Miranda Silveria Manzanares

Anàlisi i modelat de la diabetis mellitus gestacional amb regressió logística

Treball Fi de Grau

dirigit pel Dr. Agustí Solanas

Grau en Enginyeria Biomèdica

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2022

Agraïments

Al Dr. Agustí Solanas per donar-me l'oportunitat de fer les pràctiques al Smart Technologies Research Group i guiar-me per assolir els coneixements necessaris per portar endavant aquest treball.

A la Dra. Ana Megía per proporcionar-me la base de dades reals de l'hospital Joan XXIII i a totes les dones i els seus fills que formen part de l'estudi.

A la meva família i amics pel suport que m'han donat durant aquest últim any.

Resum

La diabetis mellitus gestacional és una complicació metabòlica de l'embaràs que afecta a dones sense història clínica prèvia d'hiperglucèmia. En la majoria dels casos, desapareix un temps després de donar a llum, tot i això, és imprescindible diagnosticar-la en una etapa poc avançada per poder mantenir el control i evitar efectes perjudicials en la salut de la dona i el seu fill. En aquest treball, es fa un anàlisi multivariant de dades reals de pacients de l'Hospital Joan XXIII de Tarragona, amb l'objectiu d'estudiar els factors de risc i construir un model de regressió logística per detectar la malaltia.

Paraules clau: diabetis mellitus gestacional, model regressió logística, anàlisi multivariant.

Resumen

La diabetes mellitus gestacional es una complicación metabólica del embarazo que afecta a mujeres sin historia clínica previa de hiperglucemia. En la mayoría de los casos, desaparece un tiempo después de dar a luz, sin embargo, es imprescindible diagnosticarla en una etapa poco avanzada para poder mantener el control y evitar efectos perjudiciales en la salud de la madre y el hijo. En este trabajo, se realiza un análisis multivariante de datos reales de pacientes del Hospital Joan XXIII de Tarragona, con el objetivo de estudiar los factores de riesgo y construir un modelo de regresión logística para detectar la enfermedad.

Palabras clave: diabetes mellitus gestacional, modelo de regresión logística, análisis multivariante.

Abstract

Gestational diabetes mellitus is a metabolic complication of pregnancy that affects women with no previous clinical history of hyperglycemia. In most cases, it disappears some time after giving birth, however, it is essential to diagnose it at an early stage in order to maintain control and avoid harmful effects on the health of the mother and her baby. In this work, a multivariate analysis of patient data from Hospital Joan XXIII is carried out, with the aim of studying risk factors and building a logistic regression model to detect the disease.

Keywords: gestational diabetes mellitus, logistic regression model, multivariant analysis.

Índex

1 Introducció

La diabetis mellitus gestacional (DMG) és una complicació metabòlica de l'embaràs que normalment es detecta entre la 24a i 28a setmana de gestació i afecta a dones sense història clínica prèvia d'hiperglucèmia. En la majoria dels casos, desapareix un temps després de donar a llum. Entre els efectes perjudicials de la malaltia trobem la preeclàmpsia, la macrosomia, la obesitat, i està associada a un augment del risc per les mares de desenvolupar diabetis tipus 2 en un futur.

Durant l'embaràs el cos s'adapta per generar l'entorn necessari pel correcte desenvolupament del fetus, en conseqüència apareix un desequilibri metabòlic que s'associa amb l'augment de risc d'aparició de malalties. En relació amb el metabolisme glucídic, durant les primeres setmanes de gestació els teixits materns canvien la seva sensibilitat a la insulina, això provoca la disminució del nivell de glucosa en sang en dejuni i l'augment en períodes postprandials per tenir suficients reserves de nutrients i que aquests puguin anar cap a la unitat fetoplacentària. Quan aquest desequilibri es torna constant, augmenta la resistència a la leptina i el número de citocines proinflamatòries, es genera un estat que pot desencadenar en l'aparició de la DMG [1].

En l'atenció sanitària maternofetal és primordial diagnosticar la diabetis gestacional en una etapa poc avançada de la malaltia per poder evitar les greus afectacions associades a l'exposició prollongada als nivells elevats de glucosa, com les malformacions o els avortaments [2]. Per aquesta raó, s'ha posat el focus de la investigació en desenvolupar noves tècniques que permetin detectar la malaltia aviat, tot i això, com veurem més endavant, encara no hi ha un consens universal del mètode de diagnosi ni el tractament que han de seguir les pacients.

1.1 Motivació i objectiu

Durant el curs 2021/2022 he estudiat la línia de recerca en salut intel·ligent al Smart Technologies Research Group de la URV. He fet una anàlisi de l'estat de l'art, seguint el mètode de Vom Brocke, dels models matemàtics per a la detecció de la DMG. Aquest estudi m'ha permès veure de manera global les tècniques d'anàlisi que s'han utilitzat per crear els models de diagnosi de la DMG al llarg de la història i també, conèixer els principals biomarcadors, els factors de risc i els mètodes de cribratge estàndard. A més, he complementat les pràctiques amb el treball de fi de grau. L'objectiu d'aquest és fer una anàlisi multivariant de dades reals adquirides de l'Hospital Universitari Joan XXIII de Tarragona. La principal finalitat és construir un mètode de detecció de la malaltia utilitzant R, un software de programació enfocat a l'anàlisi estadístic. El classificador que hem seleccionat és un model de regressió logística que ens permeti classificar el grup d'observació binomial, és a dir, si és control o té diabetis gestacional, a partir d'un conjunt d'entrenament de variables quantitatives.

Per poder situar l'objectiu de l'anàlisi, en primer lloc, presentem el marc teòric de la DMG i els mètodes de cribratge que s'utilitzen de manera general. Seguidament, comentem alguns conceptes d'anàlisi multivariant i les tècniques matemàtiques que utilitzem al treball. Un cop descrita la informació que envolta el context del treball, expliquem detalladament els diferents apartats per a la preparació de les dades, la construcció del model de regressió logística, i utilitzem un conjunt d'avaluació per testejar el classificador. Finalment, acabem el treball amb una valoració general dels conceptes apresos i els resultats de l'anàlisi estadística.

2 La diabetis gestacional

2.1 Introducció a la diabetis mellitus

2.1.1 Contextualització

2.1.1.1 La glucosa

La glucosa ($C_6H_{12}O_6$) és la principal font d'energia per la respiració cel·lular aeròbica i anaeròbica dels organismes. La major part és adquirida a partir de la ingesta d'aliments i queda emmagatzemada en forma d'un polímer anomenat glucogen que s'allibera quan estem en dejuni i necessitem energia. A banda, una altra font d'obtenció és la gluconeogènesis, el procés de síntesi de la glucosa a partir de la degradació de greixos i proteïnes. Podem trobar la glucosa en diferents formes isomètriques:

- Monosacàrid: glucosa, fructosa i galactosa
- Disacàrid: lactosa i sacarosa
- Polisacàrid: midó [3]

Hi ha dos tipus de famílies de proteïnes de membrana que s'encarreguen de la difusió de la glucosa a través de la membrana cel·lular, els transportadors de glucosa acoblats a sodi (SGLT) i les proteïnes facilitadores del transport de glucosa (GLUT). Trobem els SGLT als epitelis de l'intestí prim i l'epiteli tubular renal, aquí la principal funció és l'absorció i la reabsorció de nutrients. Per l'altre costat, les GLUT estan presents a totes les cèl·lules i són les encarregades de transportar la glucosa entre els diferents compartiments [4].

La glucèmia, que és el nivell de glucosa a la sang, està regulada per hormones pancreàtiques: la insulina que s'encarrega de la penetració de la glucosa dins les cèl·lules, per tant redueix la concentració i el glucagó que l'augmenta principalment a partir de reaccions al fetge [5].

2.1.1.2 El pàncrees.

El pàncrees és una glàndula situada a l'abdomen posterior, entre el duodè i la melsa, darrera de l'estómac, que participa en els sistemes digestiu i endocrí. La seva estructura es divideix en cap, coll, cos i cua i està format per diferents tipus de cèl·lules endocrines i exocrines.

Aproximadament el 98% està format de cèl·lules exocrines que produeixen els enzims del suc pancreàtic encarregats de la digestió de greixos, proteïnes i carbohidrats. El suc excretat s'aboca als conductes i aquests el condueixen cap al duodè.

El 2% restant està compost de cèl·lules endocrines encarregades de produir hormones per mantenir el control de la glucèmia. El 90% d'aquestes s'agrupen en estructures anomenades Illots de Langerhans. Aproximadament n'hi ha un milió i estan distribuïts per tota la glàndula, en major quantitat al cos i la cua [6]. Estan formats per diferents tipus de cèl·lules, les cèl·lules principals són α, β, δ i PP que secreten glucagó, insulina, somatostatina i polipèptid pancreàtic respectivament com podem veure a la Figura 1 i a la Taula 1.

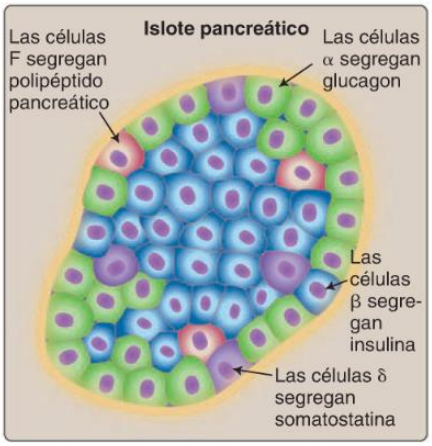

Figura 1. Illot de Langerhans amb les cèl·lules α, β i δ i les hormones que secreten indicades. Font: [5]

Taula 1. Hormones principals del funcionament pancreàtic. Font: [6]

Cèl·lula	Hormona	Funció
α	Glucagó	Augmenta el nivell de glucosa en la sang.
β	Insulina	Permet que la glucosa entri dins les cèl·lules dels teixits.
δ	Somatostatina	Inhibeix la secreció d'hormones de les cèl·lules endocrines i exocrines.
PP o F	Polipèptid Pancreàtic	Inhibeix la secreció d'hormones de les cèl·lules exocrines del pàncrees.

Existeix una comunicació directa entre aquestes cèl·lules a través de l'espai extracel·lular o mitjançant connexions intracel·lulars entre contigües, aquesta relació permet que entre elles puguin regular l'alliberació i la inhibició de les hormones [7]. Els sistemes nerviosos simpàtic i parasimpàtic també enerven l'Illot intervenint en l'alliberació (simpàtic) o en l'emmagatzematge de substrats energètics (parasimpàtic) [5].

Les hormones són alliberades dins del corrent sanguini per poder ser transportades per tot el cos, per aquesta raó els Illots estan molt vascularitzats [7]. La sang entra pel centre de l'Illot i el flux es mou cap el perímetre permetent que les hormones puguin regular totes les cèl·lules. Hi ha un gran nombre de cèl·lules beta i aquestes es troben al centre envoltades per les alpha [5].

2.1.1.3 La insulina

La insulina és una hormona de dues cadenes de pèptids formada per 51 aminoàcids que aproximadament viu entre 3 i 8 minuts, més del 50% es degrada al fetge i la resta als ronyons i als teixits perifèrics. Aquesta és l'encarregada de la disminució del nivell de glucosa a la sang [5]. Estimula l'augment d'agregació de molècules GLUT a la membrana plasmàtica de les cèl·lules del múscul esquelètic, del teixit adipós i del fetge, per tant la glucosa entra en gran quantitat dins dels teixits i disminueix la hiperglucèmia de l'organisme [4].

La insulina es sintetitza inicialment com a preproinsulina al nucli de cèl·lules β dels Illots del pàncrees. Després, al reticle endoplasmàtic rugós es plega per dos ponts disulfur per transformar-se en proinsulina. Finalment a l'aparell de Golgi s'empaqueten i es guarden als gràmuls secretors on a través de les proconvertasas 1 i 2 i la carboxipeptidasa E la separen

produint la insulina i un pèptid C, quan sigui necessari l'organisme enviarà un senyal que els alliberarà al corrent sanguini [7].

El pèptid C és inactiu i està format per 31 aminoàcids, el fetge no el degrada per tant té més temps de vida que la insulina i es pot utilitzar com a biomarcador per controlar si les cèl·lules beta funcionen adequadament [5].

La glucosa és el principal regulador de la secreció d'insulina, a partir de les molècules transportadores de glucosa tipus 2 (GLUT-2) les molècules de sucre entren dins les cèl·lules β i es fosforil·len formant la glucosa-6-fosfat (GLU-6-P) que queda atrapada. Després, als mitocondris s'inicia el metabolisme de la glucosa i augmenta la concentració de trifosfat d'adenosina (ATP), en conseqüència desencadena que es tanquin els canals de potassi dependents d'ATP (K-ATP) que hi ha a la membrana plasmàtica. Això genera una despolarització que obre els canals de calci (Ca^{2+}), dependent del voltatge i gradient generat estimula l'exocitosi dels grànuls emmagatzemats de la síntesi de la insulina que contenen la hormona i el pèptid C que són alliberats al sistema circulatori [7].

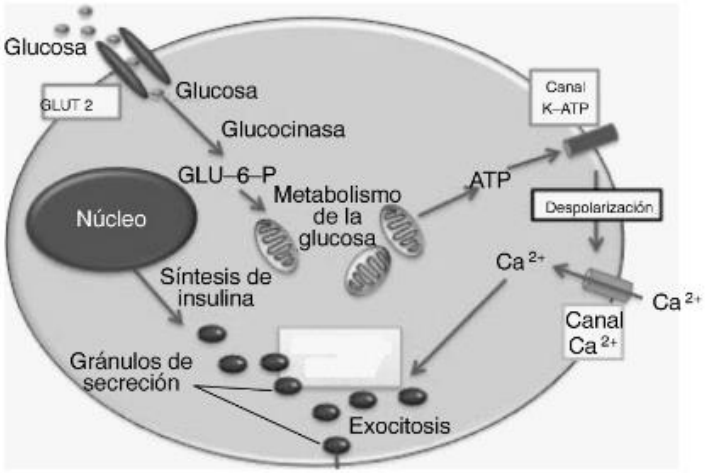

Figura 2. Esquema de la secreció d'insulina. Font: [7]

A través d'un circuit de retroalimentació negativa, quan hi ha molta concentració de glucosa a la sang, es redueix la secreció d'insulina induïda per la somatostatina que és alliberada per les es cèl·lules δ i també per l'estimulació α-adrenèrgica del sistema nerviós simpàtic [5].

