesario señalar que en varios casos es necesario complementarla con tomografías o resonancia magnética. *Cabe indicar que la presente investigación toma como referencia la información que puede proporcionar la radiografía convencional en la evaluación de la patología lumbar*

Factores de riesgo

"El dolor lumbar se puede clasificar como primario o secundario, con o sin compromiso neurológico; degenerativo mecánico; no mecánico; inflamatorio, infeccioso, metabólico, neoplásico o secundario a los efectos de enfermedades sistémicas. También hay un grupo importante de dolor lumbar no orgánico, que es extremadamente importante en un contexto ocupacional, Trabajar largas jornadas, realizar tareas pesadas, levantar carga, y problemas psicológicos … Los factores que han sido identificados que confieren riesgo de dolor lumbar ocupacional incluyen traumatismos acumulativos, actividades dinámicas relacionadas con movimientos de flexión y rotación del tronco, trabajo físico pesado, flexión o cuclillas, macro traumas, levantamiento o transporte de cargas, exposición a turnos prolongados sin pausas, vibraciones de cuerpo entero y posturas estáticas e inadecuadas." (4)

Según el artículo "Las enfermedades de la columna lumbar y su relación con el trabajo en España", El origen de las enfermedades de la columna lumbar se dice que es multifactorial, que incluye factores genéticos, degenerativos, bioquímicos, médicos, mecánicos, traumáticos y psicosociales…. Se consideran factores de riesgo laboral de dolor lumbar: traumatismos, manipulación de cargas, inclinaciones y giros, vibraciones, posturas forzadas, conducción, movimientos repetitivos, tabaquismo, obesidad, sedestación, debilidad muscular

y depresión (15). Además, el Ministerio de salud pública del Ecuador señala como factor de riesgo el incremento del índice de masa corporal. (13)

"Los factores desencadenantes de lumbalgias: factores físicos, organizativos y sociales en el lugar de trabajo, variables físicas y sociales ajenas al ámbito laboral y rasgos físicos y psicológicos de cada individuo, son complejos y están interrelacionados en la aparición de trastornos musculo esqueléticos…El dolor lumbar es más frecuente en grupo de trabajadores como obreros de la construcción, campesinos, enfermeras, trabajadores que manejan equipos pesados". (16)

En el estudio de Factores asociados a la enfermedad discal lumbar de origen laboral, se concluye señalando que "Existe asociación estadísticamente significativa entre el género y la exposición al factor de riesgo vibraciones/impacto. Las características de los factores de riesgos biomecánicos como la posición de la columna vertebral en flexión, la postura del cuerpo caminando durante la mayor parte de la jornada laboral, el levantar y depositar manualmente objetos, manipulación de carga mayor de 15 kg y la exposición a vibración de cuerpo entero mayor a 4 horas de la jornada laboral, y el tiempo de exposición laboral mayor a 1 años son elementos fundamentales que se deben tener en cuenta en el proceso de calificación del origen laboral de la enfermedad discal lumbar". (16)

4 Metodología

Los datos con los que se trabaja son obtenidos de las historias clínicas de pacientes ecuatorianos que tienen patologías en la región lumbar de la columna vertebral, cuentan con empleo y seguro privado de la empresa "Biodiment".

4.1 Tratamientos de datos

Se inicia con un tratamiento de los datos con el objetivo de visualizar en forma general, e identificar tipos de variables, categorías o re-categorizar a las variables, cambiar de nombres a las variables, entre otros aspectos importantes, que permita continuar y facilitar con análisis posteriores

La base de datos inicial obtenida consideraba una dimensión de 678 observaciones de pacientes atendidos por una sola vez durante los años 2018 al 2020 y 34 variables. A esta matriz, se le aplicó el tratamiento de datos adecuados, para obtener el conjunto definitivo a analizar. Los pasos del tratamiento fueron los siguientes:

1. Selección inicial de variables (filas=678, variables=24), se eliminan variables por ser no relevantes por su baja frecuencia o ninguna variación y/o dificultar el análisis del fenómeno de investigación, a continuación, se detalla las variables eliminadas:
   La identificación del paciente (ID_ Cliente), el examen de curvatura (EXM_FIS_CURVATURA) porque toda la muestra investigada presentaba un examen de curvatura normal, consumo de sustancias psicotrópicas (FRE_SUSTANCIA_PSICOTROPICA) porque el total de la muestra registra nunca haber consumido, por último, no se incluye los

peligros de los factores de riesgo sino únicamente los factores de riesgo: accidentes mayores, biológicos, ergonómicos, físicos, mecánicos, psicosociales y químicos

2. Ajuste, re-categorización de variables CARGO, EDAD y DEPORTE (filas=678, variables=24).
3. Filtro de la variable Patología (COD CIE10), se eliminaron de la base de datos, las patologías M512, M513, M519, M532 debido a su baja frecuencia de información, los pacientes que padecen son menos de 5 que representa menos del 0.6% (filas=669, variables=24).
4. Filtro de no fumadores, pues sólo había un caso registrado (filas=668, variables=24).
5. Ajuste, re-categorización de IMC_CATEGORICA (filas=668, variables=24).
6. Transformar a formato factor las columnas ubicadas en las posiciones de 1 a 5 y 10 a 24 (filas=668, variables=24).
7. Renombrar variables COD CIE10 por PATOLOGIA (filas=668, variables=24).
8. Transformar la variable PATOLOGIA a sólo 2 categorías, M545 / OTRA. (filas=668, variables=24).
   Cabe indicar que las dos categorías son: la más frecuente “M54.5” que hace referencia a la Lumbalgia o Lumbago y como “OTRA” las menos frecuentes como son M51.06, M51.16, M51.86 y M54.16

Previo este ajuste la base de datos con la que se cuenta para los posteriores análisis tiene 24 variables y una muestra de 668 datos que representa a los pacientes con alguna patología de la región lumbar.

Tabla3. Identificación de las variables de la base de datos (data) que se analizará

N°	VARIABLE	ETIQUETA DE LA VARIABLE	TIPO DE VARIABLE	ESCALA DE MEDIDA	NIVELES	VALORES FALTANTES
1	NUMERO	NUMERO DE IDENTIFICACIÓN				
2	AÑO	AÑO	CUALITATIVO	ORDINAL		0
3	PATOLOGÍA	PATOLOGÍA	CUALITATIVO	NOMINAL	2	0
4	CARGO	CARGO	CUALITATIVO	NOMINAL	3	0
5	SEXO	SEXO	CUALITATIVO	NOMINAL	2	0
6	EDAD	EDAD	CUANTITATIVO	RAZON		0
7	PESO	PESO	CUANTITATIVO	RAZON		0
8	ESTATURA	ESTATURA	CUANTITATIVO	RAZON		0
9	IMC	IMC	CUANTITATIVO	INTERVALO		0
10	IMC_CATEGORICA	IMC CATEGORICA	CUALITATIVO	ORDINAL	3	0
11	EDAD_CATEGORICA	EDAD CATEGORICA	CUALITATIVO	ORDINAL	5	0
12	EXM_FIS_DOLOR	EXAMEN FISICO: DOLOR	CUALITATIVO	NOMINAL	2	0
13	EXM_FIS_FLEXION	EXAMEN FISICO: FLEXION	CUALITATIVO	NOMINAL	2	0
14	EXM_FIS_LASEGUE	EXAMEN FISICO: LASEGUE	CUALITATIVO	NOMINAL	2	0
15	FRE_ALCOHOL	FRECUENCIA DE CONSUMO DE ALCOHOL	CUALITATIVO	ORDINAL	4	0
16	FRE_DEPORTE	FRECUENCIA DE REALIZAR DEPORTE	CUALITATIVO	ORDINAL	5	0
17	FRE_TABACO	FRECUENCIA DE CONSUMO DE TABACO	CUALITATIVO	ORDINAL	4	0
18	FACTOR_ACCIDENTES_ MAYORES	FACTORES DE RIESGO: ACCIDENTES MAYO	CUALITATIVO	NOMINAL	2	0
19	FACTOR_BIOLOGICOS	FACTORES DE RIESGO: BIOLOGICOS	CUALITATIVO	NOMINAL	2	0
20	FACTOR_ERGONOMICO	FACTORES DE RIESGO: ERGONOMICO	CUALITATIVO	NOMINAL	2	0
21	FACTOR_FISICOS	FACTORES DE RIESGO: FISICOS	CUALITATIVO	NOMINAL	2	0
22	FACTOR_MECANICOS	FACTORES DE RIESGO: MECANICOS	CUALITATIVO	NOMINAL	2	0
23	FACTOR_PSICOSOCIALES	FACTORES DE RIESGO: PSICOSOCIALES	CUALITATIVO	NOMINAL	2	0
24	FACTOR_QUIMICOS	FACTORES DE RIESGO: QUIMICOS	CUALITATIVO	NOMINAL	2	0

Se puede determinar que en su mayoría las variables son cualitativas nominales (columna 4) a excepción de las cuantitativas: edad, peso, estatura e IMC. Por grupos de colores se diferencian los factores: antecedentes personales (files 4 a 11), exámenes físicos (fila 12 a 14), hábitos (fila 15 a 17) y factores de riesgo en el trabajo (fila 18 a 24), y la patología como variable dependiente (fila 3).

4.2 Análisis exploratorio de datos (EDA)

Previo el ajuste de la base de datos y para demostrar los objetivos planteados se inicia con una visualización detallada de la base de datos, que permita describirlos, identificar posibles datos faltantes y atípicos y relaciones entre variables, esto se lo realiza con el análisis exploratorio de datos (EDA)

Analizar posibles datos faltantes y atípicos. - La detección de valores atípicos u outliers se puede analizar con técnicas simples como la estadística descriptiva, incluyendo mínimo, máximo, histograma, diagrama de caja y percentiles, utilizando el criterio de rango intercuartílico (IQR), para variables cuantitativas. (17)

Análisis exploratorio de datos (EDA). - Es necesario como uno de los puntos iniciales del procesamiento de datos realizar el EDA, con el objeto de identificar el comportamiento de las variables, evaluar objetivamente que variables aportar con el fenómeno investigado y la que no aportan con evidencias suficiente para incluirlas en el análisis, permite también identificar la relación que se puede tener entre variables

En la base de datos estudiada no hay datos faltantes y los datos atípicos observados se presenta en las variable IMC (relación entre peso y estatura) y edad, sin embargo, por la naturaleza de la investigación no amerita eliminarlos, ya que estos datos pueden influir en la presencia de las enfermedades de la región lumbar. (Ver *Figura 6*)

4.2.1 Análisis univariante y bivariante

4.2.1.1 Análisis de datos univariado

Se realiza un análisis univariado con las variables cuantitativas, calculando medidas descriptivas, y la prueba de normalidad, con un 5% de significancia. Mientras que con variables cualitativas se calcula frecuencias absolutas y relativas.

Es relevante señalar que la patología M54.5=Lumbalgia representa el 76.7%, a diferencia de las otras patologías menos frecuentes 23.35% (M51.86, M51.16, M54.16 y M51.06). (Ver *Tabla 1*)

- **Prueba de normalidad**

Existen varios test que permiten probar la normalidad como**:** Prueba de Chi-Cuadrado para bondad y ajuste, prueba de Shapiro-Wilks, prueba Kolmogorov-Smirnov, prueba Anderson-Darling, prueba de Jarque-Bera.

La prueba de Jarque-Bera utiliza un estadístico de prueba que involucra la curtosis y la asimetría y es adecuado cuando la muestra es mayor a 30. Mientras que Shapiro-Wilks en muestras menores a 30

- *Formulación de la hipótesis:*
  H0: Los datos se distribuyen normalmente
  H1: Los datos no se distribuyen normalmente

Con los datos analizados se evidencia que existe normalidad únicamente en la variable estatura (Ver *Tabla 5*)

4.2.1.2 Análisis de datos bivariados

Realizando un análisis bivariado se puede calcular varias pruebas de significancia que permiten enriquecer los resultados, como podemos ver más adelante

Se inicia calculando y analizando las correlaciones de Pearson, además a través de diagramas de caja se realiza el análisis de la variable cualitativa (Patología) con las variables cuantitativas: edad, peso estatura e IMC, y se puede observar los valores atípicos.

Se realiza la prueba de hipótesis de normalidad considerando como factor la patología con dos niveles (M545 y otro) y como variables respuesta cada una de las variables cuantitativas: edad, estatura, IMC y peso, aplicando el test de Shapiro-Wilks para n<30 y el test de Jarque-Bera para n>=30, con estas pruebas se determina que no existe normalidad (ver *Tabla 7*); razón por lo cual se debe aplicar pruebas no paramétricas como el test de Wilcoxon para probar la significancia de pares de medianas (niveles de patologías) por cada variable cuantitativa, a diferencia de la prueba t o Z que admiten normalidad

Prueba de medianas (Test de Wilcoxon)

Con la prueba de medianas se determina qué pares de patologías son iguales o diferentes, a través del test de Wilcoxon

Al no ser necesario asumir que las muestras se distribuyen de forma normal, se puede aplicar el test no paramétrico prueba de los rangos con signo de Wilcoxon, que permite comparar las medianas individuales

- *Formulación de la hipótesis:*
  Ho: Las medianas son iguales
  Ha: Las medianas no son iguales

El contraste de hipótesis lleva a validar que las patologías M5.45 y Otras, tienen relación con la mediana de las edades, peso, estatura e IMC de los pacientes. (Ver *Anexo 1*)

Pruebas de hipótesis con variables cualitativas

En el caso de las variables categóricas, cuando se trata de probar hipótesis utilizando pruebas de contraste se pueden utilizar: Chi cuadrada (X^2) de Pearson o Ji cuadrada, la prueba exacta de Fisher y la de Mcnemar

Se realiza la prueba de independencia con el test Chi cuadrado con la variable cualitativa (patología) vs las cualitativas, por cada par de variables, previo a utilizar el test Chi-cuadrado es necesario resaltar que se utilizará la función *(simulate.p.value)* en la fórmula de cálculo para corregir cualquier problema donde las frecuencias observadas sean pequeñas

- *Formulación de la hipótesis*.
  H0: Las variables son independientes por lo que una variable no varía entre los distintos niveles de la otra variable
  H1: Las variables son dependientes, una variable varía entre los distintos niveles de la otra variable.

De este contraste se evidencia que las variables que están relacionadas con la patología son: año, sexo, examen físico del dolor, factores mecánicos, factores psicosociales, factores químicos, el IMC categórico y la edad categórica. (Ver *Tabla 8*)

Se concluye el análisis bivariado realizando gráficos de barras apiladas de las variables significativamente asociadas con la patología (Ver *Figura 7*)

4.3 Modelos Predictivos

Se decide analizar modelos de respuesta binaria (Logit y Probit) y Árboles de clasificación (Rpart y C5.0), con el objetivo de comparar cuál es el modelo que da el mejor ajuste para predecir, previo lo trabajado anteriormente como es: el tratamiento de la base de datos, el análisis exploratorio univariante y bivariante y la aplicación de los test apropiados como Chi-cuadrado y Wilcoxon para diagnosticar las variables asociadas significativamente con la Patología, como son:

Tabla 4. Variables asociadas significativamente con la variable patología

N°	Variables	Tipo de variable	Niveles	Codificación
1	PATOLOGIA	Cualitativo	M545 OTRA	1 0
2	SEXO	Cualitativo	Masculino Femenino	1 0
3	EXM_FIS_DOLOR	Cualitativo	SI NO	1 0
4	FACTOR_MECAMICOS	Cualitativo	SI NO	1 0
5	FACTOR_PSICOSOCIALES	Cualitativo	SI NO	1 0
6	FACTOR_QUIMICOS	Cualitativo	SI NO	1 0
7	EDAD	Cuantitativo		
8	IMC	Cuantitativo		

Cabe señalar que la base de datos cuenta con 668 observaciones y 8 variables estadísticamente significativas relacionadas con las patologías.

Antes de construir los modelos, se realiza los siguientes pasos:

1. Se divide el conjunto de datos disponibles en:

- Un conjunto de datos de entrenamiento (training) que se utilizará para construir el modelo predictivo y
- Un conjunto de prueba (testing) que servirá para evaluar la eficacia del modelo y hacer predicciones correctas

Con el objetivo de evaluar la capacidad predictiva del modelo se selecciona aleatoriamente un 80% de observaciones como entrenamiento y el 20% para prueba. Previamente es necesario que las observaciones se seleccionen aleatoriamente, eligiendo una semilla

Al realizar la partición entre entrenamiento y test se visualiza que las proporciones de los niveles de la variable respuesta “PATOLOGIA” sean aproximadamente iguales

2. Desbalance y re muestreo. - Al tener un nivel mayor en una categoría (M54.5=410), provocará sesgar los resultados al nivel de la categoría mayor, lo cual repercutirá en la matriz de confusión, razón por la cual es necesario realizar un re muestreo y tener una base balanceada que permita clasificar los niveles correctamente

   El problema de desbalance se puede solucionar con: Rebalanceo de la muestra o asignar pesos o costos a las observaciones.

   La solución que se aplicó es el rebalanceo de la muestra que se puede realizar de dos formas: Sub muestreo o sobre muestreo, la que se aplicó es el sobre muestreo, replicando los casos, teniendo una dimensión de 820 observaciones y 8 variables

3. Aplicación del modelo. - La construcción del modelo se realiza a partir del conjunto de entrenamiento.
4. Evaluación del modelo. - La evaluación del modelo se realiza a partir del conjunto de prueba

4.3.1 Modelo Logit y Probit

Son modelos que permiten explicar los factores que determinan que un paciente con patología en la región lumbar presente la patología más frecuente Lumbago (M54.5) o las menos frecuente (Otra)

Son modelos no lineales que se utilizan cuando la variable dependiente que es la Patología es binaria o dummy, es decir que toma dos valores, 1=M54.5 (patología más frecuente "Lumbago"), 0= Otro (Patologías menos frecuentes: M51.06, M51.16, M51.86, M54.16) y las variables independientes pueden ser cuantitativas o cualitativas

La regresión logística es llamada también modelo Logit y la regresión Probit es llamada también modelo Probit. En ambos casos, Logit y Probit, el modelo es no lineal y, dadas las observaciones de la variable Y y del vector x, se utiliza técnicas de estimación por máxima verosimilitud, junto con algoritmos de optimización numérica, para estimar los parámetros β del modelo. La técnica de máxima verosimilitud que utiliza la función de distribución para la variable y es logística en el modelo Logit, mientras que es normal para Probit, si se dispone de una muestra de observaciones independientes

El modelo Logit pueden expresarse con la siguiente función:

$$\Pr(Z) = \frac{e^Z}{1 + e^Z}$$

Dónde: La probabilidad de éxito se evalúa en la función Pr(Z), es la función de distribución acumulada logística estándar, donde el vector Z corresponde al predictor lineal del modelo.

EL modelo específicamente será:

$$p(X) = \frac{e^{\beta_0+\beta_{1(sexo)}+\beta_{2(dolor)}+\beta_{3(factor\ mecanico)}+\beta_{4(factor\ psicosocial)}+\beta_{5(edad)}+\beta_{6(IMC)}}}{1 + e^{\beta_0+\beta_{1(sexo)}+\beta_{2(dolor)}+\beta_{3(factor\ mecanico)}+\beta_{4(factor\ psicosocial)}+\beta_{5(edad)}+\beta_{6(IMC)}}}$$

Dónde: p(X), predice la probabilidad de que un paciente tenga patología (M545), dada la información que tenemos (variables x)

El modelo Probit pueden expresarse con la siguiente función:

$$P(y = 1|x) = \Phi\left(X^{\hat{T}}\beta\right),$$

Dónde: P es la probabilidad, Φla función de distribución acumulada, β son los parámetros, $X^{\hat{T}}$ transpuesta del vector de regresores (variable x)

El modelo específicamente será:

$$P(y = 1|x) = \Phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6)$$

Aplicación del Modelo Logit y Probit

Cabe indicar que los diagnósticos realizados para el modelo Logit y Probit son similares

Se estima el modelo Logit utilizando la función *glm del programa R (modelo lineal generalizado).* (18) En el primer modelo denominado lg1, la única variable que no es significativo es: FACTOR_QUIMICOSSI, de manera que aplicando el método backward se obtiene el modelo lg2 donde todas las variables son significativas. (Ver *Tabla 9 y 10, apartado de resultados*)

Métodos Backward o Eliminación hacia atrás, es uno de varios métodos que permite seleccionar las variables a retener en el modelo, se caracteriza porque se empieza elimina la variable menos influyente, hasta que todas son significativas.

Validación del modelo:

- **Significancia de los estimadores**

Si el estimador es < 0.05 esa variable independiente explica la variable dependiente. Los signos de los estimadores indican la dirección de la relación

- **Evaluar la multicolinealidad**

La multicolinealidad se presenta cuando hay una fuerte correlación entre variables explicativas del modelo, se puede identificar que existe multicolinealidad a través de la matriz de correlaciones, de pruebas estadísticas o a través del factor de inflación de la varianza (VIF). Según Montgomery 2006, si el VIF es mayor que 5 o 10 es indicio que existe multicolinealidad En el modelo no hay indicios de multicolinealidad, ver *Tabla 11, apartado de resultados*

- **Bondad de ajuste**

Los modelos Logit y Probit al ser no lineales no tienen sentido evaluar el coeficiente de determinación R^2, existen criterios alternativos una medida es el pseudo R^2 de Mc Fadden

$$R^2 = 1 - \frac{\ln(LM)}{\ln(Lo)}$$

Dónde: ln(LM), es el logaritmo de la probabilidad para el modelo ajustado; ln(Lo), logaritmo del valor de la función de verosimilitud para un modelo sin predictores o modelo nulo o base, es similar a la suma de cuadrados residuales en la regresión lineal. (19)

A mayor capacidad explicativa del modelo el pseudo R^2 estará próximo a 1

- **Curva de ROC (Curva de característica operativa del recepto)**

Se evaluará a partir del conjunto de prueba (test), es un gráfico que muestra el rendimiento de un modelo. El AUC es el área bajo la curva de ROC, miden qué tan bien se clasifican las predicciones. Un modelo cuyas predicciones son un 100% correctas tienen un AUC=1

- **Matriz de confusión**

Permite visualizar el desempeño del modelo, permite analizar los aciertos y errores que tiene el modelo, en resumen, permite evaluar la exactitud y la precisión, la sensibilidad y la especificidad del modelo.