2.1.1.4 El glucagó

El glucagó és una hormona peptídica de 29 aminoàcids sintetitzada a partir de la proteòlisi del proglucagó, en aquesta reacció s'allibera l'hormona i dos fragments de proteïnes inactives. Aproximadament un cop s'ha alliberat dura entre 5 i 10 minuts fins que arriba al fetge que el degrada.

S'encarrega principalment d'estimular la descomposició del glucogen perquè els teixits del cos obtinguin energia en períodes de dejuni o estrès a partir de la fosforilasa de glucogen i la glucosa-6-fosfatasa. Per un altre costat inicia la gluconeogènesi provinent de fonts de lípids i les proteïnes i inhibeix reaccions de degradació de de la glucosa.

Hi ha dues maneres per les que s'estimula la secreció, per un costat amb l'augment de la concentració de colecistoquinina (CCK) i d'aminoàcids. Per l'altra banda, per retroalimentació negativa al detectar un nivell baix de glucosa en sang i també per ordre del sistema nerviós

simpàtic en moments d'estrès. Aquest circuit s'inhibeix amb l'alliberació d'insulina i somatostatina, l'increment de sucre, àcids grassos i cossos cetònics [5].

2.1.2 La diabetis mellitus

La diabetis mellitus (DM) és una malaltia metabòlica caracteritzada per l'augment crònic de la concentració de glucosa en sang, en altres paraules, la hiperglucèmia que està causada per defectes en la segregació i/o la funció de la insulina. A Catalunya hi ha aproximadament 600.000 persones que la pateixen, el 15% de la població de més de 15 anys i el 20% major de 65 [8].

El procés de regulació fisiològic es coneix com l'homeòstasi de la glucosa. En valors baixos el pàncrees segrega glucagó i aquest actua sobre el fetge perquè catabolitzi les reserves de glucogen i s'alliberi glucosa al torrent sanguini. En canvi, quan la concentració és massa alta, segrega insulina i aquesta estimula l'absorció de glucosa per part dels teixits i l'emmagatzematge al fetge en forma de glucogen. En situacions d'estrès o perill la glàndula pituïtària produeix corticotropina (ACTH) que actua sobre el còrtex suprarenal per produir hormones que augmenten la velocitat de degradació de proteïnes per transformar-les en glucosa. Finalment, també actuen fibres del sistema nerviós simpàtic que estimulen la medul·la suprarenal per l'alliberació d'adrenalina i noradrenalina augmentant també el nivell de sucre a la sang [9].

La diabetis apareix quan s'altera el circuit de retroalimentació de la insulina i en conseqüència el cos no pot disminuir la glucosa en sang i regular-lo quan aquest es troba en nivells alts [9]. Si no es detecta aviat i es tracta adequadament pot arribar a perjudicar als diferents òrgans, principalment els ulls, els ronyons, els nervis, el cor i els vasos sanguinis, per aquesta raó és essencial diagnosticar, classificar i tractar tant als pacients que ja pateixen la diabetis com els que es troben en una etapa prediabètica [10].

L'Associació de la Diabetis Americana (ADA) ha determinat quatre propostes d'estratègies per diagnosticar la DM:

1. Glucèmia ≥200mg/dL en qualsevol moment del dia. No necessita confirmació.
2. Glucèmia ≥126 mg/dL després de mínim 8 hores de dejuni. Sí necessita confirmació.
3. Glucèmia ≥200mg/dL en una prova de tolerància oral a la glucosa (PTOG), també coneguda com a sobrecàrrega oral de glucosa (SOG) després de 2 hores de prendre la solució de 75g de glucosa anhidra diluïda en 300 ml d'aigua, tal com recomana la Organització Mundial de la Salut (OMS). Sí necessita confirmació.
4. Hemoglobina A_1c(HbA_1c) ≥ 6,5%. Sí necessita confirmació.

Menys amb la primera proposta, en les altres proves es necessita confirmar els resultats per assegurar el diagnòstic. Si coincideix el resultat de dues proves s'accepta i si difereixen se'n demana una tercera [2].

Gràcies a l'augment de coneixements sobre la fisiopatologia de la diabetis, s'han pogut identificar els primers signes indicadors de l'inici de la malaltia i per tant, millorar l'atenció sanitària permetent prevenir o retardar la seva aparició [10]. Els criteris actuals de la detecció de la prediabetis són:

1. Glucosa alterada en període de dejuni: 126 mg/dL ≥ Glucèmia ≥100mg/dL
2. Intolerància a la glucosa després d'un PTOG de 2 hores amb 75g: 200 mg/dL ≥ Glucèmia ≥140mg/dL
3. 6,4% ≥ Hemoglobina A_1c(HbA_1c) ≥ 5,7% [2].

2.1.3 Tipus de diabetis mellitus

Hi ha diferents factors que poden afectar en la patogènesi de la DM, per exemple l'organisme por generar una reacció autoimmune que destrueixi les cèl·lules beta del pàncrees que són les encarregades de secretar insulina o també, el problema pot estar en la resistència que oposen els teixits a la glucosa [2]. En funció del mecanisme que causa la hiperglucèmia trobem diferents tipus de diabetis mellitus.

2.1.3.1 Diabetis mellitus tipus 1 (DM1)

Quan la hiperglucèmia és causada per la destrucció o mal funcionament de les cèl·lules β parlem de diabetis de tipus 1, aquest grup representa entre el 5 i el 10% dels casos [2]. En moltes ocasions és necessari un tractament periòdic d'injeccions d'insulina per suplir la falta de l'hormona i regular l'entrada de glucosa als teixits, per això aquest tipus també es coneix com a diabetis insulinodependent [5]. Principalment està causada per una resposta immunitària d'anticossos que destrueixen les cèl·lules secretores d'insulina. La majoria de persones que pateixen la DM1 són infants i adolescents però també és possible que es manifesti en l'edat adulta. Existeix variació en la velocitat de deteriorament, normalment en nens és ràpida mentre que en adults és més lenta. És una afectació que no es pot evitar ja que principalment està relacionada amb factors genètics i ambientals [2].

Hi ha un petit subgrup de DM1 amb alta predisposició de ser heretable en que es desconeix l'origen de la falta d'insulina, ja que no hi ha evidències d'activitat immunitària per la destrucció de les cèl·lules beta, aquest subtipus es coneix com a diabetis idiopàtica [2].

2.1.3.2 Diabetis mellitus tipus 2 (DM2)

Aproximadament el 90% de la població amb diabetis en té d'aquest tipus. En aquest cas les cèl·lules β funcionen correctament i és causada per la resistència de les cèl·lules a la insulina [2]. Normalment apareix en l'edat adulta, està relacionada amb factors hereditaris, el sedentarisme i la obesitat, hipertensió arterial i alteracions dels greixos. Es desenvolupa de forma lenta, per tant, és possible estar molt de temps sense símptomes i que passi desapercebuda [11]. Els tractaments es basen en canvis en l'estil de vida del pacient, principalment amb exercici físic i dietes equilibrades, i amb fàrmacs que disminueixin la resistència a la insulina [2].

2.1.3.3 Diabetis mellitus gestacional (DMG)

Aquesta malaltia afecta al 10% de les embarassades i és coneix com la hiperglucèmia que apareix per primer cop durant l'embaràs en dones sense història prèvia de diabetis. Durant la gestació hi ha canvis metabòlics per augmentar les reserves d'energies i quan la insulina alliberada és insuficient apareix la DMG. És molt important detectar-la i monitoritzar-la aviat, ja que si no es fa un tractament pot afectar la salut de l'infant i la mare. Alguns dels factors de risc més evidenciats són l'edat, la obesitat, antecedents familiars i DMG prèvia [11].

2.2 La diabetis mellitus gestacional

2.2.1 Definició

La diabetis gestacional és la hiperglucèmia causada per l'aparició del trastorn de la intolerància als carbohidrats [12], aquesta es detecta normalment a partir del segon trimestre de l'embaràs en dones sense història prèvia clínica de diabetis [13].

L'obesitat és el principal factor de risc pel desenvolupament de diabetis tipus 2. En els últims anys la prevalença d'aquesta condició ha augmentat considerablement a tot en món i en conseqüència, també ha augmentat el número de dones embarassades amb sobrepès i

DM2 no diagnosticada. És per això que en pacients amb factors de risc de patir diabetis tipus 2, en els primers controls es fa una prova de cribratge seguint els criteris generals. En cas que es diagnostiqui la DM2 durant el primer trimestre es considera que la malaltia ja era present abans de l'embaràs, per tant, aquestes pacients es classifiquen com a dones amb diabetis pregestacional (DPG) [10]. Tant la DMG com la DPG si no es diagnostiquen i es tracten aviat poden afectar el procés de l'embaràs i produir greus complicacions per la salut de la mare i el fill [12].

Abans del descobriment de la insulina es considerava que una dona amb diabetis no podia tenir fills ja que afectava les capacitats de reproducció i en relació, hi havia una gran taxa de mortalitat fetal i maternal per causa de la malaltia. Per tant, per evitar morts es recomanava que no quedessin embarassades i també els avortaments terapèutics. Frederick Banting va descobrir la insulina a l'any 1921, a partir d'aquest moment es van dissenyar tractaments, per un altre costat, els controls obstètrics eren més exhaustius i la taxa de mortalitat va disminuir considerablement, la de les mares va passar ràpidament del 45 al 2% mentre que la dels infants va anar baixant progressivament [12].

2.2.2 Factors de risc

Degut a l'interès en dissenyar estratègies de diagnòstic que permetin detectar la malaltia el més aviat possible, s'ha posat el focus en l'estudi de possibles factors de risc que permetin trobar la malaltia. Amb l'estudi de l'estat de l'art, en el que vaig filtrar 4131 articles seguint pautes de qualitat i interès, en vaig seleccionar 121 que consistien en aquest tipus d'estudis que vaig classificar en:

- Biomarcadors: diferents molècules útils per a la diferenciació entre grups control i amb DMG. Per exemple, adipòcits, la proteïna C reactiva, adiponenctina baixa al primer trimestre, microRNAs circulants, deficiència de la concentració de vitamina D, marcadors de la tiroides, fibronectina glicosilada, la microbiota intestinal, el perfil de la orina o factors genètics entre d'altres.

- Paràmetres relacionats amb el pes: el guany de pes al començament de l'embaràs, l'índex de massa corporal durant i abans de la gestació o l'obesitat estan clarament relacionats amb la diabetis gestacional.

- Patrons dietaris: patrons de la dieta que poden augmentar o prevenir l'aparició de la DMG. Per exemple, una dieta alta en colesterol o la consumició de refrescs estan associades amb un major risc, mentre que la dieta mediterrània i el suplement de folat el disminueix.

- Entorn ambiental: factors ambientals que poden afectar al risc de patir la malaltia, per exemple: la temperatura, l'exposició al bisfenol A o a metalls pesats com el Cadmi o la contaminació de l'aire en general per exemple amb les partícules PM2.5.

- Ètnia: hi ha una alta prevalença en dones llatinoamericanes i asiàtiques.

- Altres malalties: trobem indicats que la malaltia està relacionada amb la síndrome de l'ovari poliquístic, la preeclàmpsia, la disfunció de la tiroide i la malaltia del fetge gras no alcohòlic (NAFLD).

- Trastorns del son: problemes de respiració, la duració, la qualitat del son i els roncs.

- Altres: l'efecte de la reproducció assistida, el sexe del bebè, fumar, l'activitat física, la llargada i regularitat del cicle menstrual, la presa de medicaments antipsicòtics i l'índex inflamatori.

A la Taula 2 es mostren els factors de risc de la diabetis gestacional descrits als documents de consens de la SEGO (Sociedad Española de Ginecología y Obstetricia) i la GEDE (Grupo Español de Diabetes y Embarazo)

Taula 2. Factors de risc de la diabetis gestacional segons els documents de consens SEGO (*Sociedad Española de Ginecología y Obstetricia*) i GEDE (*Grupo Español de Diabetes y Embarazo*). Font: [12]

Factors de risc de la diabetis gestacional

• Familiar de primer grau amb història clínica de DM
• Història clínica de DMG anterior o d'altres alteracions del metabolisme glucídic.
• Obesitat (IMC > 30kg/m^2)
• Edat > 35 anys
• Haver nascut amb macrosomia
• Antecedents obstètrics:
  ○ Avortaments habituals
  ○ Mort fetal sense causa
  ○ Malformacions
  ○ Macrosomia
• Durant l'embaràs actual:
  ○ Macrosomia
  ○ Hidramnis
  ○ Malformacions
• Origen ètnic d'alt risc: dones afroamericanes, asiàtic-americanes, hispàniques, indi-americanes

2.2.3 Repercussions

2.2.3.1 Sobre l'embaràs

Algunes de les complicacions que poden aparèixer a causa de la diabetis són les infeccions urinàries, candidiasis vaginal, l'acumulació excessiva de líquid amniòtic que envolta el fetus (polihidramnis) , estats d'hipertensió (preeclàmpsia) i la provocació de parts prematurs i cesàries [14].

2.2.3.2 Sobre el fetus i el bebè

La diabetis pregestacional, durant el període d'organogènesi pot causar avortaments i malformacions. A més, si la mare té alguna vasculopatia el fetus pot patir creixement intrauterí retardat (CIR).