La sensibilidad y especificidad evalúa la capacidad que tiene el estimador para discriminar los casos positivos de los negativos; la sensibilidad es la fracción de verdaderos positivos y la especificidad es la fracción de verdaderos negativos

4.3.2 Modelo Árbol de clasificación

Los árboles de decisión son modelos de predicción, existe varias maneras de obtener los árboles de decisión las que usaremos es CART=Classification and Regression Trees, es una técnica que permite tener árboles de clasificación y regresión, el que aplicaremos es el árbol de clasificación porque permite predecir una respuesta cualitativa que en la presente investigación es la patología con dos niveles M5.45 y Otra; se va a trabajar con los siguientes paquetes:

- Árboles de clasificación, usando el paquete *rpart*. (20)
- Árboles de clasificación, usando el paquete C*5.0* (21)

4.3.2.1 Árboles de clasificación, usando el paquete *rpart*

“La implementación particular de CART que usaremos es conocida como Recursive Partitioning and Regression Trees (Árboles de regresión y particionamiento recursivos) o RPART**.**

De manera general, lo que hace este algoritmo es encontrar la variable independiente que mejor separa nuestros datos en grupos, que corresponden con las categorías de la variable objetivo. Esta mejor separación es expresada con una regla. A cada regla corresponde un nodo, hasta llegar a un nodo terminal u hoja y es cuando el algoritmo se detiene, una característica es que cuando una variable es elegida para separar los datos no es utilizada por otra

ocasión. Cabe indicar que lo que ocurre entre grupos es independiente entre sí y las reglas que se aplican no afectan en nada a los demás". (20)

Una de las principales limitaciones del modelo es que es un tipo de clasificación "débil", pues sus resultados pueden variar mucho dependiendo de la muestra de datos usados para entrenar un modelo

- **Aplicación del modelo**

Para implementar el modelo, se trabaja con la data de entreneamiento, se utiliza la función rpart y se elige el método class, se excluye la variable independiente factores de riesgo químico que como vimos anteriormente no fue significativo y para que sea comparable con todos los modelos planteado

El modelo muestra el esquema del árbol de clasificación implementado, (ver *Anexo 5*), cada inciso nos indica un nodo junto con la regla de clasificación que le corresponde, siguiendo los nodos se llega a las hojas del árbol que corresponde a la clasificación de los datos, esto se visualiza mejor en la gráfica del modelo que se visualiza en los resultados (Ver *Figura 10 o Anexo 6*)

- **Evaluación del modelo**

Realizamos las predicciones con la base de datos de prueba (test). Se genera la curva de ROC y la matriz de confusión cruzando las predicciones con los datos de prueba

4.3.2.2 Árboles de clasificación, usando el paquete C*5.0*

Se procede con el modelo de clasificación en un árbol C5.0, publicado por John Ross Quinlan (1992), este algoritmo crea árboles de decisión simples, es un modelo basado en reglas, es accesible mediante el paquete C5.0. Entre las particularidades que tiene este algoritmo son:

- Utiliza la entropía como medida de pureza para dividir los árboles
- El método de podado se realiza por defecto
- El algoritmo de boosting se detiene si la incorporación de nuevos modelos no aporta mejoras
- Permite asignar diferentes pesos a cada tipo de error. (21)

- **Aplicación del modelo**

Para implementar el modelo se utiliza el algoritmo C5.0 con la base de entrenamiento a excepción de la variable independiente factores de riesgo químico que como vimos anteriormente no era significativo y para que sea comparable con todos los modelos planteado

- **Evaluación del modelo**

La capacidad predictiva del modelo se evalúa con el conjunto de prueba. Se evalúa la curva de ROC (Ver *Figura 12, apartado de resultados*) y la matriz de confusión (*Anexo 9*).

Finalmente se procede a realizar una comparación de los resultados obtenidos con los cuatro modelos predictivos de respuesta binaria: Logit, Probit, Árboles de

clasificación Part y C5.0, cabe señalar que los modelos utilizan diferentes algoritmos y metodologías, sin embargo, únicamente se pretende probar cual tiene mejor ajuste y predecir en base a ese modelo, dando mejores resultados el modelo Logit.

5 Resultados

Previo el tratamiento que se hace con la base de datos inicial, detallado en el primer apartado de la metodología, se procede a realizar el análisis exploratorio de datos

5.1 Análisis exploratorio de datos (EDA)

Análisis de datos univariado de variables cuantitativas

El análisis exploratorio de datos se inicia con una descripción univariante de las variables cuantitativas, como: media, desviación típica, mediana, mínimo, máximo, rango, sesgo, curtosis, el valor p de la prueba de normalidad y el coeficiente de variación

Tabla 5. Medidas descriptivas y prueba de normalidad de variables cuantitativas

Variables	n	mean	sd	median	min	max	range	skew	kurt	valor.p	c.var
EDAD	668	36.92	9.00	36.00	19.00	66.00	47.00	0.6	-0.08	0.00	24.38
PESO	668	72.53	13.57	71.25	43.00	138.60	95.60	0.9	2.04	0.00	18.70
ESTATURA	668	1.64	0.09	1.64	1.41	1.92	0.51	0.2	-0.17	0.08	5.28
IMC	668	26.77	3.92	26.34	17.22	46.85	29.63	1.0	2.67	0.00	14.63

La edad, peso, estatura e IMC tienen sesgo positivo es decir tiene cola más larga a derecha los valores están sobre la media, con respecto a la curtosis la edad y la estatura tienen un aplanamiento son platicúrtica, mientras el peso e IMC tienen un apuntamiento son leptocúrticas, en general la variabilidad es baja en todas las variables cuantitativas, sin embargo, los datos de estatura están más concentrados alrededor de su media 1.64 cm, a diferencia de la edad que presenta mayor variación, registrando una edad mínima de 19 años y una máxima de 66 años.

Existe evidencia significativa que la estatura se distribuye normalmente considerando un 5% de significancia, según el test de Jarque-Bera, mientras que edad, peso e IMC no siguen una distribución normal.

Análisis de datos univariado de variables cualitativas

Se interpreta las frecuencias absolutas y relativas de las variables cualitativas, para tener un entendimiento básico de los datos

Tabla 6. Distribución de frecuencias de variables cualitativas

Variable	Nivel	Frecuencia	Porcentaje
AÑO	2018	408	61.08%
AÑO	2019	186	27.84%
AÑO	2020	74	11.08%
PATOLOGIA	OTRA	156	23,35%
PATOLOGIA	M545	512	76.65%
SEXO	F	255	38.17%
SEXO	M	413	61.83%
CARGO	A	180	26.95%
CARGO	B	267	39.97%
CARGO	C	221	33.08%
EXM_FIS_DOLOR	NO	349	52.25%
EXM_FIS_DOLOR	SI	319	47.75%
EXM_FIS_FLEXION	ANORMAL	22	3.29%
EXM_FIS_FLEXION	NORMAL	646	96.71%
EXM_FIS_LASEGUE	NO	625	93.56%
EXM_FIS_LASEGUE	SI	43	6.44%
FRE_ALCOHOL	DIARIO	270	40.42%
FRE_ALCOHOL	MENSUAL	39	5.84%
FRE_ALCOHOL	QUINCENAL	346	51.8%
FRE_ALCOHOL	SEMANAL	13	1.95%
FRE_DEPORTE	A lo menos 1 vez al mes	611	91.47%
FRE_DEPORTE	Nunca	57	8,53%
FRE_TABACO	DIARIO	561	83.98%
FRE_TABACO	MENSUAL	11	1.65%
FRE_TABACO	QUINCENAL	69	10.33%
FRE_TABACO	SEMANAL	27	4.04%
FACTOR_ACCIDENTES_MAYORES	NO	614	91.92%
FACTOR_ACCIDENTES_MAYORES	SI	54	8.08%
FACTOR_BIOLOGICOS	NO	604	90.42%
FACTOR_BIOLOGICOS	SI	64	9.58%
FACTOR_ERGONOMICO	NO	70	10.48%
FACTOR_ERGONOMICO	SI	598	89.52%
FACTOR_FISICOS	NO	354	52.99%
FACTOR_FISICOS	SI	314	47.01%
FACTOR_MECANICOS	NO	258	38.62%
FACTOR_MECANICOS	SI	410	61.38%
FACTOR_PSICOSOCIALES	NO	438	65.57%
FACTOR_PSICOSOCIALES	SI	230	34.43%
FACTOR_QUIMICOS	NO	545	81.59%
FACTOR_QUIMICOS	SI	123	18.41%
IMC_CATEGORICA	Normal	229	34.28%
IMC_CATEGORICA	Obesidad	121	18.11%
IMC_CATEGORICA	Sobre peso	318	47.6%

EDAD_CATEGORICA	Entre 31 y 40 años	275	41.17%
EDAD_CATEGORICA	Entre 41 y 50 años	147	22.01%
EDAD_CATEGORICA	Entre 51 y 60 años	50	7.49%
EDAD_CATEGORICA	Mayores de 60 años	10	1.5%
EDAD_CATEGORICA	Menores de 31 años	186	27.84%

En el año 2018 se registró mayormente pacientes con patologías de la región lumbar, disminuyendo cada año hasta registrar un 11.1% en el 2020, hay que destacar que en este año se presentó un fenómeno particular que fue la pandemia.

La mayoría de pacientes el 76.7% presentan la patología codificada como M5.45 que representa la Lumbalgia y Otras patologías recoge el 23.35%, cabe indicar que esto corresponde a las patologías M51.86 que representa el 7.2%, M51.16 el 6.9%, M54.16 el 5.1% y M51.06 el 4.2%. (Ver *Tabla 1*)

Con respecto al sexo el 61.8% son hombres y el 38.2% mujeres. El cargo tiene 3 categorías, A con el 27% corresponde al personal de administración, contabilidad dirección e informática, B con el 40% corresponde al personal de producción y servicios generales y C con el 33.1% incluye a los profesionales y técnicos y ventas.

En lo que respecta a los exámenes físicos, el 96,7% de pacientes presenta normal el examen de flexión y el 93.6 su examen de lasegue, con respecto al examen físico de dolor se registra un 47,8% de pacientes que en algún momento han presentado dolor en la región lumbar.

Con respecto a hábitos, el consumo de alcohol y tabaco al menos una vez al mes es del 100% en los pacientes con empleo sin distinción del sexo, el 91.5% realiza deportes al menos una vez al mes.

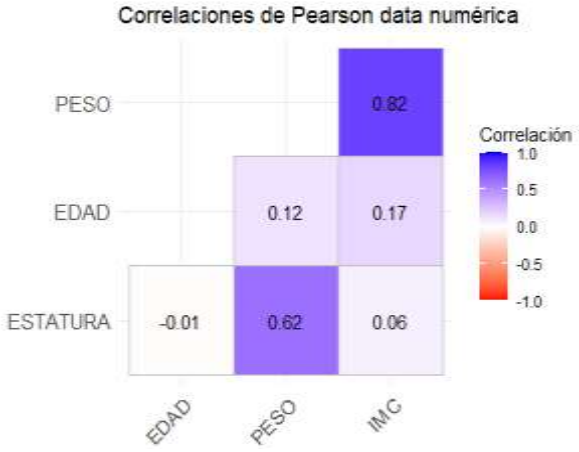
Uno de los riesgos mayores que los pacientes están expuestos en sus trabajos es el riesgo ergonómico 89.5%, seguido de mecánicos 61.4%, físicos 47%, psicosocial 34.4%, químico 18.4%, biológicos 9.6% y finalmente accidentes mayores 8.1%

El 65,7% de pacientes tienen un IMC (relación entre el peso y la estatura) sobre lo normal, sin embargo, hay que destacar que la mayoría de pacientes hace deporte, de manera que el índice de masa muscular aumenta y este factor no se incluye en la fórmula de cálculo del IMC, además la correlación entre el peso y la estatura es del 62%, como se puede apreciar en la *Figura 5*

Análisis de datos bivariado de variables cuantitativas

Se analiza las relaciones entre variables cuantitativas: peso, edad estatura e IMC