Per altra banda, tant per DMG com DPG els efectes són: l'alteració del benestar fetal; macrosomia que pot provocar traumatismes i distòcies fetals que augmenten la taxa de cesàries; miocardiopatia hipertròfica; immaduresa fetal que pot desencadenar en la síndrome

de distrès respiratori (SDR), és a dir, en el moment de néixer el bebè encara no ha desenvolupat totalment els pulmons, o també altres alteracions metabòliques [14].

Amb el pas dels anys, els nens que han nascut d'embarassos afectats per la DM materna, tenen predisposició a patir obesitat, alteracions del metabolisme hidrocarbonat i la síndrome metabòlica [14].

2.2.3.3 Sobre la mare

La diabetis pregestacional pot afavorir el deteriorament del control metabòlic i el desenvolupament de la retinopatia diabètica. Per l'altra banda, patir DMG durant l'embaràs augmenta les possibilitats de patir DM2, alteració de la concentració en sang de lípids i proteïnes (dislipèmia), obesitat i hipertensió arterial (HTA), en definitiva malalties associades a la síndrome metabòlica. Ocasionalment es genera una resposta autoimmune de destrucció de les cèl·lules β donant lloc a una futura diabetis de tipus 1 [14].

2.2.4 Etiologia

Durant l'embaràs l'organisme genera una sèrie de canvis cardiovasculars, respiratoris i metabòlics per tal de mantenir l'equilibri adequat entre la mare i el fetus i permetre el correcte desenvolupament d'aquest [1]. L'estratègia que utilitza per assegurar que hi hagi reserves és generar resistència a la insulina per poder emmagatzemar els nutrients i utilitzar-los per alimentar el fetus en períodes de dejuni. Aquest fenomen és coneix com l'anabolisme característic de la primera meitat de la gestació. Fisiològicament les cèl·lules β, estimulades per les hormones de l'embaràs, augmenten la seva producció d'insulina. Metabòlicament, els estrògens i la progesterona mantenen la glucèmia i ajuden a la insulina en la seva funció al fetge per activar l'emmagatzematge de glucogen [12].

Al voltant de la setmana 26 la concentració de cortisol, lactogen placentari humà (HPL) i prolactina es troben en nivells molt alts. Aquestes hormones s'encarreguen de produir la resistència a la insulina per afavorir el metabolisme catabòlic i que les substàncies emmagatzemades s'utilitzin per nodrir el fetus i la placenta [12]. Durant l'última etapa de l'embaràs és quan més s'utilitza la glucosa fetal per aquesta raó augmenta la gluconeogènesi i disminueix la sensibilitat a la insulina [1].

Finalment, la diabetis gestacional apareix quan el cos de la dona no pot mantenir l'equilibri entre la secreció i la resistència a la insulina i s'expressa com una hiperglucèmia crònica durant la gestació [12].

2.2.5 Mètodes i criteris de diagnosi

2.2.5.1 Propostes de les diferents organitzacions internacionals

Universalment no existeix consens sobre els mètodes de diagnosi de la diabetis gestacional, és per això que es un tema amb controvèrsia dins de la comunitat mèdica i depenent del territori consideren uns criteris o uns altres [15]. En la majoria de casos s'utilitza la PTOG però hi ha variacions en la dosi administrada i els llindars de diagnosis establerts [16]. Les principals organitzacions que han proposat estratègies són:

- ACOG (American College of Obstetricians and Gynecologists)
- ADA (American Diabetes Association)
- CDA (Canadian Diabetes Association)
- GEDE (Grupo Español de Diabetes y Embarazo)
- OMS (Organització Mundial de la Salut)

- NICE (National Institute for Health and Clinical Excellence
- SIGN (Scottish Intercollegiate Guidelines Network)

Les principals diferències en les estratègies proposades per les diferents institucions són: si el cribratge ha de ser universal per totes les embarassades o només per les que presenten factors de risc ja que la prevalença de la malaltia és baixa; la setmana en la que s'ha de fer la prova; el protocol més adequat i els llindars de detecció [17].

El test més acceptat i utilitzat internacionalment és la PTOG amb 75 o 100g de glucosa segons l'estratègia. Depenent de cada organització s'utilitzen dos tipus de procediments, d'un o dos passos. L'estratègia de dos passos consisteix en començar amb la prova de O'Sullivan per fer un cribratge inicial [17], aquest test consisteix en administrar oralment 50g de glucosa pel matí sense necessitat de fer dejuni, evitar l'activitat física o els hidrats de carboni els dies d'abans [12]. Si el resultat és positiu fer una prova de tolerància oral a la glucosa amb 75 o 100g i mesurar la glucèmia basal més el valor a 1, 2 o 3 hores després. Per altra banda, en el procediment d'un pas només es realitza en dejuni una PTOG amb 75g de glucosa i si compleixen els criteris per la diabetis mellitus de la OMS es considera un test positiu de DMG [17]. A la Taula 3 s'indiquen les característiques de les propostes de les diferents organitzacions.

Taula 3. Tipus de d'estratègia les organitzacions. Font: [17]

	Cribratge		Procediment	
Organització	Universal	Selectiva	1 pas	2 passos
ACOG	X			X
ADA	X		X	X
CDA	X			X
GEDE	X			X
OMS	X		X	
NICE		X		X
SIGN		X		X

Aquestes institucions han anat supervisant i modificant els seus propis punts de tall a partir de diferents estudis que s'han fet al llarg del temps. Per poder arribar a un acord internacional i determinar el llindar de detecció més efectiu es va realitzar l'Estudi d'Hiperglucèmia i Resultats Adversos de l'Embaràs (HAPO), basant-se en l'aparició d'afectacions de l'embaràs [16], va determinar l'existència d'una alta relació lineal i contínua entre la glucèmia i certes complicacions obstètriques com la preeclàmpsia en la mare i

macrosomia, distòcia i hipoglucèmia en l'infant, fins i tot amb valors de glucosa considerats fisiològics en l'embaràs [15].

A partir dels resultats de l'estudi HAPO es va concloure que l'efecte del nivell de sucre en sang de la mare sobre l'aparició de dolències relacionades és un fenomen biològic. A banda, també es va veure que construir criteris de diagnòstic per la DMG és un procés molt complicat si es vol fer a partir de relacions significatives entre el nivell de glucosa en sang matern i els resultats obstètrics. En conseqüència, es va crear l'Associació Internacional de Grups d'Estudi de Diabetis i Embaràs (IADPSG) per coordinar i enfocar internacionalment la investigació i l'educació en el camp de la diabetis gestacional i millorar l'atenció sanitària a les embarassades que pateixen la malaltia [16].

Els grups d'experts confirmen que hi ha dades que recolzen cadascuna de les propostes de les diferents organitzacions encara que les recomanacions siguin contradictòries i la decisió de l'estratègia ha de tenir en compte altres paràmetres com la relació cost-benefici o les infraestructures disponibles de cada regió [15]. A continuació, detallem els procediments de la IADPSG, la ADA i la GEDE.

2.2.5.2 Estratègia proposada per la IADSPG

A la primera visita amb el metge es realitzarà la prova de cribratge:

1. Es considera pregestacional si compleix algun dels següents criteris:
   - Glucosa plasmàtica en dejuni ≥ 126 mg/dL
   - HbA_1c 6,5%.
   - Glucèmia a l'atzar ≥200 mg/dL

2. Es considera DMG si:
   - 126mg/dl ≥ Glucosa plasmàtica en dejuni ≥ 92 mg/dL

3. Si la glucèmia en dejú ≤ 92 mg/dL entre les setmanes 24 i 28 es realitzarà una SOG de 75g prenent mostres en dejuni, 1 i 2 hores i es considera DMG si es compleix algun valor del següent criteri:
   - Glucosa plasmàtica en dejuni ≥ 192 mg/dL
   - 1h ≥ 180 mg/dL
   - 2h ≥ 153 mg/dL [10]

2.2.5.3 Estratègia proposada per la ADA

Seguint les indicacions de la IADPSG, al 2015 l'Associació de Diabetis Americana va establir dues estratègies aplicables entre les setmanes 24 i 28 de gestació [15].

1. Estratègia d'un pas: es mesura la glucèmia en dejú, 1 i 2 hores després d'una PTOG de 75 g de glucosa anhidra. Es considera la DM quan:

   - Glucèmia en dejú ≥92 mg/dL
   - Glucèmia 1h PTOG ≥180 mg/dL
   - Glucèmia 2h PTOG ≥153 mg/dL [2]

2. Estratègia de dos passos: Primer es realitza una PTOG amb 50g de glucosa anhidra en qualsevol moment del dia i es determina el valor de la glucèmia 1 hora després de la càrrega, aquesta prova es realitza a totes les embarassades entre les setmanes 24 i 28, o a la primera consulta si la pacient presenta factors de risc de diabetis gestacional [18]. Si el resultat del test és superior o igual a 140mg/dL, es torna a fer la prova però amb 100 g. Finalment s'accepta el diagnòstic si al menys dues mesures son iguals o

superiors als criteris de Carpenter/Coustan i de la National Diabetes Data Group [2]. Aquests llindars s'indiquen a la Taula 4.

Taula 4. Criteris pel diagnòstic de la diabetis mellitus gestacional de Carpenter/Coustan i la National Diabetes Data Group. Font: [2]

	Carpenter/Coustan	National Diabetes Data Group
Dejú (mg/dl)	95	105
1 hora (mg/dl)	180	190
2 hores (mg/dl)	155	165
3 hores (mg/dl)	140	145

2.2.5.4 Proposta del Grup Espanyol de Diabetis i Embaràs (GEDE)

La primera prova és el Test de O'Sullivan amb 50g de glucosa. Un cop la dona pren la solució, es mesura la glucèmia basal amb una extracció sanguínia venosa i una hora després, es torna a avaluar la glucèmia en sang. Normalment el cribratge es fa entre la setmana 24 i 28 de gestació, però si la dona presenta factors de risc es fa durant el primer trimestre [12]. El Test de O'Sullivan té una sensibilitat del 80% i una especificitat del 87% i presenta pocs falsos negatius i com molts mètodes de cribratge molts falsos positius, per aquesta raó després és necessari fer una PTOG per confirmar el resultat [19]. Per tant, si la segona extracció és ≥140 mg/dL es considera un valor anormal i es realitza una SOG amb 100g de substrat diluït en 250 ml de líquid.

En aquest segon examen és necessari estar entre 8 i 14 hores en dejú i ha de menjar els dies anteriors mínim 150g d'hidrats de carboni diàriament. Es fan quatre extraccions, al moment de prendre la glucosa i cada hora durant les 3 següents. Els llindars que s'utilitzen són els següents:

- Basal ≥ 105 mg/dL
- 1 hora ≥ 190 mg/dL
- 2 hores ≥165 mg/dL
- 3 hores ≥145 mg/dL [12]

Si al menys dues de les mesures compleixen el criteri s'accepta l'existència de la malaltia, en canvi, si només hi ha un valor anormal es considera que hi ha intolerància als carbohidrats i es recomana que després de 3 setmanes es torni a repetir el procediment [12]. Per un altre costat, si per alguna raó no es va fer la prova en el segon trimestre o va donar negatiu i durant el tercer període s'han desenvolupat complicacions associades a la DMG, per exemple macrosomia o polihidramnis directament es fa la PTOG amb 100g [19]. A la Figura 3 es mostra l'esquema de tot el procediment.

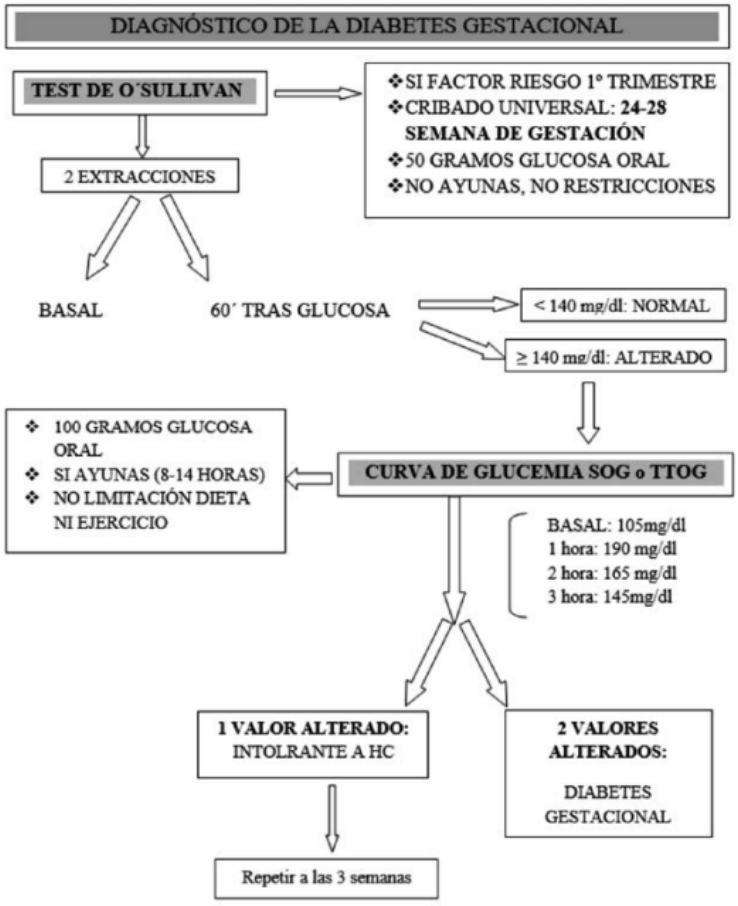

Figura 3. Esquema de l'estratègia de diagnosis de la DMG de la GEDE. Font: [12]

2.2.6 Control i tractament

Un cop es diagnostica la malaltia és imprescindible començar ràpidament un tractament per aconseguir l'equilibri glucídic. Aquesta teràpia consisteix en establir unes pautes nutricionals, exercici físic, controlar l'augment de pes i monitoritzar el nivell de glucosa [10].