Figura 5. Correlaciones entre variables cuantitativas

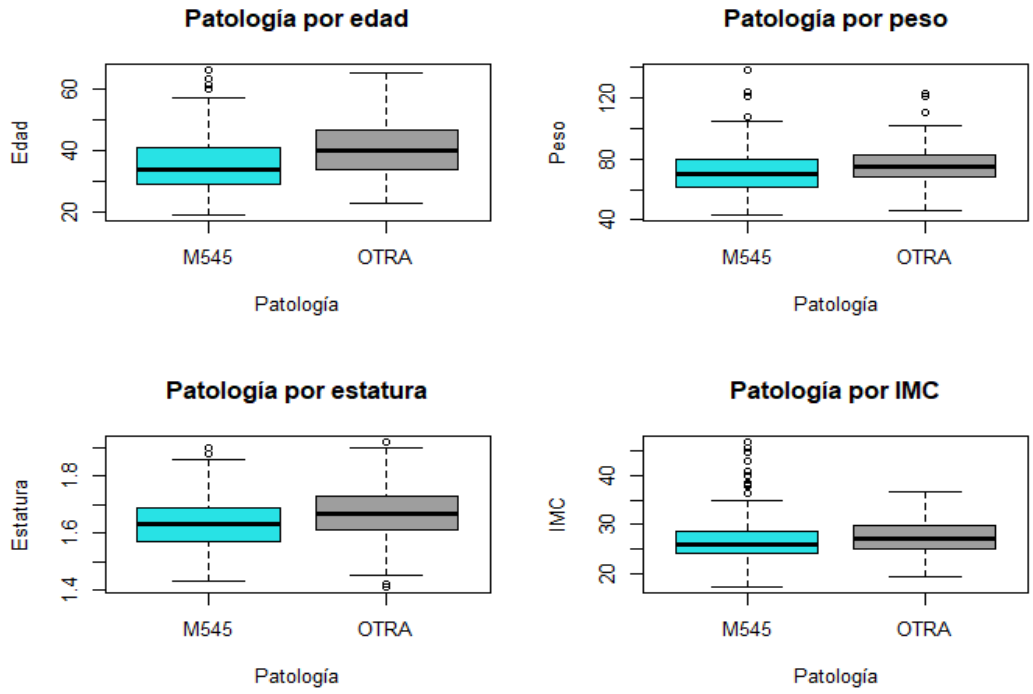


Hay que destacar que las variables altamente correlacionadas positivamente son: peso e IMC con el 82% y el peso con la estatura 62%

Análisis de datos bivariado de variable cualitativa (patología) vs cuantitativas

Se identifica a través del gráfico de caja los valores atípicos de las variables cuantitativas según las patologías

Figura 6. Diagrama de caja de la variable cualitativa (patología) vs cuantitativas



Los puntos que salen fuera de los bigotes en el diagrama de caja, se consideran como atípicos, se observa que existen datos atípicos por encima de

la media, en la edad, la estatura, el peso y el IMC de los pacientes medidos según la patología, se observa que los datos atípicos se registran generalmente en la patología M54.5 que representa a Lumbalgia y es la más frecuente en los pacientes

Para la presente investigación no se considera necesario eliminar o tratar a los datos atípicos, ya que el IMC que es la relación entre el peso y la talla, puede ser influyente en las enfermedades de la región lumbar, al igual que la edad.

Prueba de normalidad

Se aplica el test de Shapiro-Wilks para n<30 y el test de Jarque-Bera para n ≥ 30, para probar normalidad en los niveles de patologías considerados como factores y la variable respuesta la edad, estatura, IMC y peso

Tabla 7. Prueba de normalidad por niveles de la variable patología

VARIABLE	PATOLOGIA	n	VALOR P	DIAGNÓSTICO
EDAD	M545	512	0	NO NORMALIDAD
EDAD	OTRA	156	0.0373	NO NORMALIDAD
ESTATURA	M545	512	0.0337	NO NORMALIDAD
ESTATURA	OTRA	156	0.860	NORMALIDAD
IMC	M545	512	0	NO NORMALIDAD
IMC	OTRA	156	0.214	NORMALIDAD
PESO	M545	512	0	NO NORMALIDAD
PESO	OTRA	156	0	NO NORMALIDAD

Al ser al menos un nivel del factor Patología no normal en las variables cuantitativas: edad, estatura, IMC y peso, entonces no se puede aplicar pruebas paramétricas, pero si el test de Wilcoxon

Prueba de pares de medianas

Al no cumplir el supuesto de normalidad, se realiza la prueba de medianas con el test de Wilcoxon, con cada variable respuesta: edad, peso, estatura e IMC y como factor la patología con 2 niveles. (Ver *Anexo 1*)

- Edad: al ser el valor p= 8.923e-09 < 0.05 se evidencia que existe diferencia significativa en la mediana de la edad de los pacientes debido a las patologías.
- Peso: al ser el valor p= 3.088e-06 < 0.05 se evidencia que existe diferencia significativa en la mediana de peso de los pacientes debido a las patologías.
- Estatura: al ser el valor p= 1.837e-05 < 0.05 se evidencia que existe diferencia significativa en la mediana de la estatura de los pacientes debido a las patologías.

- IMC: al ser el valor 0.002364 < 0.05 se evidencia que existe diferencia significativa en la mediana del IMC de los pacientes debido a las patologías

Se concluye que las patologías M5.45 y Otras, tienen relación con la mediana de las edades, peso, estatura e IMC de los pacientes

Variable cualitativa (patología) vs cualitativas

Prueba de independencia

Se realiza la prueba de independencia con tablas de contingencia con variables cualitativas a través del test Chi cuadrado. Es necesario resaltar que uno de los requisitos para aplicar esta prueba es que las frecuencias observadas no deben ser pequeñas, sin embargo, para solucionar este problema de ser necesario utilizaremos la función (*simulate.p.value*) que corrige este problema. En el *Anexo 2*, se puede visualizar las tablas de contingencia de frecuencias observadas y esperadas

Tabla 8. Prueba de independencia con el test Chi cuadrado

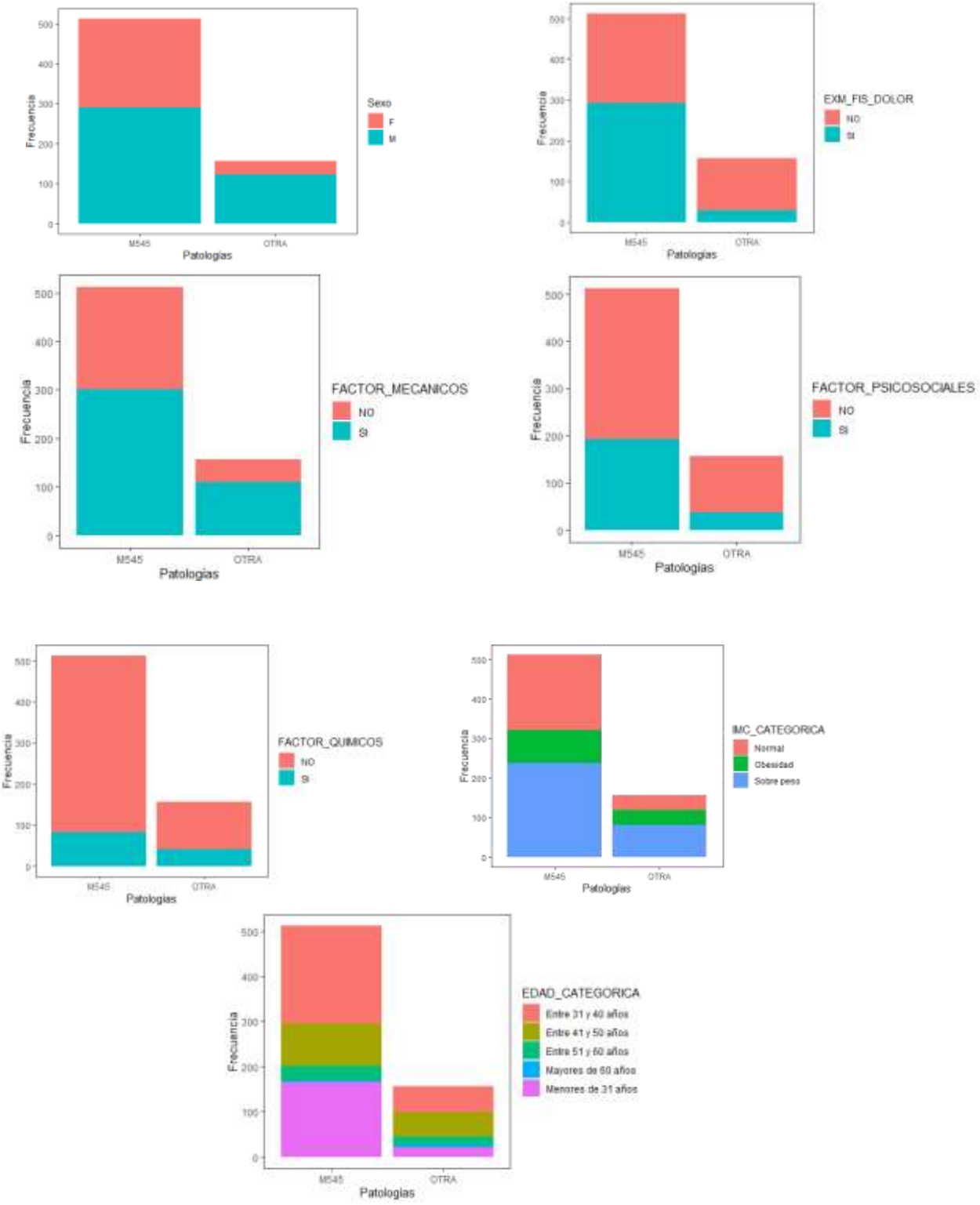
Variable	Valor P	Diagnóstico	
AÑO	0.00600	Se rechaza Ho	(Significativa)
CARGO	0.657	No se rechaza	(NO significativa)
SEXO	0.000500	Se rechaza Ho	(Significativa)
EXM_FIS_DOLOR	0.000500	Se rechaza Ho	(Significativa)
EXM_FIS_FLEXION	0.321	No se rechaza	(NO significativa)
EXM_FIS_LASEGUE	0.0940	No se rechaza	(NO significativa)
FRE_ALCOHOL	0.676	No se rechaza	(NO significativa)
FRE_DEPORTE	0.0555	No se rechaza	(NO significativa)
FRE_TABACO	0.275	No se rechaza	(NO significativa)
FACTOR_ACCIDENTES_MAYORES	0.751	No se rechaza	(NO significativa)
FACTOR_BIOLOGICOS	0.0660	No se rechaza	(NO significativa)
FACTOR_ERGONOMICO	0.132	No se rechaza	(NO significativa)
FACTOR_FISICOS	0.154	No se rechaza	(NO significativa)
FACTOR_MECANICOS	0.00700	Se rechaza Ho	(Significativa)
FACTOR_PSICOSOCIALES	0.00250	Se rechaza Ho	(Significativa)
FACTOR_QUIMICOS	0.00400	Se rechaza Ho	(Significativa)
IMC_CATEGORICA	0.00200	Se rechaza Ho	(Significativa)
EDAD_CATEGORICA	0.000500	Se rechaza Ho	(Significativa)

- Las variables: año y sexo revela dependencia con las patologías
- El examen físico de dolor revela dependencia con las patologías
- Los factores de riego que revelan dependencia con las patologías son: mecánicos, psicosociales y químicos
- El IMC (categórica) y edad (categórica) revela dependencia con las patologías

Análisis gráfico de variable dependientes cualitativa (patología) vs cualitativas

Se analiza el gráfico de barras de las variables cualitativas asociadas significativamente con la patología

Figura 7. Gráfico de barras apilados de la variable patología con las variables cualitativas relacionadas



Se puede apreciar que la patología más frecuente (M54.5) no existe mayor diferencia entre hombres y mujeres, mientras en otras patologías se presenta mayormente en hombres, la presencia de dolor es mayormente en la patología M54.5 que, en otras, además los factores químicos que están expuestos es menor comparado con factores mecánica y psicosociales. Con las dos categorías de patología se visualiza que existen más pacientes que tienen un IMC sobre lo normal, la mayoría de pacientes tienen entre 31 a 40 años con la patología M5.45, mientras que con la patología otra la mayoría de pacientes tiene 41 a 50 años de edad.