A la següent llista s'indiquen els objectius metabòlics que ha de complir una embarassada amb qualsevol trastorn diabètic:

- Glucèmia capil·lar abans de menjar < 95 mg/dL
- Glucèmia capil·lar 1h després de menjar < 140 mg/dL
- Setmanalment glucèmia mitja entre 80 i 100 mg/dL
- Hemoglobina (Hb) glicosilada normal
- Que no hi hagi hipoglucèmies
- Que no hi hagi cetonúria principalment al matí després del dejuni nocturn [12]

Molts assajos han demostrat que entre el 70 i 80% de les dones poden reduir la possibilitat de desenvolupar DMG si durant el primer i el segon trimestre milloren el seu estil de vida, fent activitat física i amb una dieta adequada [10].

2.2.6.1 Dieta

Les necessitats energètiques augmenten a partir del segon trimestre i per evitar l'augment excessiu del pes gestacional i proveir els nutrients necessaris tant al fetus com la

mare, es segueixen unes pautes nutricionals [10]. Es recomana distribuir els àpats i les calories durant tot el dia, amb un increment de 300kcal/dia respecte la ingesta pregestacional. En dones amb obesitat l'increment de calories diàries es redueix a 100 kcal i en cas que tinguin un índex de massa corporal (IMC) superior a 27 kg/m^2, la ingesta diària total es recomana que sigui inferior a 25kcal/k per evitar que augmenti el seu pes [15].

Les guies indiquen una dieta normocalòrica, no restrictiva i equilibrada d'aproximadament 30-35 kcal/kg al dia i un augment de pes total de l'embaràs d'entre 9 i 10 kg. Segons el pla nutricional la dieta ha de ser:

- 20% proteïna.
- 30% greixos: es recomanen els monoinsaturats i disminuir el consum de colesterol, greixos saturats i poliinsaturats.
- 50-55% carbohidrats: principalment d'absorció lenta. S'aconsella restringir els hidrats de carboni d'acció ràpida, per exemple, sucres, refrescs o la brioixeria industrial.
- Fibra vegetal per ajudar a l'absorció intestinal i el buidament gàstric [12].

2.2.6.2 Exercici físic

L'activitat física aeròbica moderada aporta beneficis per tots els embarassos ja que s'ha vist que serveix per prevenir l'aparició de la diabetis gestacional i millorar la sensibilitat a la insulina. La ACOG recomana fer cada dia almenys 30 minuts d'exercici [10], per exemple anar a passejar [12], restringint els moviments que promouen la contracció de la musculatura abdominal [10]. L'esport promou l'augment de consum de glucosa per tant millora el control glucídic, això s'aconsegueix ja que hi ha un augment de transportadors de glucosa sensibles a la insulina (GLUT4), en conseqüència augmenta la vascularització dels teixits i disminueix la quantitat d'àcids grassos lliures i de greix intraabdominal i finalment, augmenta la sensibilitat a la insulina [12].

2.2.6.3 Fàrmacs

Només amb els canvis en la dieta i l'exercici físic la majoria de pacients aconsegueixen mantenir el control glucídic en valors adequats. Per contra, les dones amb diabetis gestacional més agressiva (30-40%) també necessiten un tractament farmacològic per poder dominar la malaltia.

Actualment, l'únic fàrmac aprovat per l'Administració d'Aliments i Medicaments (FDA en anglès) per utilitzar-lo per controlar el nivell de glucosa en embarassades és la insulina [10], ja que no traspassa la barrera placentària i és segur tant per la mare com pel fetus. A partir de tècniques de DNA recombinant es reprodueix la mateixa seqüència d'aminoàcids i es creen les solucions que simulen la insulina humana natural. Tenen una capacitat antigènica baixa, per tant, no solen provocar reaccions patològiques com al·lèrgies o lipodistròfies. N'hi ha de tres tipus segons la velocitat d'acció: ràpida, intermèdia i prolongada, però les utilitzades per embarassades són les d'acció ràpida i intermèdia ja que l'última presenta més variabilitat en l'absorció i la funció [12].

L'Australian Carbohydrate Intolerance Study in Pregnant Women (ACHOIS) va comparar els resultats perinatals entre grups control sense teràpia amb grups de dones a les que s'estava fent el tractament amb insulina i va demostrar que amb el fàrmac va disminuir la morbiditat i la mortalitat dels fills [15].

Uns fàrmacs que tenen una acció molt semblant són els anàlegs d'insulina que s'han desenvolupat en els últims anys. Aquests són de temps d'acció més curt i funcionen com a màxim 1h després d'haver menjat, per tant, són de gran utilitat per controlar la hiperglucèmia postprandial [12].

Hi ha controvèrsia en l'ús d'altres tipus de medicines. La FDA ha classificat algunes propostes en les categories farmacològiques de l'embaràs B (glibenclamida, metformina, acarbosa) i C (altres sulfonilureas, glitazones, meglitinidas). Tot i això, hi ha estudis de l'última dècada que recolzen l'ús de fàrmacs antidiabètics que poden ser una alternativa a la insulina segura, eficaç, confiable, econòmica i amb risc baix per la mare i el bebè i que ha estat ben acceptada per les dones implicades [10]. A la Taula 5 es descriuen els nivells de la classificació del risc dels fàrmacs en l'embaràs segons l'Administració d'Aliments i Medicaments.

Taula 5. Classificació de la FDA de categories de risc a l'embaràs. Font: [20]

Categoria	Descripció
A	Estudis adequats i ben controlats en dones embarassades no han mostrat un risc augmentat d'anormalitats fetals.
B	Estudis en animals han mostrat que no hi ha evidència de mal sobre el fetus. Però, no hi ha estudis adequats i ben controlats sobre dones embarassades.
C	Estudis en animals han mostrat algun efecte advers i no hi ha estudis adequats i ben controlats en dones embarassades.
D	Estudis adequats i ben controlats o estudis observacionals en dones embarassades han demostrat algun risc pel fetus; tot i això, els beneficis del tractament són superiors als riscs potencials.
X	Estudis adequats i ben controlats o estudis observacionals en animals o dones embarassades han demostrat que produeixen anormalitats fetals. L'ús d'aquests productes està contraindicat en dones embarassades.

A la Fourth International Workshop-Conference on Gestational Diabetes (1997) es va recomanar començar el tractament si la glucosa plasmàtica és >95mg/dL, >140 mg/dL, >120mg/dL; en dejuni, 1 hora i 2 hores després de la ingesta respectivament. Per una altra banda, algunes característiques associades a la mida fetal també condicionen l'inici de la teràpia amb insulina.

La dosi inicial és 0,2 UI de protamina neutra de Hagedorn (NPH), que és una insulina d'acció intermèdia, per cada kg de la dona, pel matí o la nit depenent del moment on la hiperglucèmia és màxima. En cas de que sigui necessari s'afegeix una altra dosi en moments puntuals d'insulina d'acció ràpida. En total es fan entre 50 i 90 injeccions durant tot el tractament depenent de les característiques pes o la ètnia de la mare i l'estat del fetus [12].

2.2.7 Prevalença

A Espanya el percentatge de dones amb DMG es troba entre el 7,6 i el 10,6% [18]. La prevalença depèn del criteri que s'utilitza, per tant, degut a la falta de consens en els mètodes de cribratge es desconeix amb exactitud la prevalença mundial de la diabetis gestacional. Segons l'estimació de l'Associació de la Diabetis Americana la prevalença mundial està entre l'1 i el 14% de la població estudiada [10].

En els últims anys el percentatge de dones amb DMG ha augmentat considerablement en relació amb l'epidèmia de diabetis mellitus i l'augment de persones amb factors associats a l'aparició de la malaltia, com tenir fills cada cop més tard, la obesitat pregestacional, la hipertensió arterial crònica o tenir familiars de primer grau amb DM. És important dissenyar estratègies de diagnosi personalitzades per la pacient tenint en compte les seves

característiques clíniques i els seus factors de risc associats. Això millorarà els resultats perinatals i disminuirà la prevalença de la DM2 [18]. Per poder construir aquests nous mètodes és imprescindible estudiar els factors relacionats amb l'aparició de la malaltia.

3 Conceptes d'anàlisi multivariant

Existeixen diverses característiques que afecten la possibilitat de patir la diabetis gestacional. En aquest treball utilitzem l'entorn del llenguatge de programació R per analitzar un conjunt de més de 1000 variables a partir de les quals construirem un model ens permet classificar si una pacient té o no la malaltia. Abans, és convenient conèixer el marc teòric de l'anàlisi multivariant i les tècniques matemàtiques amb les que treballarem.

3.1 Introducció

L'estadística és una part de les matemàtiques que serveix per obtenir informació d'un grup d'individus, en un espai i temps determinats, a partir de mesures que s'han fet sobre aquests. Les dades s'emmagatzemen en forma de variables per poder ser analitzades, segons el número d'atributs observats sobre cada unitat de la població, parlem d'estructures de dades univariants si només s'estudia una variable, o multivariant en cas de que n'hi hagi múltiples. Segons l'interès de l'investigador, aquestes estructures es poden utilitzar per trobar associacions dins d'un mateix conjunt de casos o entre diferents poblacions [21].

La branca de l'estadística que proporciona els mètodes necessaris per estudiar les relacions entre una gran quantitat de mesures és l'anàlisi multivariant (AM) [22]. Aquestes tècniques serveixen per poder explicar una realitat complexa a partir de variables indicadores més senzilles, conegudes com factors. Els seus objectius són:

1. Resumir i descriure les dades originals estudiades, a partir d'un número inferior de noves variables que s'han obtingut a partir de modificacions de les primeres perdent la mínima informació possible.
2. Separar-les en diferents grups representatius.
3. Poder classificar noves mesures en els grups trobats.
4. Trobar relacions entre dos conjunts de variables [23].

Segons el tipus d'investigació i l'enfoc que posen els autors trobem diferents definicions a la literatura. De manera general l'AM engloba totes les tècniques estadístiques que permeten analitzar més de dues variables, alguns consideren que el punt important és que les variables han de ser aleatòries i han d'estar relacionades entre elles, de manera que no es poden estudiar els seus efectes de manera independent. Altres consideren que es basen en combinacions lineals obtingudes a partir de sumes de de les variables originals ponderades i permeten descriure la relació de forma matemàtica [24].

Per treballar amb l'AM habitualment trobarem les dades representades en una matriu $\mathbf{X}$ de dimensió $(n \times p)$ on p correspon al número de variables escalars, també conegudes com univariants, que han sigut observades sobre els *n* elements de la població. Per un altre costat, el conjunt de variables de cada cas formen el que es coneix com la variable vectorial o multivariant. Les dades es poden escriure en diferents estructures com s'indica a les Equacions (1), (2) i (3) [25].

- En tota la matriu:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \quad (1)$$

- Per files: on $\mathbf{x}'_i$ és un vector $(p \times 1)$ que conté les mesures de totes les variables de la instància i.

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}'_1 \\ \vdots \\ \boldsymbol{x}'_p \end{bmatrix} \quad (2)$$

- Per columnes: on $\mathbf{x}(\mathrm{j})$ és un vector $(1 \times n)$ que conté els valors de l'atribut escalar j per cadascun els objectes de la població.

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_{(1)} & \dots & \boldsymbol{x}_{(n)} \end{bmatrix} \quad (3)$$

3.2 Tipus de variables

3.2.1 Segons l'escala numèrica

Segons el tipus d'escala que s'ha utilitzat per quantificar les observacions les variables es classifiquen en:

- Mètriques o quantitatives: s'expressen en forma de valor numèric.
  - Contínues o d'interval: qualsevol valor real dins d'un interval.
  - Discretes: només valors enters.

- No mètriques o qualitatives: s'expressen en forma de categoria, poden mantenir un ordre entre elles o no [22].
  - Binaries o dicotòmiques: estan limitades a dues etiquetes.
  - Generals: qualsevol etiqueta és possible [25].

3.2.2 Classificació segons el rol

Segons el paper en la relació amb la resta es classifiquen en:

- Variables independents (v.i.): dins dels experiments són aquelles situacions a les que són sotmesos els individus, si parlem d'un estudi observacional són els atributs que els diferencien. També s'anomenen variables predictores o explicatives.

- Variables dependents (v.d.): són les variables de les que es pot predir el seu comportament a partir del valor d'un grup de v.i.. També s'anomenen variables criteri o resposta.

El rol no sempre és fàcil de determinar, ja que depèn del tipus d'investigació i de les seves característiques i en alguns casos, segons la situació que s'estigui analitzant pot ser l'una o l'altra [22].

3.3 Anàlisi descriptiva multivariant

3.3.1 Mesures de distribució

En aquest apartat es descriuran els diferents tipus de mesura que serveixen per veure de quina manera estan distribuïdes les dades respecte el centre.

3.3.1.1 Mesures de centralització: El vector de mitjanes

Considerant $\mathbf{X}$ la matriu de dades, el vector $\bar{\mathbf{x}}$ recull la mitjana de cadascuna de les variables observades, vegeu l'Equació (4). És una mesura de centralització ja que es busca que la suma de les desviacions sigui 0 i en conseqüència, estarà al centre de les dades. Si a les observacions originals restem les mitjanes s'obté la matriu de dades centralitzades $\widetilde{\mathbf{X}}$, representada a l'Equació (5).

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} = \frac{1}{n}\boldsymbol{X}'\mathbf{1} \qquad (4)$$

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\overline{\boldsymbol{x}}' = \boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X} \qquad (5)$$

Per un altre costat, la mediana també és un valor important que ens dona informació sobre la centralitat de les mostres. A l'anàlisi descriptiu univariant, si la mitjana i la mediana d'un conjunt són molt diferents això pot indicar que són dades poc homogènies, asimètriques o que hi ha valors atípics, per aquesta raó és útil calcular aquestes dues mesures a l'estudi inicial. En el cas multivariant, les dades no tenen un ordre natural per tant encara que calculem el vector de les medianes aquestes no estan obligades a ser el centre [25].