5.2 Modelos Predictivos

Previo a realizar el tratamiento de datos, un análisis univariante y bivariante, y validando a través de las pruebas no paramétricas adecuadas que permitieron diagnosticar las variables asociadas significativas con la patología, se cuenta con una base de datos con 668 observaciones y 8 variables, donde la variable dependiente es la patología y las independientes: sexo, examen físico de dolor, factores mecánicos, factores psicosociales, factores químicos, edad e IMC, como se detalla en la *Tabla 4*.

Con todo lo trabajado hasta el momento y para demostrar los objetivos planteados en la investigación se decide evaluar los modelos predictivos de respuesta binaria: Logit y Probit; y Árboles de clasificación usando el paquete Rpart y C5.0.

Cabe señalar que previo a construir los modelos se selecciona aleatoriamente un 80% de observaciones como un conjunto de datos de entrenamiento (training) que se utilizará para construir el modelo predictivo y un 20% de observaciones como un conjunto de prueba (testing) que servirá para evaluar la eficacia del modelo y hacer predicciones correctas

De las 534 observaciones del conjunto de datos de entrenamiento se observa que se tiene un nivel mayor en una categoría lo cual afecta a los resultados posteriores, razón por lo cual se procede a realizar un re muestreo con el objetivo de tener una base balanceada que permita clasificar los niveles correctamente, aplicando específicamente un sobre muestreo para balancear la base de datos

Solo después de realizar los puntos anteriores se construyen los modelos a partir del conjunto de entrenamiento.

5.2.1 Modelo Logit

Aplicación y validación del Modelo Logit

- **Significancia de los estimadores**

La siguiente tabla muestra los coeficientes, sus errores estándar, el estadístico Z y los valores de p Asociados

Tabla 9. Primer modelo Logit (lg1)

Modelo Logit (1)					
Coefficients:	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	4,15552	0,672885	6,176	6,59E-10	***
SEXOM	-1,05222	0,187368	-5,616	1,96E-08	***
EXM_FIS_DOLORSI	1,97373	0,183081	10,781	< 2e-16	***
FACTOR_MECANICOSSI	-0,390524	0,181658	-2,150	0,03157	*
FACTOR_PSICOSOCIALESSI	0,794626	0,188536	4,215	2,50E-05	***
FACTOR_QUIMICOSSI	-0,073914	0,224253	-0,330	0,7417	
EDAD	-0,05711	0,009592	-5,954	2,61E-09	***
IMC	-0,072077	0,022435	-3,213	0,00132	**

Nota: "*" significancia del 0,05; "**" significancia del 0,01; "***" significancia del 0,001; " " no hay significancia

Tabla 10. Segundo modelo Logit (lg2)

Modelo Logit (2)					
Coefficients:	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	4,15303	0,67231	6,177	6,52E-10	***
SEXOM	-1,06511	0,18332	-5,810	6,24E-09	***
EXM_FIS_DOLORSI	1,97872	0,18252	10,841	< 2e-16	***
FACTOR_MECANICOSSI	-0,40388	0,17722	-2,279	0,02267	*
FACTOR_PSICOSOCIALESSI	0,79216	0,18832	4,206	2,59E-05	***
EDAD	-0,05687	0,00955	-5,955	2,60E-09	***
IMC	-0,07223	0,02245	-3,218	0,00129	**

Nota: "*" significancia del 0,05; "**" significancia del 0,01; "***" significancia del 0,001

En el primer modelo Logit la única variable que no es significativa es el factor químico, aplicando el método Backward o eliminación hacia atrás se empieza eliminando la variable menos significativa y se obtiene el segundo modelo donde todas las variables individualmente son significativas al 5%, además los signos de los coeficientes tienen sentido lógico, más adelante lo detallamos

- **Evaluación de la multicolinealidad**

Tabla 11. Evaluación del factor de inflación de la varianza (VIF). Modelo Logit

Vif(lg2)					
SEXO	**EXM_FIS_ DOLOR**	**FACTOR_ME CANICOS**	**FACTOR_PSICO SOCIALES**	**EDAD**	**IMC**
1,025960	1,106178	1,058588	1,034774	1,098243	1,081351

De la salida anterior vemos que los VIF de todas las variables, no son grandes, eso indica que no hay un problema de multicolinealidad entre esas variables, ninguna variable independiente está correlacionada

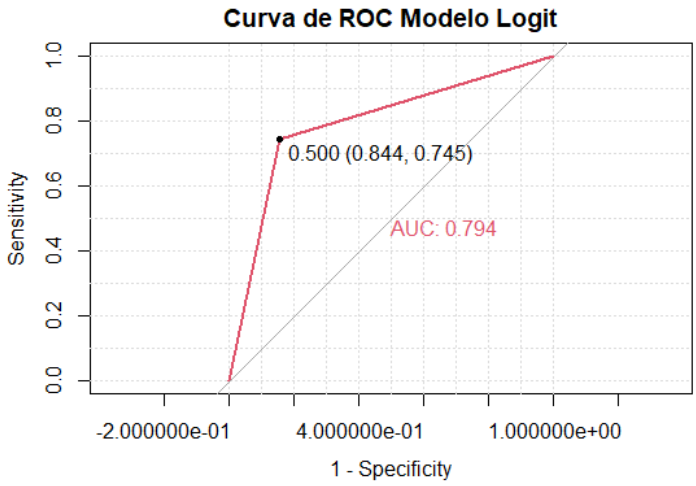
- **Bondad de ajuste del modelo**

Una medida que evalúa la bondad de ajuste del modelo es el pseudo R^2 de Mc Fadden que es 0.2326 y el valor ajustado 0.2203, lo cual quiere decir que el 22% de la patología está siendo explicado por las variables independientes incluidas en el modelo, sin embargo, es necesario evaluar algunos otros indicadores

- **Curva de ROC**

La Curva de ROC, se evaluará a partir del conjunto de prueba

Figura 8. Curva de ROC del Modelo Logit



Esta curva muestra por un lado la sensibilidad y por el otro 1 – el valor de la especificidad, si la curva se encuentra próxima a la línea de 45° diremos que nuestro modelo no discrimina nada, si por el contrario más alejado está de la línea el modelo discrimina mejor es decir si el área bajo la curva ROC es próximo a 1 discrimina bien, en nuestro gráfico el área bajo la curva ROC es de 0.794 discrimina correctamente y es un modelo aceptable, indica un buen valor diagnóstico, un buen ajuste del modelo, es decir las predicciones se clasifican bien

- **Matriz de confusión**

Permite evaluar la predicción (Ver *Anexo 3*), lo que ha clasificado el modelo frente a lo que realmente es, así tenemos 27 pacientes que el modelo ha clasificado como otras patologías cuando realmente formaba parte de ella, mientras que 26 pacientes que el modelo ha clasificado como otras patologías cuando realmente no formaban parte de ella, es decir lo que está dentro de la diagonal principal es lo que está correctamente clasificado, lo que está fuera de la diagonal principal es lo que no predijo el modelo correctamente.

Sensibilidad hace referencia al porcentaje de pacientes que se estima que forma parte de los pacientes con patología M54.5 cuando realmente forman parte de ella, equivale al 84%

Especificidad hace referencia al porcentaje de pacientes que se estimó que no formaban parte de los pacientes con patología M54.5 y realmente no formaban parte de ella, equivale al 75%

Una medida general de ver si el modelo está correctamente clasificado (Accuracy) es del 77% y un error del 23%

5.2.2 Modelo Probit

Aplicación y validación del Modelo Probit

- **Significancia de los estimadores**

La siguiente tabla muestra los coeficientes, sus errores estándar, el estadístico Z y los valores de p asociados

Tabla 12. Primer modelo Probit (pr1)

Modelo Probit (pr1)					
Coefficients:	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	2,493892	0,389865	6,397	1,59E-10	***
SEXOM	-0,621323	0,109669	-5,665	1,47E-08	***
EXM_FIS_DOLORSI	1,183510	0,105346	11,235	< 2e-16	***
FACTOR_MECANICOSSI	-0,251198	0,107787	-2,330	0,01978	*
FACTOR_PSICOSOCIALESSI	0,463794	0,110908	4,182	2,89E-05	***
FACTOR_QUIMICOSSI	-0,04713	0,131235	-0,359	0,7195	
EDAD	-0,034482	0,005574	-6,187	6,15E-10	***
IMC	-0,042696	0,013189	-3,237	0,00121	**

Nota: "*" significancia del 0,05; "**" significancia del 0,01; "***" significancia del 0,001; " " no hay significancia

Tabla 13. Segundo modelo Probit (pr2)

Modelo Probit (pr2)					
Coefficients:	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	2,490624	0,389426	6,396	1,60E-10	***
SEXOM	-0,629042	0,107441	-5,855	4,78E-09	***
EXM_FIS_DOLORSI	1,187061	0,10492	11,314	< 2e-16	***
FACTOR_MECANICOSSI	-0,259953	0,105172	-2,472	0,0134	*
FACTOR_PSICOSOCIALESSI	0,462568	0,110822	4,174	2,99E-05	***
EDAD	-0,034342	0,005556	-6,181	6,38E-10	***
IMC	-0,042724	0,013187	-3,240	0,0012	**

Nota: "*" significancia del 0,05; "**" significancia del 0,01; "***" significancia del 0,001

En el primer modelo Probit la única variable que no es significativa es el factor químico, aplicando el método Backward o eliminación hacia atrás, se obtiene el segundo modelo donde todas las variables individualmente son significativas al 5%, además los signos de los coeficientes son equivalentes al modelo Logit y tiene sentido lógico.

- **Evaluación de la multicolinealidad**

Tabla 14. Evaluación del factor de inflación de la varianza (VIF). Modelo Probit

Vif(pr2)					
SEXO	**EXM_FIS_DOLOR**	**FACTOR_MECANICOS**	**FACTOR_PSICOSOCIALES**	**EDAD**	**IMC**
1,012607	1,056291	1,057708	1,028729	1,082509	1,072747

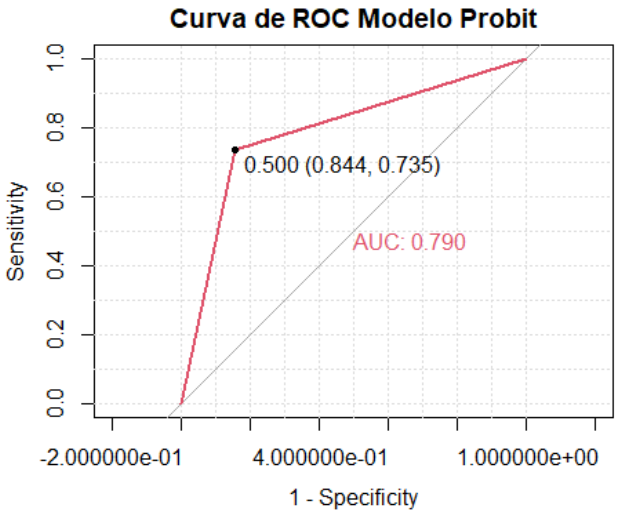
Vemos que los VIF de todas las variables, son pequeños, eso indica que no hay un problema de multicolinealidad entre esas variables, ninguna variable independiente está correlacionada

- **Bondad de ajuste del modelo**

El pseudo R^2 de Mc Fadden es 0.2326 y el valor ajustado 0.2203, lo cual quiere decir que el 22% de la patología está siendo explicado por las variables independientes incluidas en el modelo, sin embargo, es necesario evaluar algunos otros indicadores

- **Curva de ROC (Curva de característica operativa del recepto)**

Figura 9. Curva de ROC del Modelo Probit



En nuestro gráfico el área bajo la curva ROC es de 0.790 discrimina correctamente y es un modelo aceptable, es decir las predicciones se clasifican bien, cabe indicar que comparado con el modelo Logit la curva de ROC es algo mejor

- **Matriz de confusión**

Permite evaluar la predicción (Ver *Anexo 4*), tenemos 27 pacientes que el modelo ha clasificado como otras patologías cuando realmente formaba parte de ella, mientras que 27 pacientes que el modelo ha clasificado como otras patologías cuando realmente no formaban parte de ella, es decir lo que está dentro de la diagonal principal es lo que está correctamente clasificado, lo que está fuera de la diagonal principal es lo que no predijo el modelo correctamente.