3.3.1.2 Coeficients d'asimetria i kurtosis

El coeficient d'asimetria A_p determina la falta de simetria de les dades, és a dir, la diferència de la distribució entre la esquerra i la dreta del centre. Per tant, com més a prop sigui de 0 la funció de distribució serà més simètrica.

Per un altre costat, el coeficient de kurtosis K_p serveix per mesurar la relació entre la variabilitat de les desviacions i la de la mitjana. Geomètricament representa si la funció de distribució és més o menys punxeguda, quan hi ha valors atípics augmenta la variabilitat, en conseqüència el coeficient de Kurtosis serà alt i tindrà més forma de pic, mentre que si no hi ha variabilitat la distribució serà més plana.

La distància euclidiana mesura la distància entre dos elements i j , i es calcula com s'indica a l'Equació (6). En estudis multivariants, els coeficients d'asimetria i kurtosis es poden calcular a partir de les distàncies d_{ij},vegeu les Equacions, (7) i (8) [25].

$$d_{ij} = \sqrt{\left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2} \qquad (6)$$

$$A_p = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^3 \qquad (7)$$

$$K_p = \frac{1}{n}\sum_{i=1}^{n} d_{ii}^2 \qquad (8)$$

3.3.2 Mesures de dispersió

3.3.2.1 La matriu de variàncies i covariàncies

La variabilitat de les variables normalment és mesurada per la variància o per la desviació estàndard, i la covariància és el paràmetre que determina la dependència lineal entre dues variables escalars. Les característiques de la matriu de variàncies i covariàncies ($\mathbf{S}$) són les següents: en primer lloc és quadrada i simètrica, en segon lloc, trobarem indicades les variàncies a la diagonal i les covariàncies a la resta de la matriu [25]. A l'Equació (9) es descriu la matriu $\mathbf{S}$.

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(x_i - \bar{x})' = \begin{bmatrix} {s_1}^2 & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & {s_p}^2 \end{bmatrix} = \frac{1}{n}\mathbf{1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}} \qquad (9)$$

Per evitar el biaix de l'estimador de la matriu de població, en alguns llibres es divideix la matriu de covariàncies per $n - 1$ ja que aquest és el número real de desviacions independents [25].

3.3.3 Mesures de dependència lineal

3.3.3.1 Regressió per parells: Matriu de correlació

El coeficient de correlació lineal r és un valor que representa el nivell de dependència lineal entre dues variables expressat en un rang entre 0 i 1. Aquest es pot utilitzar per determinar el grau de relació per parells de totes les variables analitzades i representar-les a la matriu de correlació $\mathbf{R}$. A l'Equació (10) s'indica l'expressió matemàtica del coeficient r entre les variables x_j i x_k.

$$r_{jk} = \frac{s_{jk}}{s_j s_k} \quad ; \quad 0 \le |r_{jk}| \le 1 \qquad (10)$$

La matriu $\mathbf{R}$ és quadrada i simètrica, té uns a la diagonal que representen la relació entre una variable i ella mateixa, per l'altre costat a la resta d'espais trobarem els coeficients entre els diferents atributs [25].

3.3.3.2 Regressió múltiple

La regressió múltiple serveix per determinar les relacions d'una variable amb la resta. La variable que volem estudiar respecte la resta es la variable resposta y, i la resta són les explicatives o regressores x . El model matemàtic de predicció de la variable explicativa està representat a l'Equació (11).

$$\hat{\boldsymbol{y}}_{\boldsymbol{1}} = \bar{y} + \hat{\beta}_1(x_{i1} - \bar{x}_1) + \ldots + \hat{\beta}_p(x_{ip} - \bar{x}_p), \qquad i = 1, \ldots, n \tag{11}$$

Els coeficients $\hat{\beta}$ es defineixen a partir de l'error de l'equació de predicció buscant que aquest sigui el més pròxim a 0. La manera de calcular aquest error és sumant el quadrat de les diferències entre els valors originals i els predits $e_i = y_i - \hat{y}_1$, també coneguts com residus. S'utilitza la tècnica dels mínims quadrats per eliminar la condició del signe dels residus.

Si a partir d'un subconjunt de dades centralitzades $\widetilde{\mathbf{X}}$ determinem una matriu $\mathbf{X_R}$ de dimensió $(n \times p-1)$. Podem expressar l'equació de regressió múltiple entre la variable resposta $y = x_j$ i la resta de variables x_k a partir de la matriu de covariàncies sense la columna referent a la variable resposta, $\mathbf{S_{p-1}^{-1}}$, multiplicada per la columna eliminada $\mathbf{S_{xy}}$, vegeu l'Equació (12).

$$\hat{\beta} = (\mathbf{X_R'}\,\mathbf{X_R})^{-1}\mathbf{X_R'}\mathbf{y} = \mathbf{S_{p-1}^{-1}}\mathbf{S_{xy}} \tag{12}$$

Aquesta mesura ens ajudarà a veure quines són les variables que poden ser descrites a partir de la resta, és a dir, quines són més predictibles i redundants [25].

3.3.4 La distància de Mahalanobis

La distància entre dos punts és una mesura útil per estudiar la variabilitat entre les observacions d'un conjunt de dades. Hi ha diverses maneres per calcular la distància entre una observació escalar i la mitjana de la variable, tal com s'indica a l'Equació (13).

$$d = \sqrt{(x_i - \bar{x})^2} = |x_i - \bar{x}| \tag{13}$$

La mitjana dels valors de les distàncies entre els diferents punts i el valor mitjà correspon a la variància de l'atribut. Si generalitzem l'expressió anterior pel cas multivariant definim la distància com la mitjana de cada punt de la variable vectorial i el vector de mitjanes d'aquesta, quan ens referim a aquest últim concepte parlem de la distància de Mahalanobis (MSD), la seva expressió està definida a la l'Equació (14).

$$d_i = \sqrt{(\mathbf{x_i} - \bar{\mathbf{x}})'\mathbf{S^{-1}}(\mathbf{x_i} - \bar{\mathbf{x}})} \tag{14}$$

Cal destacar que podem trobar que alguns autors per simplificar les descripcions denominen amb el mateix nom al quadrat de la distància d_i^2 [25]. Aquest concepte, ens indica la distància dels punts amb el centre de massa de la variable, considerant com a valors atípics

les dades que es trobin més lluny. És important detectar els outliers per poder eliminar-los, ja que poden afectar les mesures de distribució, dispersió i correlació, per tant, modifiquen els resultats de l'estudi.

Poden aparèixer errors causats per conjunts de valors atípics, per un costat pot ser que condueixin la distribució i la variància cap a la seva direcció i en conseqüència disminueixi la MSD. Per l'altre costat, pot augmentar la distància de les dades correctes si el conjunt atrau el vector de mitjanes i aparta la variància de la distribució de dades que no són outliers [26].

3.4 Classificació de les tècniques multivariants

Per escollir la millor tècnica d'anàlisi s'ha de tenir en compte el tipus d'investigació, els seus objectius, els tipus de dades que s'estudiaran i si es busca trobar relacions entre variables o entre casos de la població. Per exposar de manera global els mètodes més utilitzats, en aquesta memòria utilitzem la classificació proposada el 1998 per Hair, Anderson, Tatham i Black que es basa en agrupar les tècniques basades en models lineals segons el tipus de relació i l'escala [24].

1. Relació de dependència: es busca predir o explicar l'efecte de les v.i. a partir d'un conjunt de v.d [24].

    - 1 variable dependent
        - Quantitativa:
            - Anàlisi de regressió múltiple
            - Anàlisi conjunt
        - Qualitativa:
            - Anàlisi discriminant
            - Anàlisi conjunt
            - Regressió logística
            - Models lògit

    - Més d'una variable dependent
        - Quantitativa amb una única relació amb
            - Variables independents quantitatives: Anàlisi de correlació canònica.
            - Variables independents qualitatives: Anàlisi de variància multivariant: MANOVA

    - Moltes variables dependents i independents quantitatives amb relacions múltiples: Models d'equacions estructurals

2. Relació d'interdependència: l'objectiu és analitzar a la vegada totes les variables sense tenir en compte si són dependents o independents per trobar un model que representi el conjunt de variables o objectes [24].

    - Relacions entre variables:
        - Anàlisi de components principals
        - Anàlisi factorial
        - Anàlisi de conglomerats

    - Relacions entre casos: Anàlisi de conglomerats

- Relacions entre objectes:
  - Mesurats de forma quantitativa: Escalat multidimensional
  - Mesurats de forma qualitativa: Anàlisi de correspondències

Hi ha molts tipus de mètodes, cadascun dissenyat per resoldre problemes concrets i depenent de l'objectiu de la investigació s'escull un o altre. Al nostre estudi, en primer lloc reduïm el número de variables amb un Anàlisi de Components Principals per poder treballar amb una dimensionalitat més manejable, amb una pèrdua mínima d'informació. Després construïm el model matemàtic: un classificador del grup d'observació de les pacients que classifica en dones control o amb DMG. La variable resposta és qualitativa, per tant decidim que utilitzarem un model de regressió logística.

3.5 Anàlisi de Components Principals (ACP)

3.5.1 Autovalors i autovectors

En aquest apartat introduirem els conceptes d'autovectors i autovalors d'una matriu quadrada. Depenent de la bibliografia els autovalors es defineixen com a valors propis, característics o arrels latents i els autovectors com vectors propis, característics o latents. En una matriu $\mathbf{A}$ d'ordre p, si es compleix que quan és multiplicada per un vector $\mathbf{x}$ el resultat és aquest mateix vector multiplicat per un escalar λ , real o complex, $\mathbf{x}$ i λ representen respectivament un autovector i un autovalor d'A, tal com mostra l'Equació (15).

$$\boldsymbol{Ax} = \lambda \boldsymbol{x} \qquad (15)$$

Per calcular els autovalors s'utilitza l'equació característica d'$\mathbf{A}$, vegeu l'Equació (16), en la que els resultats són p arrels $(\lambda_1, \ldots, \lambda_p)$. Per exemple, si $\mathbf{A}$ és una matriu d'ordre 2, la fórmula consisteix en una equació de segon grau en la que trobarem dos autovalors [27].

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = (a_{11} - \lambda)(a_{22} - \lambda) - (a_{21} \cdot a_{12}) = 0 \qquad (16)$$

Les propietats dels valors propis són les següents:

- El sumatori d'autovalors és igual a la suma de la diagonal principal d'$\mathbf{A}$.
- El productori d'autovalors és igual al determinant d'$\mathbf{A}$.
- Si el determinant d'$\mathbf{A}$ és nul com a mínim un autovalor serà 0.
- El rang d'$\mathbf{A}$ és el número d'autovalors diferents de 0.

Un cop s'han calculat els autovalors de la matriu, per trobar els autovectors respectius d'aquests, es substitueixen les arrels i es busquen els resultats del sistema homogeni de l'Equació (17).

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = 0 \;\; ; \;\; \begin{cases} (a_{11} - \lambda)x_{11} + a_{12}x_{21} = 0 \\ a_{21}x_{11} + (a_{22} - \lambda)x_{21} = 0 \end{cases} \qquad (17)$$

Si $\mathbf{A}$ és una matriu simètrica els autovectors són ortogonals. Considerant $\mathbf{\Lambda}$ la matriu diagonal que conté les arrels latents i $\mathbf{U}$ la matriu de vectors característics normalitzats es compleixen les propietats de l'Equació (18) [27].

$$\mathbf{A} = \mathbf{U\Lambda U'} \rightarrow \ \mathbf{A^{-1}} = \mathbf{U\Lambda^{-1}U'} \rightarrow \mathbf{\Lambda} = \mathbf{UAU'} \qquad (18)$$

3.5.2 Introducció a l'ACP

L'Anàlisi de Components principals, Principal Component Analysis (PCA) en anglès, és una tècnica multivariant que pot ser utilitzada amb diferents finalitats: per transformar un grup de variables correlacionades en un altre d'independents, per trobar dins d'un conjunt de variables combinacions lineals amb molta o poca variabilitat, o per reduir el número de variables de manera que es reprodueixi idènticament la matriu de covariàncies.

Les components principals (CP) són combinacions lineals de les variables del conjunt original, estan incorrelades i definides jeràrquicament de tal manera que començant per la primera expliquen de més a menys la variància de les dades. Aquestes es calculen a partir dels autovalors i els autovectors normalitzats de les matrius de covariàncies o correlacions, els valors propis representen la variància i els vectors característics la direcció en que la variabilitat dels components és màxima [28].

A la Figura 4, es mostra l'exemple d'un ACP bidimensional en el que tenim un conjunt de n mostres amb mitjana 0 de manera que l'origen es troba al centre de gravetat i cadascun dels punts està definit pel valor de les variables x_1 i x_2. El que és pretén es trobar una combinació lineal de les variables que expressi la major part de la informació, en unes altres paraules, una direcció en la que la variància és màxima quan els punts es projecten sobre el nou eix. Aquesta és la denominada la primera component principal, un cop tenim la CP_1 es busca la CP_2, una nova direcció ortogonal a l'anterior que contingui la major part de la variabilitat que queda per descriure.

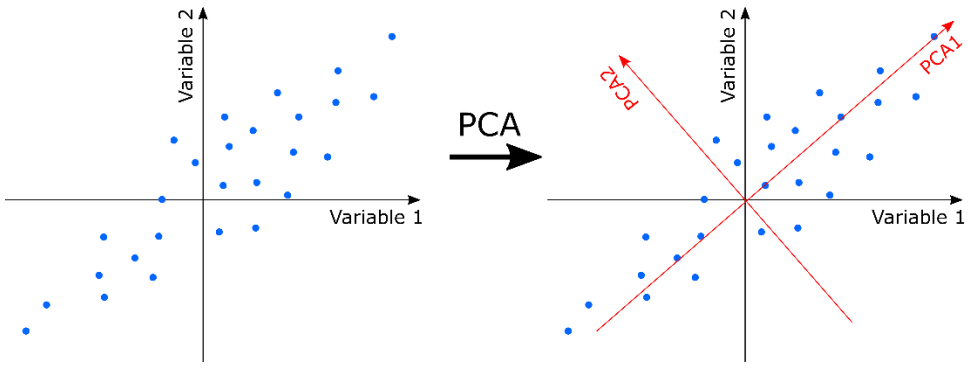

Figura 4. Anàlisi de Components principals d'un cas bidimensional. Font: [29]

Hi ha el mateix número de components principals que d'atributs inicials, després, un cop calculades es seleccionarà el número de CP suficients per descriure les dades amb una pèrdua mínima d'informació. El següent pas consisteix en definir cada punt sobre el nou eix de coordenades i els anomenarem puntuacions sobre les components principals.