Sensibilidad hace referencia al porcentaje de pacientes que se estima que forma parte de los pacientes con patología M54.5 cuando realmente forman parte de ella, equivale al 84%

Especificidad hace referencia al porcentaje de pacientes que se estimó que no formaban parte de los pacientes con patología M54.5 y realmente no formaban parte de ella, equivale al 73%

Una medida general de ver si el modelo está correctamente clasificado (Accuracy) es del 76% y un error del 24%

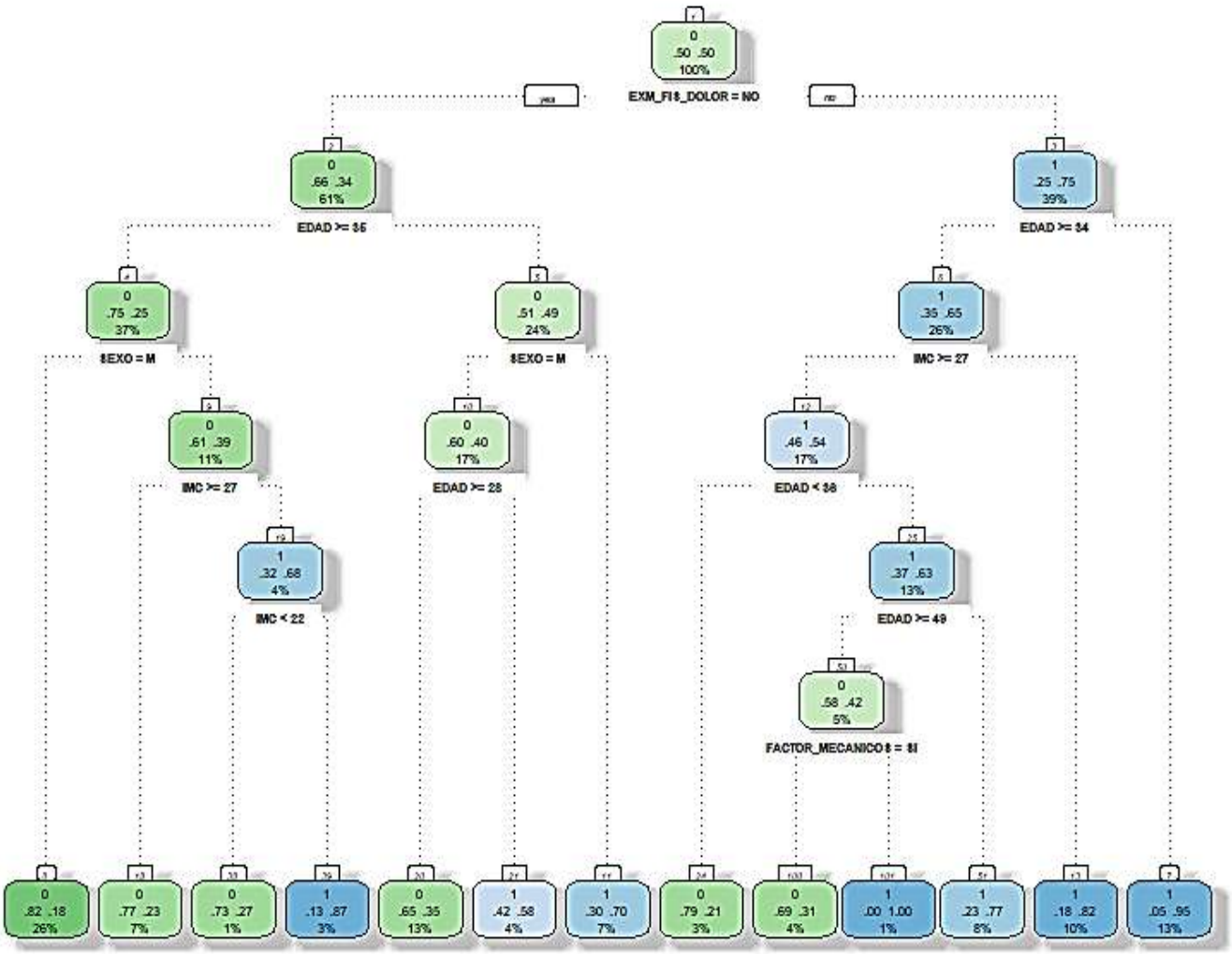
5.2.3 Árboles de clasificación, usando el paquete Rpart

El modelo se puede ver en el *Anexo 5*. Cabe indicar que no se aplicaron criterio de poda al árbol ya que el objetivo es hacer comparaciones equivalentes entre los cuatro modelos.

- **Diagrama de Árbol Inicial**

  Los rectángulos representan a un nodo del árbol con su regla de clasificación y coloreado según la categoría mayor, dentro de cada nodo se muestra la proporción de casos perteneciente a cada categoría y la proporción del total de casos que ha sido agrupada.

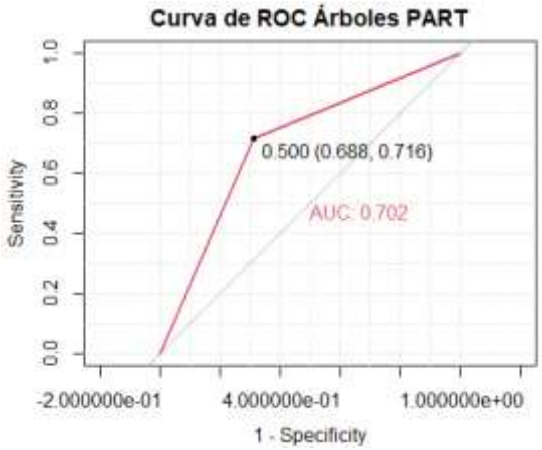
Figura 10. Diagrama de árbol. Modelo de clasificación (Rpart)



Hay que destacar que una de las hojas del árbol ha logrado un 100% de clasificaciones correctas para los pacientes con patología M54.5, y una también el 95% de clasificación correcta de la patología M54.5 (último lado derecho), cuya interpretación es, para pacientes que tienen dolor existe una probabilidad del 75% de que tenga la patología M54.5 y si es < 34 años existe un 95% de tener la patología M54.5. (Ver *Anexo 6*). Cabe indicar que no se aplicaron criterio de poda al árbol ya que el objetivo es hacer comparaciones equivalentes entre los cuatro modelos, como se mencionó anteriormente.

- **Curva de ROC (Curva de característica operativa del recepto)**

Figura 11. Curva de ROC. Árbol de clasificación (Rpart)



En nuestro gráfico el área bajo la curva ROC es de 0.70 es un modelo regular, es decir las predicciones se clasifican como regulares, cabe indicar que es más bajo que los modelos Logit y Probit

- **Matriz de confusión**

La matriz de confusión está en el *Anexo 7*. Tenemos una precisión (Accuracy) del 71% y un error del 29%, cabe señalar que el algoritmo del modelo utilizado busca la mejor separación para crear grupos

Sensibilidad hace referencia al porcentaje de pacientes que se estima que forma parte de los pacientes con patología M54.5 cuando realmente forman parte de ella, equivale al 69%

Especificidad hace referencia al porcentaje de pacientes que se estimó que no formaban parte de los pacientes con patología M54.5 y realmente no formaban parte de ella, equivale al 72%

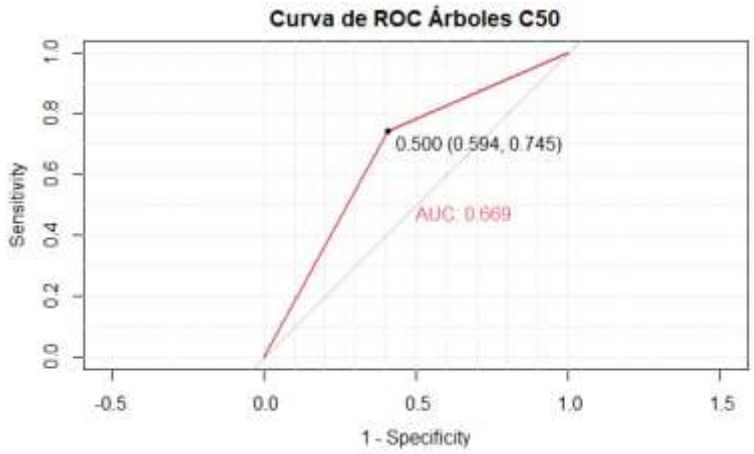
5.2.4 Árboles de clasificación, usando el paquete C5.0

Se crea un modelo de clasificación basado en un Árbol C 5.0. Al aplicar el summary al modelo creado, (Ver *Anexo 8*), muestra la información sobre el número de observaciones de entrenamiento que corresponde a 820 observaciones y 7 variables, igual que los modelos anteriores, adicionalmente se puede apreciar la división del árbol, los errores de entrenamiento y la importancia

de los predictores, con respecto a los errores son del 66 de las 820 observaciones que han sido clasificados incorrectamente, que representa el 8%

- **Curva de ROC (Curva de característica operativa del recepto)**

Figura 12. Curva de ROC. Árbol de clasificación (C5.0)



El área bajo la Curva ROC es de 0.67 es un modelo regular, es decir las predicciones se clasifican como regulares, cabe indicar que es más bajo que los modelos Logit y Probit y Árbol de clasificación con Rpart

- **Matriz de confusión**

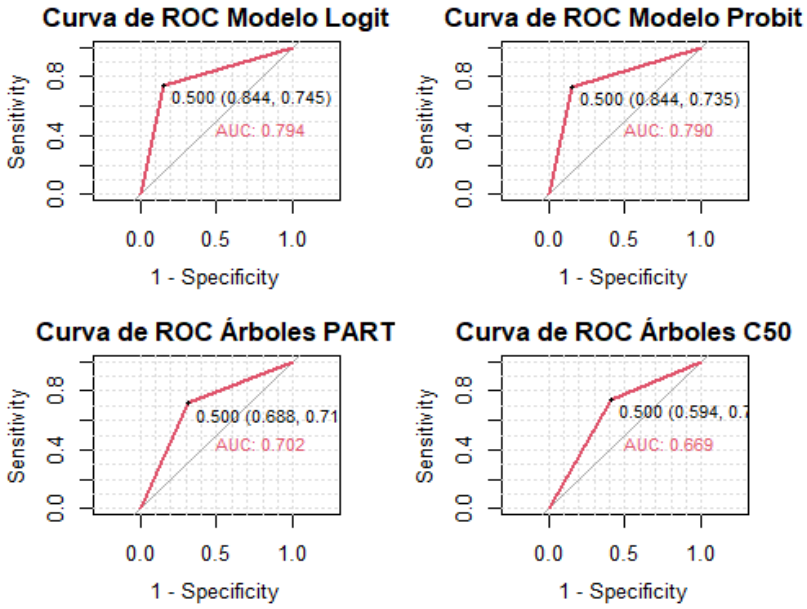
Evaluando el modelo a través de la matriz de confusión se ha obtenido una exactitud (Accuracy) de aproximadamente el 71% con un error del 29%, este error es mayor que el obtenido al ajustar el modelo con la base de entrenamiento, lo cual es de esperarse. Es importante señalar que este modelo tiene una sensibilidad del 59%, lo cual quiere decir que se puede predecir el 59% de los pacientes que tiene la patología M54.5, con este resultado se ve necesario mejorar este indicador (Ver *Anexo 9*)

5.2.5 Comparación de los modelos y análisis del mejor modelo

Se inicia comparando la curva de ROC de los cuatro modelos: Logit, Probit, Árbol de clasificación Part y C5.0