De manera general l'expressió matemàtica de les components principals, vegeu l'Equació (19), és igual a la suma lineal de les variables multiplicades pels coeficients de la component principal que són vectors unitaris en la mateixa direcció, restringits en aquesta condició per aconseguir que es compleixi amb la màxima variabilitat. Si Y és una component principal, **v** els coeficients i X les variables:

$$Y_j = \mathbf{v_{1j}}X_1 + \mathbf{v_{2j}}X_2 + \ldots + \mathbf{v_{Pj}}X_P \quad (19)$$

Podem descriure les coordenades de cada observació individualment o de forma matricial com s'indica a les Equacions (20) i (21).

$$y_{ij} = x_{i1}v_{1j} + x_{i2}v_{2j} + \cdots + x_{ip}v_{p1} \quad (20)$$

$$\mathbf{y}_j = \boldsymbol{X}\mathbf{v}_j \quad (21)$$

També es pot expressar de manera general, observeu l'Equació (22), amb Y com la matriu de puntuacions de les components principals, X el conjunt de dades original i V una matriu que conté el vector de coeficients de les components a la columna corresponent [30].

$$\boldsymbol{Y} = \boldsymbol{XV} \quad (22)$$

3.5.3 Càlcul de les components principals

En aquesta secció descriurem un procediment matemàtic per calcular les components principals. En primer lloc, es vol calcular la primera component Y_1 ja que aquesta és la que explica la major part de la variància. Tal com hem vist abans **v** és un vector unitari que està restringit en aquesta condició per aconseguir la maximització de la variància. Seguint la fórmula de la variància descrivim l'Equació (23).

$$\boldsymbol{Var}(\mathrm{Y}_1) = \mathbf{v_1'Sv_1} \quad (23)$$

Utilitzem el mètode de Lagrange per optimitzar la variància. Recordem la fórmula del multiplicador de Lagrange considerant que $f(x,y)$ és la funció $\mathbf{v_1'Sv_1}$ i (x,y) la restricció igualada a 0, a l'Equació (24).

$$\begin{aligned} L(\lambda, x, y) &= f(x,y) + \lambda g(x,y) \\ L(\mathbf{v_1}) &= \mathbf{v_1'Sv_1} - \lambda(\mathbf{v_1'v_1} - 1) \end{aligned} \quad (24)$$

Finalment, si derivem respecte $\mathbf{v_1}$ i igualem a 0 trobem que:

$$\frac{\mathrm{L}(\mathbf{v_1})}{\partial \mathbf{v_1}} = 2\mathrm{Sv}_1 - \lambda 2\mathbf{v_1} = 0$$
$$\mathbf{S}\mathbf{v_1} = \lambda \mathbf{v_1} \qquad (25)$$

En definitiva, a partir de l'Equació (25) es comprova que $\mathbf{v_1}$ és un autovector de la matriu de covariàncies $\mathbf{S}$ i λ és un valor propi associat, tal que:

$$\boldsymbol{Var}(\mathbf{X}\mathbf{v_1}) = \mathbf{v_1'}\mathbf{S}\mathbf{v_1} = \mathbf{v_1'}\lambda\mathbf{v_1} = \lambda \qquad (26)$$

Els autovalors de $\mathbf{S}$ estan ordenats de més gran a més petit, per tant podem interpretar que el primer valor propi de la matriu de covariància és igual a la variància de la primera component sent $\mathbf{v_1}$ el vector associat.

A continuació, es repeteix el mateix procediment per buscar la segona component principal, tenint en compte que CP1 i CP2 són perpendiculars entre elles, i buscant que la suma de les variàncies entre les dues components sigui màxima. Finalment, s'obté una expressió, observeu l'Equació (27), similar a l'Equació (26) en la que es comprova que λ de la segona component és igual al segon autovalor de la matriu de covariàncies i $\mathbf{v_2}$ és el vector associat.

$$\mathbf{S}\mathbf{v_2} = \lambda \mathbf{v_2} \qquad (27)$$

En definitiva, cadascuna de les components principals s'extreuen a partir dels autovalors i autovectors de la matriu S i es compleix l'Equació (28).

$$\mathbf{S} = \mathbf{V}'\boldsymbol{\Lambda}\mathbf{V} \quad \begin{cases} \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_{\mathbf{2}}, \dots, \lambda_{\mathbf{p}}) \\ \mathbf{V}'\mathbf{V} = \mathbf{I} \end{cases} \qquad (28)$$

Per tant, si volem seleccionar les puntuacions $\mathbf{Y}$ només de les q primeres components a partir de la matriu $\mathbf{X}$ centralitzada, utilitzem l'Equació (29) [30].

$$\mathbf{Y_q} = \mathbf{X}\mathbf{V_q} \qquad (29)$$

Si estem analitzant un conjunt en que les escales de les variables són molt diferents, els atributs que tenen valors més alts tindran major efecte per tant el resultat es veuria modificat. Per evitar-ho, es recomana treballar amb la matriu de correlació, és a dir, estandarditzant els valors inicials, d'aquesta manera les components principals estaran condicionades a les correlacions i no a les variàncies de les dades [31].

3.5.4 Càlcul de la bondat de l'ajust

Per mesurar la proporció de variància explicada per cadascuna de les CP, es divideix l'autovalor de la component entre la suma de tots els valors propis de l'anàlisi, observeu l'Equació (30).

$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \quad (30)$$

Per un altre costat, si el que volem és calcular la bondat de tot l'ajust, escollim les q primeres components i es divideix la suma dels autovalors del subconjunt entre el sumatori total, com s'indica a l'Equació (31) [30].

$$\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{i=1}^{p} \lambda_i} \quad (31)$$

3.6 Model de regressió logística.

3.6.1 Model de regressió

Els models de regressió serveixen per determinar la relació entre un conjunt de variables independents x_i amb una dependent anomenada Y, expressada a l'Equació (32).

$$Y = f(x_1, x_2, \ldots, x_i) + \varepsilon \quad (32)$$

Aquest tipus de mètode es pot utilitzar amb intenció explicativa si el que es vol és analitzar com afecta el grup de variables explicatives sobre Y o amb intenció predictiva, si el que es busca és estimar el resultat a partir del conjunt d'atributs independents. Existeixen diferents tipus de models de regressió en els que destaquen el lineal i el logístic ja que són fàcils d'utilitzar i comprendre. S'utilitza un o l'altre depenent del tipus de la variable resposta que volem predir, si aquesta és numèrica contínua normalment es selecciona el model de regressió lineal i en cas de que sigui dicotòmica es fa servir la regressió logística.

Encara que les dues tècniques es construeixen a partir de models matemàtics similars, no és correcte utilitzar el mètode lineal si el que es vol estudiar és una variable qualitativa. En la part pràctica del treball volem determinar si una dona té diabetis gestacional per tant utilitzarem el model de regressió logística, aquest es basa en càlculs probabilístics d'esdeveniments definits, en comptes de determinar el valor real de la variable dependent. D'aquesta manera per construir el model es poden utilitzar tècniques de regressió tradicionals aplicables al lineal, ja que Y no està definida pels nivells de la variable dicotòmica. El model de regressió es construirà sobre la funció de probabilitat de la variable que volem predir i la única diferència amb el lineal serà la interpretació dels resultats [32].

3.6.2 Model de regressió logística

L'equació que segueix el model logístic és una exponencial que pot ser representada com una funció lineal a partir d'una transformació logarítmica. Considerant que volem predir una variable dicotòmica que prendrà el valor 1 amb una probabilitat p i el valor 0 amb

probabilitat $1-p$, si definim aquestes probabilitats amb la funció de distribució logística, vegeu l'Equació (33), tenim: [33]

$$p_i = \frac{1}{1+e^{-\boldsymbol{\beta_0}+\boldsymbol{\beta_1'}\mathbf{x_i}}}$$
$$1-p_i = \frac{1}{1+e^{\boldsymbol{\beta_0}+\boldsymbol{\beta_1'}\mathbf{x_i}}} \quad (33)$$

Les fórmules anteriors es transformen en un model lineal a través del Logit, aneu a l'Equació (34), que és el logaritme de la ratio de l'èxit, en altres paraules: el quocient de la probabilitat de que passi un esdeveniment i la probabilitat de que no passi, que dona lloc a una combinació lineal de les variables independents [34].

$$Logit(p) = log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \quad (34)$$

A l'hora de construir el model, l'objectiu consisteix en estimar els coeficients que determinen l'efecte de cadascuna de les variables independents a través d'un procés iteratiu. [35]. A la Figura 5 és representa el model Logit, en aquest exemple s'observa l'efecte de la variable x, per exemple el pes, sobre la possibilitat d'un esdeveniment, per exemple que una persona pateixi una malaltia. El model es defineix a partir d'individus en que l'atribut que volem predir és conegut, per tant, si estem calculant la probabilitat de la malaltia, la població control es troba a baix a l'esquerra on és 0 i el grup observació a dalt a la dreta on la probabilitat és 1. Cal destacar que es comprova que la variable x és útil per separar les dues poblacions ja que veiem que en un els valors s'agrupen sota del -4 i en l'altre per sobre del 4.

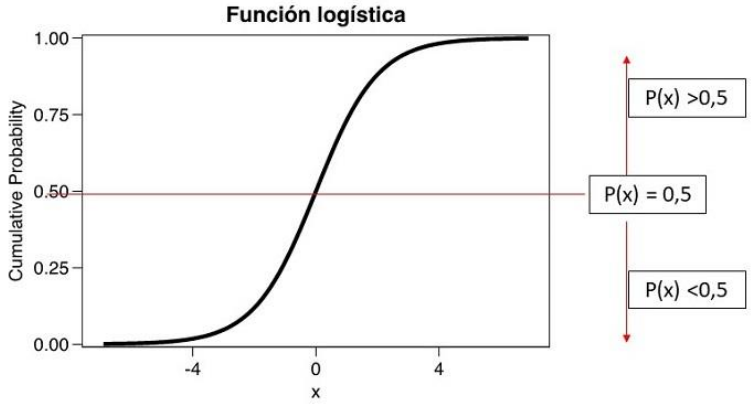

Figura 5. Funció Logit en el cas univariant. Font: [35]

A partir dels individus coneguts es calcula la corba i després quan volem classificar un nou pacient, es fa la observació de x i es busca la intersecció entre la corba del model i l'eix vertical que passa per aquest valor. Finalment, llegim la probabilitat d'aquest punt que correspon a que el pacient tingui la malaltia, en definitiva es diagnostica que té la malaltia si la probabilitat és superior a 0,5.

Si tenim més variables, podem considerar que x_1 és la que major pes té sobre el resultat de l'esdeveniment i que la resta de combinacions lineals modifiquen el coeficient β_1, és a dir, s'ajusta l'efecte de x_1 sobre la possibilitat de l'esdeveniment i en conseqüència millora

l'eficiència de la predicció, en definitiva com més variables s'utilitzin més ajustat estarà el model [34].

4 Anàlisi i modelat de la DMG amb dades reals

4.1 Introducció

Als apartats anteriors, hem exposat els diferents conceptes en que es fonamenta el nostre estudi i que són necessaris per construir el model de predicció de la diabetis gestacional, a partir d'un estudi analític d'un conjunt de dades reals. A continuació, detallem el procediment i els resultats obtinguts de l'anàlisi multivariant d'un conjunt de dades proporcionat per la Dra. Ana Megía de l'Hospital Universitari Joan XXIII de Tarragona. Aquesta base de dades està composta per 1090 variables mesurades sobre 3854 dones embarassades.

Primer prepararem el dataset per poder treballar amb ell, per un costat, eliminem els atributs i els casos amb més número de valors nuls i també, fem una anàlisi de dependències lineals per disminuir la redundància de les variables. Després, calculem la distància de Mahalanobis per detectar els valors atípics que també eliminem. Finalment, seleccionem el 80% de les dades per construir el model i el 20% restant servirà per comprovar l'eficàcia del model, com s'indica a la Figura 6.

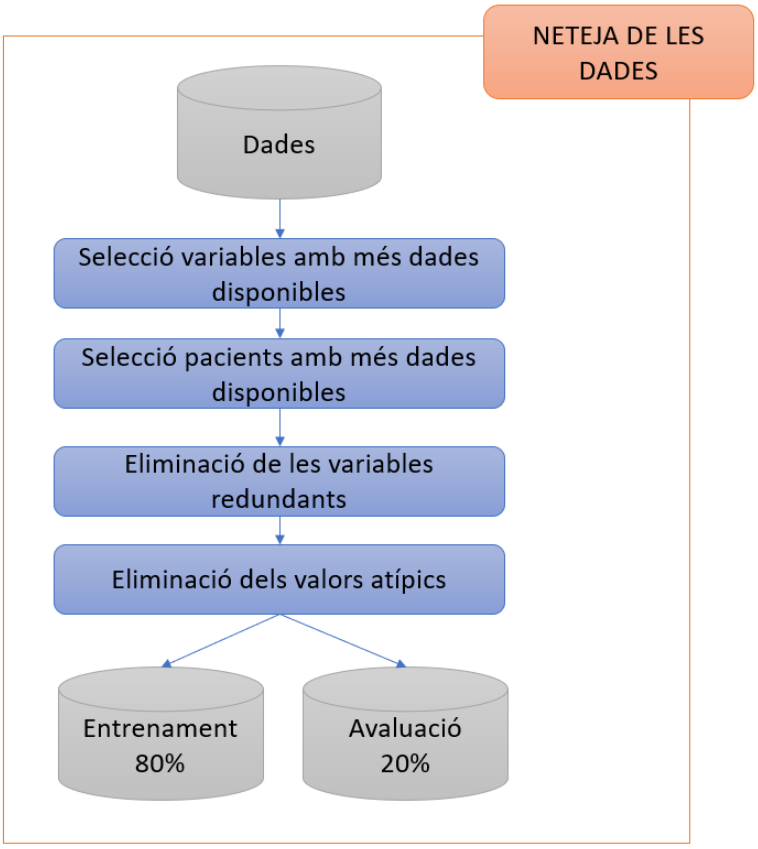

Figura 6. Esquema del procediment de preparació de les dades. Font pròpia.