- **Curva de ROC**

Figura 13. Curva de ROC de los Modelos analizados

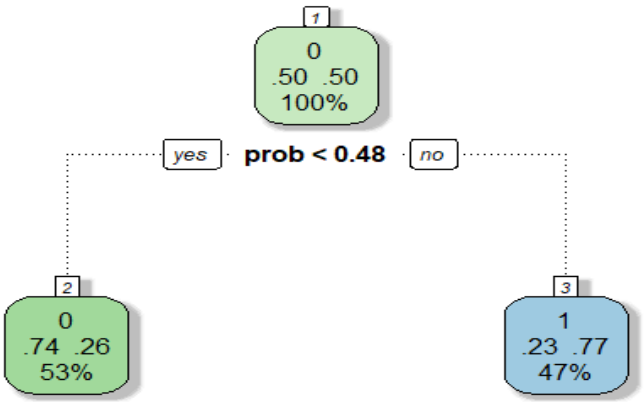


Aunque el análisis Logit y Probit producen resultados próximos, evaluando las curvas de ROC el mejor modelo es el Logit, además según la matriz de confusión se ha obtenido una mayor exactitud (Accuracy) con el modelo Logit de 0.7687, los otros indicadores también son mejores como la sensibilidad y la especificidad (Ver *Anexos 3,4,7 y 9*), con estos resultados se pone énfasis en el modelo Logit, se evalúa el corte óptimo y se analiza el modelo

- **Corte óptimo**

Es necesario evaluar cuál es el corte óptimo para obtener datos más representativos, con la data de entrenamiento y la función *optCutoff* se calcula el límite óptimo para una clasificación binaria, corte óptimo=0.4813417, adicionalmente para validar este resultado se evalúa también a través de un árbol de decisión

Figura 14. Árbol de decisión para evaluar el corte óptimo



Encontrar el corte óptimo nos permite hacer una estimación más precisa, se estima que los pacientes que tienen una probabilidad menor 48% en el modelo, van a tener el 74% de tener otras patologías y donde el modelo estima una probabilidad mayor o igual al 48%, van a tener el 77% de tener la patología M545

Matriz de confusión con corte óptimo

Se evidencia según el *Anexo 10*, que se tiene buenos resultados. La precisión (Accuracy) = 79%, la sensibilidad = 84%, lo cual quiere decir que puede predecir 84% de los pacientes que tienen la patología M54.5 y la especificidad = 77%, lo cual quiere decir que puede predecir el 77% de los pacientes que tienen las patologías menos frecuentes “Otra” y el área bajo la curva de ROC = 0.81

- **Significancia de los estimadores**

Según la *Tabla 10* presentada anteriormente se evidencia la significancia estadística de los estimadores individuales del modelo como son; sexo, examen físico del dolor, factores mecánicos, factores psicosociales, edad e IMC; y evaluando los signos de los coeficientes tiene sentido lógico:

Modelo Logit (lg2)						
Coefficients:	**Estimate**	**Exp(coefficient)-1**	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	4,15303	62,62676	0,67231	6,177	6,52E-10	***
SEXOM	-1,06511	-0,65531	0,18332	-5,81	6,24E-09	***
EXM_FIS_DOLORSI	1,97872	623.344	0,18252	10,841	< 2e-16	***
FACTOR_MECANICOSSI	-0,40388	-0,33227	0,17722	-2,279	0,02267	*
FACTOR_PSICOSOCIALESSI	0,79216	120.817	0,18832	4,206	2,59E-05	***
EDAD	-0,05687	-0,05528	0,00955	-5,955	2,60E-09	***
IMC	-0,07223	-0,06968	0,02245	-3,218	0,00129	**

Nota: “*” significancia del 0,05; “**” significancia del 0,01; “***” significancia del 0,001

Para facilitar la interpretación de los coeficientes calculamos el exponencial y analizamos el aumento o la disminución (columna 3)

- Acorde al modelo, el tener la patología M54.5 está negativamente relacionado con que sea hombre. Los hombres en comparación con las mujeres tienen menos probabilidad de tener la patología más frecuente M54.5 (frente a tener otras patologías de la columna lumbar). Si es hombre se tiene un 66% menos de tener la patología M54.5 que una mujer
- El tener la patología M54.5 está positivamente relacionado con la presencia de dolor. Si los pacientes tienen dolor tienen más probabilidad de tener la patología M54.5 (frente a tener otras patologías de la columna lumbar). Si presenta dolor el paciente tiene 6.2 veces más la posibilidad de tener la patología M54.5 que un paciente que no tiene dolor
- El tener la patología M54.5 está negativamente relacionado con que el paciente esté expuesto a factores de riesgo mecánico en su trabajo. Si los pacientes tienen factores de riego mecánicos en su trabajo tienen menos probabilidad de tener la patología M54.5 (frente a tener otras patologías de la columna lumbar). Si está expuesto a factores de riesgo mecánico se tiene un 33% menos de tener la patología M54.5 que un paciente que no está expuesto a este riesgo
- El tener la patología M54.5 está positivamente relacionado con que el paciente esté expuesto a factores de riesgo psicosociales en su trabajo. Si los pacientes tienen factores de riego psicosocial en su trabajo tienen

más probabilidad de tener la patología M54.5 (frente a tener otras patologías de la columna lumbar). Si está expuesto a factores de riesgo psicosocial se tiene 1,2 veces más (o el 121% más) de tener la patología M54.5 que un paciente que no está expuesto a este riesgo

- El tener la patología M54.5 está negativamente relacionado con la edad. A menos edad más probabilidad de tener la patología M54.5 (frente a tener otras patologías de la columna lumbar)
  Al aumentar un año de vida del paciente se disminuye en un 5.5% la posibilidad de tener la patología M54.5, que equivale a decir que por cada año de vida se espera que se disminuya en un 5.5% la posibilidad de tener la patología M54.5, manteniendo constante todo lo demás
- El tener la patología M54.5 está negativamente relacionado con IMC. A menor IMC más probabilidad de tener la patología M54.5 (frente a tener otras patologías de la columna lumbar). Si el IMC se incrementa en un punto porcentual entonces se tiene un 7% menos de tener la patología M54.5 frente a tener otras patologías de la columna lumbar

Podemos decir en base a los datos analizados que el perfil de los pacientes que tienen la patología más frecuente M54.5 son de sexo femenino, tienden a exponerse a factores de riesgo psicosocial en su trabajo, con presencia de dolor, de menor edad, menor IMC y menor riesgo de tener factores mecánicos en su lugar de trabajo

- **Evaluación de las variables más importantes**

Tabla 15. Importancia de las variables en el modelo

Variables	Overall
SEXOM	5.810.080
EXM_FIS_DOLORSI	10.840.958
FACTOR_MECANICOSSI	2.278.956
FACTOR_PSICOSOCIALESSI	4.206.448
EDAD	5.955.059
IMC	3.217.871

Los factores que pueden elevar el riesgo de que un paciente presente las patologías lumbares más frecuente M54.5-Lumbalgia, frente a las menos frecuentes Otra, son en orden de importancia o impacto: la presencia de dolor, la edad, el sexo, los factores psicosociales que están expuestos los pacientes en el trabajo, este tipo de riesgo implica alta responsabilidad, relaciones interpersonales, sobrecarga laboral, monotonía, minuciosidad en la tarea, el IMC y los factores mecánicos, este tipo de riesgo implica caídas de nivel, resbalones, golpes y/o cortes por objetos o herramientas, caída de objetos, entre otros , cabe indicar que el sexo masculino, la presencia de factores mecánica, la edad y el IMC se relacionan negativamente con la patología M54.5

Predicción con el mejor Modelo

El método de regresión logística permite estimar la probabilidad de la variable cualitativa patología (M545-Otra) en función de las seis variables independientes significativas del modelo, como venos en los siguientes ejemplos:

Ejemplo de predicción ($\hat{y}$) de varios casos evaluados a través del Modelo Logit

Tabla 16. Predicciones de la probabilidad de la patología (M545), según el modelo Logit

	Variables independientes						
Paciente N°	SEXO	EXM_FIS_DOLOR	FACTOR_MECANICOS	FACTOR_PSICOSOCIALES	EDAD	IMC	P (M545)
1	F	SI	NO	SI	20	30	**0,9739**
2	F	NO	NO	SI	20	30	**0,8377**
3	M	SI	NO	SI	20	30	**0,9279**
4	M	NO	NO	SI	20	30	**0,6401**
5	F	NO	NO	SI	60	30	**0,3467**

En base a lo analizado anteriormente era de esperarse estos resultados, en forma general se puede estimar que los pacientes que tienen dolor tienen una probabilidad más alta en padecer de Lumbalgia. Además, la edad también es importante, en pacientes que tienen mayor edad la probabilidad es más baja, que son las variables que se han variado en los ejemplos resueltos

6 Discusión

La lumbalgia constituye un problema de salud, frecuente en las consultas médicas, su etiología puede estar ligada a varios factores, como la edad, peso, IMC, riesgos laborales. (estrés, levantamiento de pesos, caídas, resbalones, entre otros).

En la presente investigación se encontró que la lumbalgia catalogada en el CIE 10 como M54.5 estuvo presente en el 76.7% de pacientes esto concuerda con el estudio realizado por Seguí y Gérvas, el cual manifiesta que alrededor del 60-80% de las personas tendrá al menos un episodio de lumbalgia en su vida. (22)

El examen físico es de vital importancia para el diagnóstico de la lumbalgia en los pacientes, se debe considerar que el principal síntoma es el dolor, sin embargo, la exploración física puede llevar a dar un diagnóstico más certero, en esta investigación el 96.7% presenta el examen de flexión normal y en cuanto al signo de lasegue el 93.6% fue negativo, lo cual demuestra que los pacientes no presentaron hernia discal, esto tiene relación con el estudio realizado por Sánchez Fernández (2012) donde manifiesta que el signo de lasegue fue positivo en el 91% de los pacientes que si tenían una hernia discal.(23)

En el presente estudio se encontró que el IMC está relacionado negativamente con la presencia de la patología (Lumbago=M54.5) en comparación con otras patologías menos frecuente de columna lumbar (Otro= M51.06, M51.16, M51.86, M54.16), resultado que llamó la atención, otro efecto es el consumo de tabaco, que no influye significativamente en la presencia de patologías de la columna lumbar, diferente con respecto a otros estudios en los cuales los resultados describen lo contrario

Según el artículo “Las enfermedades de la columna lumbar y su relación con el trabajo en España”, señala que son: …, el tabaquismo, obesidad, sedentarismo, debilidad muscular, entre otras, lo que se relaciona con las

enfermedades de columna lumbar. (15). Además, el Ministerio de Salud Pública del Ecuador, señala como factores de riesgo el incremento del IMC (IMC=peso [kg]/ estatura [m2]). (13), sin embargo, en el presente estudio no se encontraron resultados similares, pero, hay que destacar que el 92% de pacientes investigados realizan deporte al menos una vez al mes, lo cual incrementa la masa muscular y en consecuencia el IMC, adicionalmente hay que destacar que en el presente estudio hay una correlación más alta 82% entre peso con IMC, mientras que en un 62% entre peso con estatura.

Por otro lado, hay que señalar que el 100% de los pacientes investigados consumen al menos una vez al mes tabaco lo cual limita hacer una comparación con los que no consumen, de manera que en la presente investigación el consumo de tabaco no ayuda a discriminar la presencia de patologías de la columna lumbar

7 Conclusiones

En el presente trabajo de investigación se evalúa estadísticamente los factores de riesgo asociados a patologías en la región lumbar de la columna vertebral, diagnosticadas en radiografías, en pacientes ecuatorianos que cuentan con empleo y seguro privado; para evaluar la asociación, se comparan los modelos predictivos de respuesta binaria: Logit y Probit; y Árboles de clasificación usando el paquete Part y C5.0, teniendo como variable respuesta: la patología más frecuente Lumbalgia “M54.5” y las menos frecuentes “M51.06, M51.16, M51.86 y M54.16” codificada como “Otra”, finalmente se evidencia los mejores resultados con el modelo Logit, se tuvo una sensibilidad del 84%, con una especificidad del 77%, con un mejor punto de corte de 0.48, un área bajo la curva de ROC de 0.80 y una precisión del 79%

La mayoría de pacientes el 76.7% presentan la patología codificada como M54.5 que representa a la Lumbalgia y Otras patologías recoge el 23.35%, cabe indicar que esto corresponde a las patologías M51.86 que representa el 7.2%, M51.16 el 6.9%, M54.16 el 5.1% y M51.06 el 4.2%. (Ver *Tabla 1*)

Los factores que se deben tener en cuenta en la presencia de las patologías de columna lumbar más frecuente “Lumbalgia=M54.5”, frente a las menos frecuentes “Otra” por tener una asociación estadísticamente significativa, mediante el modelo predictivo Logit, de mayor a menor importancia son: la presencia de dolor, la edad, el sexo, los factores psicosociales que están expuestos los pacientes en el trabajo, este tipo de riesgo implica alta responsabilidad, relaciones interpersonales, sobrecarga laboral, monotonía, minuciosidad en la tarea; el IMC y los factores mecánicos, este tipo de riesgo implica caídas de nivel, resbalones, golpes y/o cortes por objetos o herramientas, caída de objetos, entre otros; cabe indicar que el sexo masculino, la presencia de factores mecánica, la edad y el IMC se relacionan negativamente con la patología M54.5

Las probabilidades de tener Lumbalgia son: Si los pacientes tienen dolor tienen 6,2 veces más la posibilidad de tener esta patología, que si no presentan dolor; por cada año de vida del paciente se espera que se disminuya en un 5.5% la probabilidad de tener esta patología, manteniendo constante todo lo demás; los hombres en comparación con las mujeres tienen un 66% menos de probabilidad de tener Lumbalgia; si los pacientes tienen factores de riego psicosocial en su

trabajo tienen 1,2 veces más (o el 121% más) de tener Lumbalgia, que un paciente que no está expuesto a este riesgo; si el IMC se incrementa en un punto porcentual entonces se tiene un 7% menos de probabilidad tener Lumbalgia (frente a tener otras patologías de la columna lumbar); finalmente si los pacientes están expuestos a factores de riego mecánicos en su trabajo tienen un 33% menos de probabilidad de tener Lumbalgia (frente a tener otras patologías de la columna lumbar), que un paciente que no está expuesto a este riesgo

Con los resultados obtenidos podemos analizar que el perfil de los pacientes que tienen la patología más frecuente Lumbago son de sexo femenino, tienden a exponerse a factores de riesgo psicosocial en su trabajo, con presencia de dolor, de menor edad, menor IMC y menor riesgo de tener factores mecánicos en su lugar de trabajo. *Cabe indicar que el perfil puede variar dependiendo de las condiciones laborales y de salud de los pacientes*

Es necesario incluir para futuras investigaciones más factores que pudieran influir en la presencia de patologías de columna lumbar como: factores genéticos, bioquímicos y nutricionales, lo cual permitirá identificar con más precisión los factores determinantes en patologías lumbares. Además, también que se puedan hacer comparaciones de pacientes que practican los hábitos de tomar alcohol y consumir cigarrillos frente a los que no consumen; con la nueva información se extenderá la posibilidad de nuevos estudios.

Adicionalmente se recomendaría aplicar para futuras investigaciones como estrategia el modelo predictivo de regresión logística multinomial como extensión de la regresión logística binaria, ya que la variable dependiente (patología) es de tipo nominal con más de dos categorías (M54.5, M51.16, M54.16, M51.06 y M51.86), lo cual permitiría tener una interpretación de la influencia de las variables explicativa (X1, X2, …, X6) sobre las probabilidades de todas las patologías de la región lumbar, a diferencia de una regresión logística binaria

Otra estrategia que se podría aplicar es el análisis discriminante simple o múltiple muy similar al análisis de regresión logística, la única diferencia es que las variables independientes deben ser cuantitativas que, aunque son limitadas en nuestro estudio, sí se dispone, este análisis permite identificar qué variables pertenecen a los grupos de patologías y que variables son significativas en la clasificación

En general el enfoque metodológico y la planificación planteada en la presente investigación ha sido el adecuado para lograr procesar y analizar la información, y demostrar los objetivos planteados

8 Glosario

CIE 10.- Clasificación internacional de enfermedades

M54.5.- Lumbago no especificado o lumbalgia

M51.06.- Trastornos de disco intervertebral con mielopatía, región lumbar

M51.16.- Trastornos de disco intervertebral con radiculopatía, región lumbar

M51.86.- Otros trastornos de disco intervertebral, región lumbar

M54.16.- Radiculopatía, región lumbar

IMC. - Índice de masa corporal

AUC. - Área bajo la curva de ROC

ROC: Curva de característica operativa del recepto (receiver operating characteristic)

9 Bibliografía

1. Moley P. Dolor lumbar - Trastornos de los huesos, articulaciones y músculos - Manual MSD versión para público general. Staff Manual's Editor [Internet]. 2020 [cited 2021 Mar 2]; Available from: https://www.msdmanuals.com/es-ec/hogar/trastornos-de-los-huesos,-articulaciones-y-músculos/dolor-lumbar-y-dolor-cervical/dolor-lumbar
2. Ministerio de Salud pública. Dolor lumbar: Guía de práctica Clínica. 2015;80. Available from: http://books.google.com/books?id=lwzKOgAACAAJ&pgis=1
3. (OMS) OM de la S. OMS | Atención del dolor lumbar: ¿los sistemas de salud son eficaces? Atención del dolor lumbar ¿los Sist salud son eficaces? [Internet]. 2021 [cited 2021 Mar 2];Volumen 99(Número 3):169–240. Available from: https://www.who.int/bulletin/volumes/97/6/18-226050-ab/es/
4. Oliveros Ribero Z, Delgado Páez R, Ramirez J. Signos radiológicos más frecuentes relacionados con dolor lumbar y su aplicabilidad en valoración pre-ocupacional. Rev Colomb salud Ocup [Internet]. 2018;8(1):1–7. Available from: http://revistas.unilibre.edu.co/index.php/rc_salud_ocupa/index Articulo
5. Vargas Gayón MR, Wilches MC, Estrada Orozco K. Radiografía de columna lumbosacra en dolor lumbar agudo: ¿uso o sobreuso? experiencia en el servicio de urgencias de dos centros de alta complejidad en Bogotá, Colombia. 2019;5126–31. Available from: http://contenido.acronline.org/Publicaciones/RCR/RCR30-2/03_Columna.pdf
6. Cáncer) N (Instituto N del. Definición de espina dorsal - Diccionario de cáncer del NCI - Instituto Nacional del Cáncer [Internet]. [cited 2021 Apr 9]. Available from: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/espina-dorsal
7. Enfermería y el dolor lumbar - Revista Electrónica de Portales Medicos.com [Internet]. [cited 2021 Apr 9]. Available from: https://www.revista-portalesmedicos.com/revista-medica/enfermeria-dolor-lumbar/
8. Cantos E, Moreira M, Rodríguez T, Aguayo J. Team building en la prevención de trastornos músculo esqueléticos en el personal administrativo de empresa atunera Seafman C . A . Salud & Ciencias médicas. 2021;1:28–34.
9. Chancasanampa Vega J, Díaz Lazo A. Patologías de columna lumbar diagnosticadas por radiografía convencional lumbar pathologies diagnosed by conventional radiography. 2017;8(2):21–6. Available from: http://revistas.udh.edu.pe/index.php/udh/article/view/109
10. INEC. Indicadores Laborales - encuesta nacional de empleo, desempleo y subempleo [Internet]. 2018. Available from: https://www.ecuadorencifras.gob.ec/documentos/web-inec/EMPLEO/2018/Marzo-2018/032018_Presentacion_M_Laboral.pdf
11. Tuotromedico. Enfermedades y síntomas comunes según CIE-10

(Clasificación Internacional de Enfermedades) [Internet]. 2021 [cited 2021 May 5]. Available from: https://www.tuotromedico.com/CIE10/

12. OMS. Tabla de IMC 2021 de la OMS (hombres y mujeres adultos) [Internet]. 2020. 221AD [cited 2021 Apr 9]. Available from: https://www.enterat.com/salud/imc-indice-masa-corporal.php
13. Ministerio de Salud Pública del Ecuador. Dolor lumbar: Guia práctica clínica. 2016;80. Available from: https://www.salud.gob.ec/wp-content/uploads/2017/02/GUÍA-DOLOR-LUMBAR_16012017.pdf
14. Madrid P. ¿Qué es el signo de Lasègue positivo? - Rehabilitación Premium Madrid [Internet]. [cited 2021 May 5]. Available from: https://rehabilitacionpremiummadrid.com/blog/jose-ignacio-diaz/que-es-el-signo-de-lasegue-positivo/
15. Herrero T, Íñiguez de la Torre V, Capdevila García L, López González Ángel, Terradillos García J, Aguilar Jiménez E, et al. Las enfermedades de la columna lumbar y su relación con el trabajo en España. Salud y Medio Ambient Fund MAPFRE Inst Prevención [Internet]. 2012; Available from: https://www.fundacionmapfre.org/documentacion/publico/i18n/consulta/registro.cmd?id=138943
16. Contreras Pinto WJ. Factores Asociados a la Enfermedad Discal Lumbar de Origen Laboral , Calificados por la Junta de Calificación de Invalidez Regional de Meta (Colombia). 2015;5(4):18–22.
17. Outliers detection in R - Stats and R [Internet]. [cited 2021 Apr 19]. Available from: https://statsandr.com/blog/outliers-detection-in-r/#introduction
18. Función glm - RDocumentation [Internet]. [cited 2021 Jun 3]. Available from: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm
19. What’s the Best R-Squared for Logistic Regression | Statistical Horizons [Internet]. [cited 2021 Jun 5]. Available from: https://statisticalhorizons.com/r2logistic
20. RPubs - Arboles de decisión con R - Clasificación [Internet]. [cited 2021 May 17]. Available from: https://rpubs.com/jboscomendoza/arboles_decision_clasificacion
21. Amat Rodrigo J. Arboles de decision, Random Forest, Gradient Boosting y C5.0 [Internet]. Ciencias de datos. [cited 2021 Jun 5]. Available from: https://www.cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_c50#C50
22. Seguí Díaz M, Gérvas J. Dolor Lumbar. Elsevier [Internet]. 2002;28(1):21. Available from: https://www.elsevier.es/es-revista-medicina-familia-semergen-40-articulo-el-dolor-lumbar-13025464
23. Sanchez Fernandez J, Serdeira A, Ziegler M, Donazar Severo C, Abreu Zardo E. Correlação do sinal de lasègue e manobra da elevação da perna, retificada com os achados cirúrgicos em pacientes com ciatalgia portadores de hérnia discal lombar. Scielo. 2012;11(1):2010–2.

9.1 Anexos

Anexo 1 Prueba de medianas Wilcoxon
Anexo 2 Frecuencias observadas y esperadas de v. cualitativas
Anexo 3 Matriz de confusión Modelo Logit
Anexo 4 Matriz de confusión Modelo Probit
Anexo 5 Modelo Árbol de Clasificación rpart
Anexo 6 Árbol de clasificación rpart
Anexo 7 Matriz de confusión Modelo Árbol de clasificación rpart
Anexo 8 Modelo Árbol de clasificación C5.0
Anexo 9 Matriz de confusión C5.0
Anexo 10 Matriz de confusión Modelo Logit con corte óptimo
Anexo 11 Informe con RStudio