Seguidament, utilitzem un algoritme per imputar els valors que falten. Després, fem un ACP sobre les dades d'entrenament, actualitzem les observacions dels dos conjunts amb les 4 components principals i construïm el model amb els nous valors d'entrenament, vegeu la Figura 7.

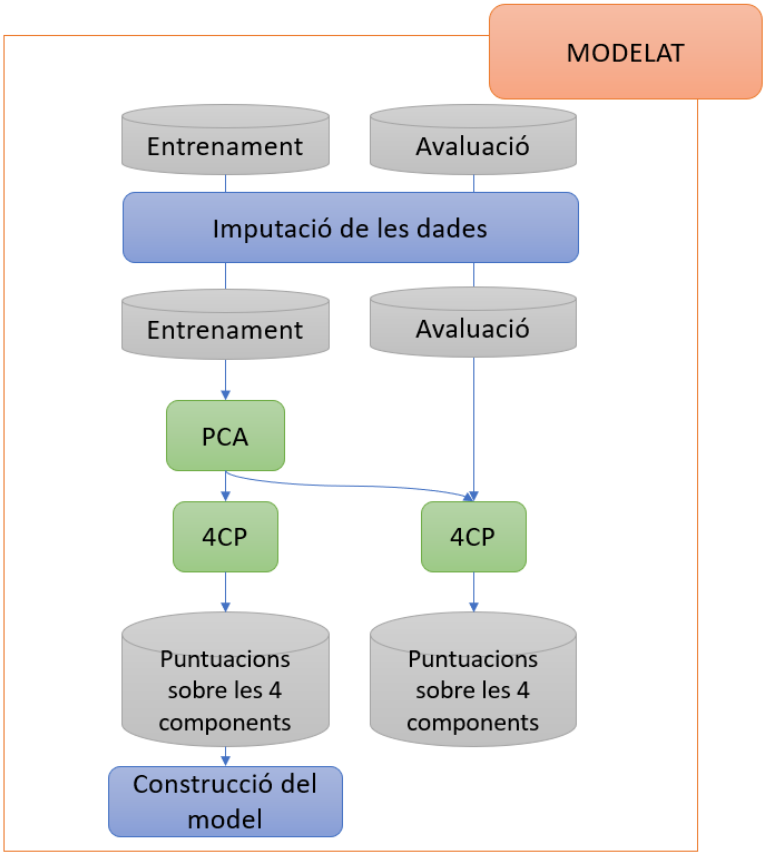

Figura 7. Esquema del procediment de modelat del classificador. Font pròpia.

Finalment, avaluem l'eficàcia tant amb el conjunt de dades d'entrenament com el de test com s'indica a la Figura 8 amb la matriu de confusió i la corba ROC.

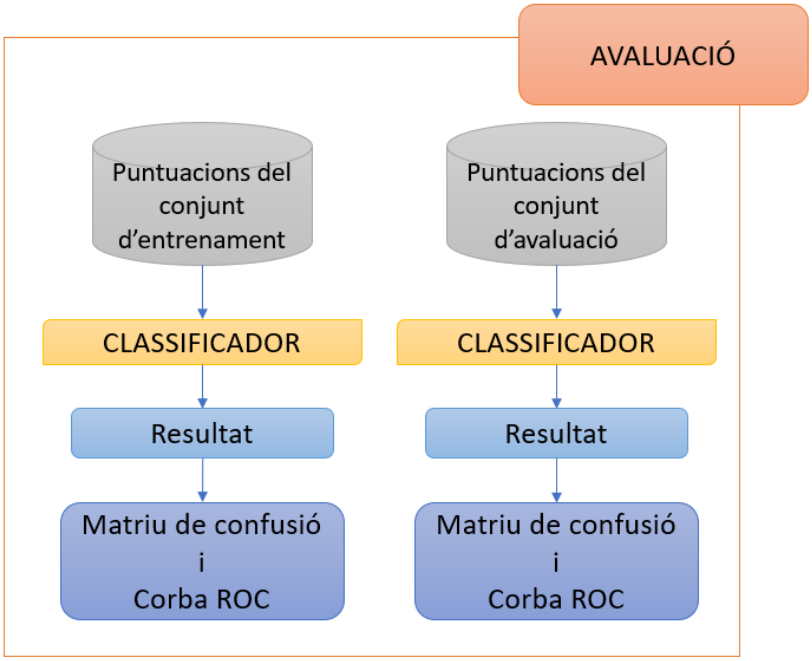

Figura 8. Esquema del procediment d'avaluació del model. Font pròpia.

A l'exploració inicial observem que la variable qualitativa grupo_dg és la que indica el grup d'observació de la dona tal com s'indica a la Taula 6.

Taula 6. Diferents grups d'observació i el número de casos. Font pròpia.

Variable grupo_dg		
Grup	Significat	Número de dones
DG	Pacients amb diabetis gestacional	1992
controles	Pacients control	1202
Intolerante	Intolerància als carbohidrats	515
No corresponde	Desconegut	9
DG por autocontrol	Diagnosis per l'autocontrol de glucèmia capil·lar	35
DG_GLUCEMIA BASAL	Alteració només del nivell de glucèmia basal	4
Diabetes franca	Diabetis preexistent definida a la primera visita	15
ITG	Intolerants a la glucosa	2
9	Error	1
NA	Valor no registrat	79

S'observa que hi ha molts grups diferents, per facilitar l'anàlisi i apropar-nos més al nostre objectiu ens quedarem només amb els casos de DMG normal i els controls, que corresponen al 82,9% de les dades totals. Per una altra banda, un cop hem eliminat els nivells que no necessitem canviem l'etiqueta "controles" a "CONTROLS".

4.2 Neteja de les dades

En aquesta secció expliquem les transformacions aplicades sobre la nostra base de dades de manera que no perdem molta informació rellevant i ens faciliti l'anàlisi. En primer lloc, ens quedarem amb les variables i els casos de dones que tenen més dades registrades, després farem una anàlisi de dependències i eliminarem les variables redundants, finalment descartarem els valors atípics (outliers).

4.2.1 Reducció del número de variables per número de dades disponibles

Per cadascuna de les 1090 variables comptem el número de valors nuls (NA) i calculem el percentatge de dades no disponibles. En una taula guardem el nom de la variable, el número de NA detectats i el percentatge. Després ordenem de menys a més (NA)s i a cada fila li posem un indicador de la posició.

El primer cop que compilem detectem que hi ha 18 variables que segons el nostre codi no tenen cap valor nul. Ho comprovem a la base de dades i ràpidament detectem que no és correcte, excepte per CODI_ID i source_01 que sí és veritat que no tenen NA. El problema ha sigut que són variables de tipus caràcter que tenen els nuls codificats com cadenes buides (" ") de mides diferents segons la variable. Per tant, abans de continuar amb la neteja de dades hem de solucionar aquest problema, les variables que hem de modificar són: OTRASCOMPL, PAQUETES, OTROSTX, ENFERMASOC, TIPO, PESO_M3, TIPOMALF, TIPOCOMPL, HIJO, ttos, MED_2A, MEDICACION, VAR00011 i VAR00018.

Primer busquem un valor buit de cadascuna de les variables i apuntem les cadenes, després construïm un bucle que recorri totes les files i per cada columna corresponent converteixi les cadenes sense valors en NA si troba les cadenes apuntades anteriorment. A continuació, comprovem si s'ha fet correctament, per cadascuna de les 16 variables comptem els NA, els valors disponibles i la suma entre els dos, aquest últim ha de sumar 3194 que és el número de pacients que hi ha.

Ja hem arreglat el problema, per tant, podem continuar amb la reducció de variables. Tornem a executar el primer codi. A la base de dades hi ha molts espais buits per tant ens quedem amb el 10% de les variables més plenes, que correspon a 109 variables on el percentatge de dades disponibles va de 100 a 32,75%.

4.2.2 Reducció del número de casos per número de dades disponibles

Tornem a aplicar el procediment anterior però en aquest cas comptarem els NA que hi ha per cada pacient. Un cop ordenada la base de dades ens quedem amb les dones que tenen com a mínim un 70% de les variables. Aquestes són 1818 dones les quals corresponen al 47,17% de les originals.

Ara tenim una base de dades de 109 variables per 1818 casos. Per continuar amb l'estudi, separem les variables quantitatives de les qualitatives i estudiarem les relacions que hi ha entre les numèriques. Dins de les variables quantitatives en tenim unes que representen diferents dates però no es mostren amb el format correcte, per tant les convertim en format any-mes-dia i les separem en un altre conjunt. A la Taula 7 descrivim el significat de les variables quantitatives que analitzarem i les seves unitats de mesura.

Taula 7. Descripció de les variables numèriques. Font pròpia.

Variable	Significat	Unitats
CODIGO_ID	Codi de la pacient	-
edad	Edat	Anys
EMB_PREV	Número d'embarassos previs	Núm. Embarassos
TALLA	Altura en cm	cm
TALLA_M	Altura en m	m
TAS	Tensió arterial sistòlica	mmHg
TAD	Tensió arterial diastòlica	mmHg
PESO	Pes a la primera visita	kg
SEMANA_P	Setmanes gestacionals que han sigut completades	setmanes
PESO_FET	Pes de naixement	g
ABORTOS_PREV	Número d'avortaments previs	Núm. Avortaments
TTOGB	Glucosa basal TTOG 100g	mg/dl
TTOG1H	Glucosa 1h després de TTOG 100g	mg/dl
TTOG2H	Glucosa 2h després de TTOG 100g	mg/dl
TTOG3H	Glucosa 3h després de TTOG 100g	mg/dl
TALLA2	Altura al quadrat	m2
PESO_PR	Pes previ a l'embaràs	kg
AUMENTOP	Augment de pes	kg
BMIPV	Índex de massa corporal de la primera visita	kg/m2
SEMANA_TTOG	Setmana en la que s'ha fet la TTOG	Setmanes
PESO_F	Pes final	kg
BMIPREG	BMI previ a la gestació	kg/m2
GRUPOBMIPREG2	Tercera classificació segons el BMI	-
DMGPESO	Desconegut	-
GRUPOBMIPREG1	Segona classificació segons el BMI	-
GRUPOBMIPREG	Classificació segons el BMI inicial	-
INCPESOPV	Increment de pes primera visita	kg
INCBMIPV	Increment de l'índex de massa corporal	kg/m2

APGAR1	Puntuació de la primera prova APGAR (1 minut després del naixement)	-
APGAR2	Puntuació de la segona prova APGAR (5 minuts després del naixement)	-
IVE	Interrupció voluntària de la gestació	-
BMIFINAL	Índex de massa corporal final	kg/m2
INCPESOTOTAL	Increment de pes total	kg
INCBMITOTAL	Increment de IMC total	kg/m2
N_PUNTOS_NDDMG	nº pts alterats de la PTTOG segons el criteri de la NDDMG	Núm. punts
N_PUNTOS_CC	nº pts alterats de la PTTOG segons el criteri de CARPENTER Y COUSTAN	Núm. punts
SEMPV	Setmana de la primera visita	Setmanes
TALLA_F	Altura de naixement en cm	cm
GLUC_SULL	Glucèmia del Test de O'Sullivan	mg/dl
GLUCOSA_INICIAL	Glucèmia basal	mg/dl
SEMANA_EXTRACCIONsemext	Setmana d'extracció	Setmanes
VAR00006	Desconegut	-
PLACENTA	Pes de la placenta en g	g
PESO_M1	Pes del primer fill al néixer en g	g
tshges	Concentració de TSH (Tirotropina)en sèrum (mUI/L) durant l'embaràs	mUI/L
MENARQUIA	Edat de la menarquia	Anys
Hba1C_INICIOCON	Hemoglobina a l'inici del control	%
Source_01	Font de les dades	-
DMGTTOG	Número de PTOG	Núm. PTOG

En primer lloc, detectem que el codi de les pacients no ens aporta informació per tant l'eliminem. En segon lloc podem observar que hi ha variables redundants que es poden calcular a partir d'altres, per tant, les podríem descartar per reduir la dimensió de les dades. Per saber quines variables eliminar farem un anàlisi de dependències lineals. A la Taula 7, tenim variables que representen conceptes qualitatius, com és el cas de la classificació segons el BMI, que estan codificats com a variables numèriques per tant les utilitzarem en el nostre anàlisi.

4.2.3 Anàlisi de dependències lineals

Calcularem tant la matriu de correlacions per parells com els quadrats dels coeficients de correlació múltiple. Aquests últims ens indicaran les variables que poden ser explicades en relació a totes les altres, mentre que els primers ens mostraran quines són les que causen la relació. Quan construïm la matriu de correlacions apareix el símbol "?" a les columnes de source_01 i DMGTTOG, aquests ens provoquen problemes a l'hora d'executar els següents càlculs i per tant, decidim eliminar les dues variables.

Per calcular els coeficients de correlació múltiple al quadrat, primer obtenim la matriu de covariàncies i la seva inversa, després, agafem els valors de les seves diagonals i els multipliquem. El següent pas consisteix en calcular la inversa dels elements i restarem cadascun dels resultats a 1. Ajuntem en un dataframe el nom de les variables amb el seu coeficient i ordenem de més a menys, tal com es mostra a la Taula 8.

Taula 8. Quadrat dels coeficients de correlació múltiple de cada variable. Font pròpia.

Variable	Coeficients $\hat{\beta}^2$
TALLA2	1,05771487
INCBMIPV	1,01549781
BMIPREG	1,0130271
BMIPV	1,00620959
TALLA	1,00211045
TALLA_M	1,00208355
INCPESOTOTAL	0,9909454
INCBMITOTAL	0,9859738
PESO_PR	0,98596303
PESO_F	0,97317999
PESO	0,97181869
BMIFINAL	0,96184856
GRUPOBMIPREG	0,95883601
N_PUNTOS_NDDMG	0,9513227
GRUPOBMIPREG1	0,95037447
DMGPESO	0,94911856
N_PUNTOS_CC	0,92220825
SEMPV	0,90825541
GRUPOBMIPREG2	0,88968579
INCPESOPV	0,88305022
SEMANA_TTOG	0,85759586
TTOG1H	0,73866721
TTOG2H	0,71610297
TTOG3H	0,66243595
PESO_FET	0,65522979
SEMANA_EXTRACCIONsemext	0,64073311
TTOGB	0,59634
AUMENTOP	0,57922384
Hba1C_INICIOCON	0,5679746
GLUCOSA_INICIAL	0,5527699
GLUC_SULL	0,54863293
TALLA_F	0,49030543
EMB_PREV	0,4458748
ABORTOS_PREV	0,4426865
SEMANA_P	0,4089685
APGAR1	0,36757173
PESO_M1	0,36594106
APGAR2	0,35837764
PLACENTA	0,35175197
VAR00006	0,32150025
TAD	0,1951652
edad	0,14114267
IVE	0,12931441
tshges	0,09381196
MENARQUIA	0,0905856
TAS	0,06583455

Amb la taula anterior es demostra que hi ha molta dependència entre les variables, principalment les associades al pes i l'altura, ja que la distribució de moltes d'elles poden ser explicades pràcticament a partir d'altres.

A continuació, representem la matriu de correlacions en un gràfic de colors que ens ajudi a veure de manera ràpida les tendències. A primera vista podem detectar ràpidament que hi ha una alta dependència entre les diferents variables, no hi ha vermells molt forts, és a dir no hi ha correlacions negatives molt intenses, en canvi, sí que hi ha tendències de dependències positives més altes. Volem desfer-nos de les variables redundants, per això ens fixarem en les que tenen coeficients de valor absoluts alts, ja que com més pròxim a 1 siguin, més linealment dependents són. Per poder estudiar aquestes relacions separem en grups, tal com veiem a la Figura 9.

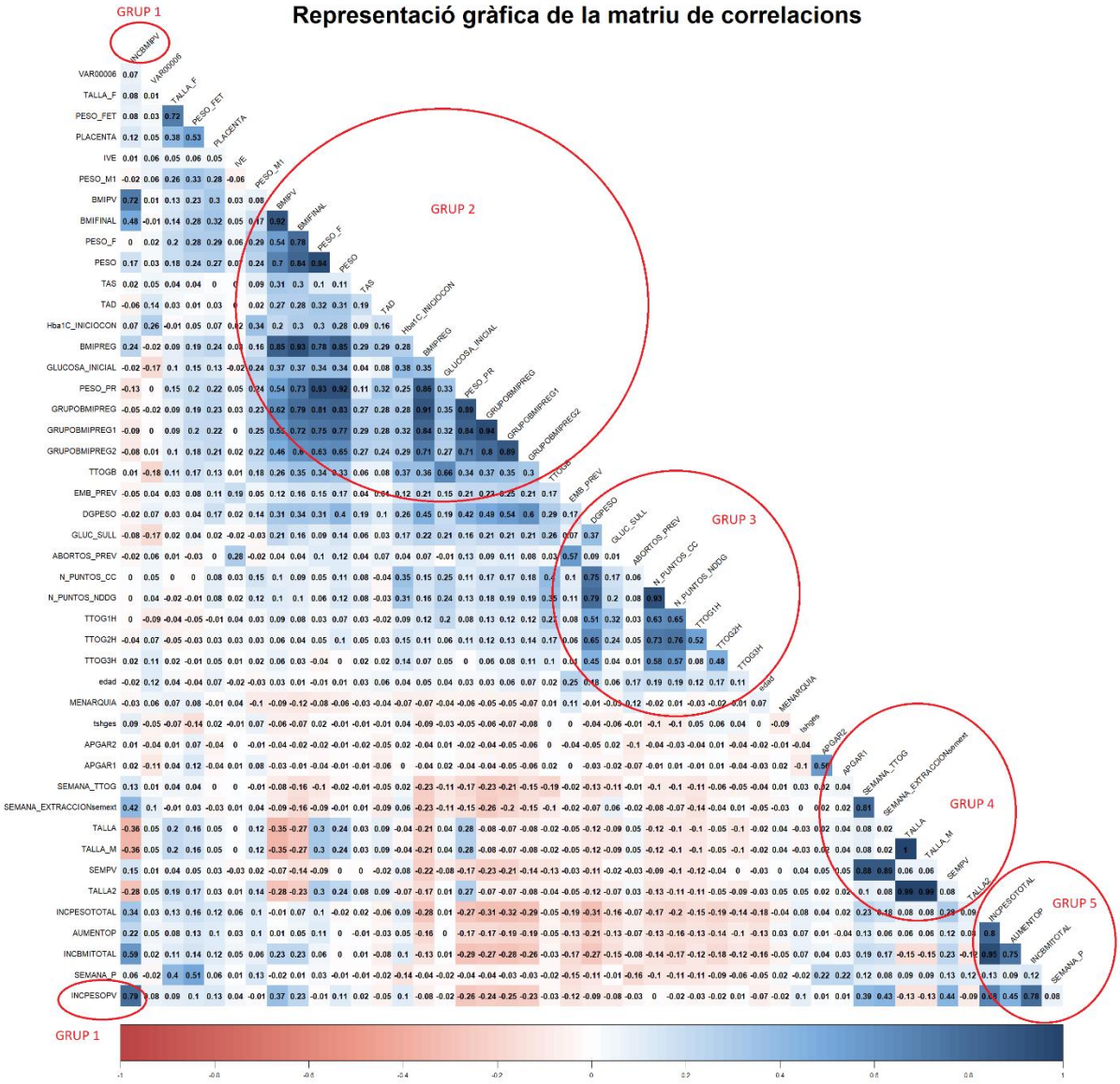

Figura 9. Representació gràfica de la matriu de correlacions. Font pròpia.

- GRUP 1: INCPESOPV té un coeficient de 0,79 amb INCBMIPV. El BMI es calcula dividint el pes entre l'altura elevada al quadrat en metres. Les dues són valors de la primera visita, per tant estan associades amb el mateix increment de pes. En definitiva si tenim aquest increment podrem calcular el del BMI, per tant, podem prescindir de INCBMIPV.

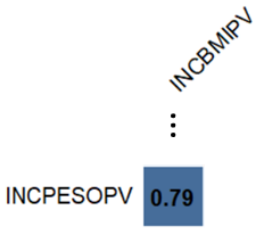

Figura 10. Representació gràfica del grup 1 de la matriu de correlacions. Font pròpia.

- GRUP 2: Hi ha moltes variables que representen el mateix però en moments diferents de l'embaràs. Per un costat, tenim el pes previ, el de la primera visita i el final, a banda amb aquests valors s'han obtingut els diferents BMI i depenent d'aquests s'han fet les classificacions en grups. Com podem observar al gràfic entre aquestes variables les correlacions són bastant fortes i podem eliminar-ne algunes. Entre les tres variables de pes (PESO_PR, PESO i PESO_F) hi ha dependències molt fortes, si guardem els increments podem eliminar el pes final i el pregestacional per tant, només ens quedem amb el pes de la primera visita. Igual que en el grup anterior els índex de massa corporal els eliminem ja que els podem obtenir amb variables existents. Finalment, tenim les variables on es guarden les classificacions segons el BMI, que com és d'esperar, depenen dels pesos i dels índex de massa corporal. Per un altre costat, també tenen coeficients alts entre ells, per tant podem considerar que la classificació no varia molt en els diferents grups. Entre aquests decidim quedar-nos amb GRUPOBMIPREG1 ja que és el que té una correlació més forta entre els altres 2.

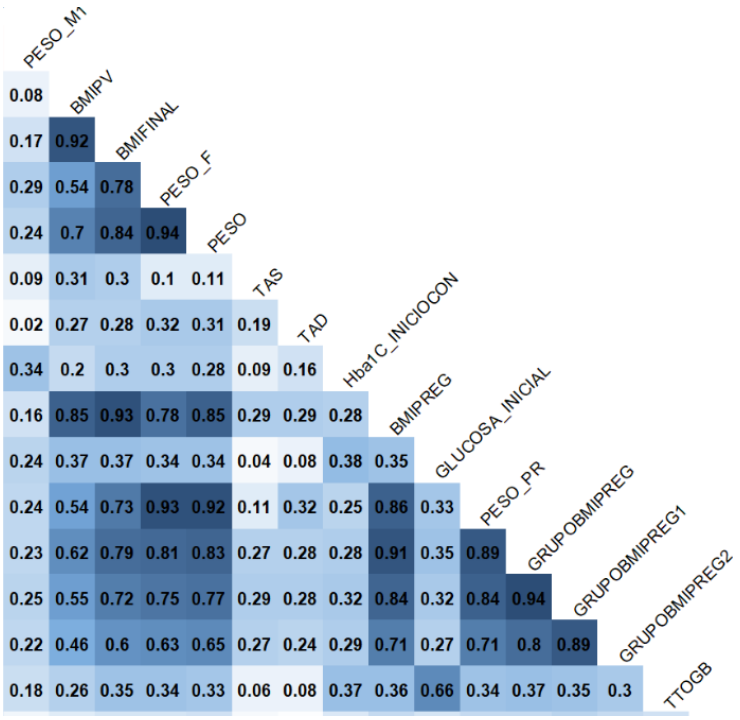

Figura 11. Representació gràfica del grup 2 de la matriu de correlacions. Font pròpia.

- GRUP 3: Aquí tenim les variables relacionades amb la prova de tolerància oral a la glucosa que serveix per diagnosticar la malaltia. Tenim les glucèmies mesurades 1h, 2h i 3h després de la ingesta i el número de punts alterats dels criteris de Carpenter/Coustan i de la National Diabetes Data Group, aquests valors són els que s'utilitzen en la proposta de detecció de la DMG de la American Diabetes Association,

si en algun d'aquests criteris s'alteren com a mínim dos valors s'accepta el diagnòstic, tal com s'indica a la Taula 4. No hi ha molta diferència entre els llindars de C/C i la NDDMG per això tenen tanta relació. En aquest estudi ens quedarem amb els punts alterats del criteri de Carpenter/Coustan ja que com té límits més baixos si es diagnostica seguint l'altre criteri també es complirà en aquest.

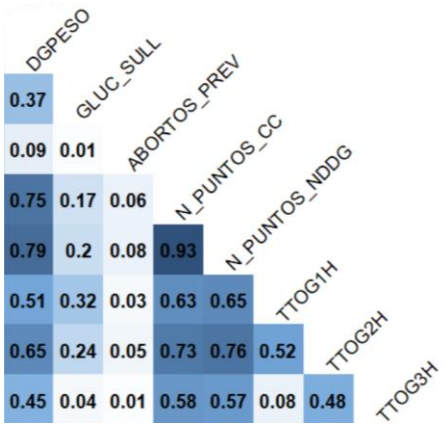

Figura 12. Representació gràfica del grup 3 de la matriu de correlacions. Font pròpia.

- GRUP 4: Detectem que el coeficient entre TALLA i TALLA_M és 1, això es deu a que representen la mateixa magnitud però expressada en unitats de mesura diferents, la primera és l'altura en cm i la segona en m, per tant, és una combinació lineal i podem eliminar-ne una. Per un altre costat, TALLA_2 és l'altura elevada al quadrat, es a dir, també esta 100% relacionada amb les altres dues. D'aquestes tres variables ens quedem amb TALLA_M ja que la necessitem en metres per calcular el BMI.

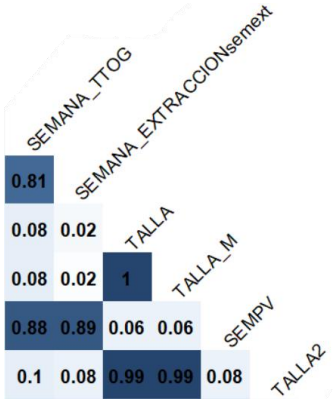

Figura 13. Representació gràfica del grup 4 de la matriu de correlacions. Font pròpia.

- GRUP 5: Com hem dit abans l'increment de BMI es pot obtenir a partir de l'increment de pes, per tant podem eliminar-la. Per un altre costat, entre INCPESOTOTAL i AUMENTOP hi ha un coeficient de correlació de 0.95 i ens quedem només amb la primera.

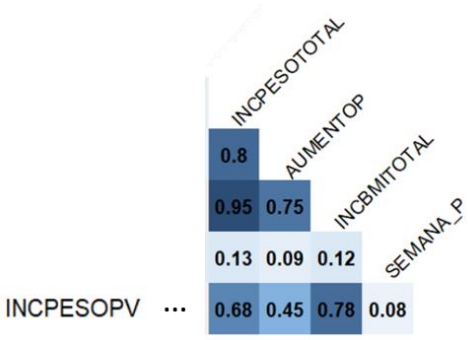

Figura 14. Representació gràfica del grup 5 de la matriu de correlacions. Font pròpia.

Finalment les variables descartades són: AUMENTOP, INCBMITOTAL, TALLA, TALLA2, N_PUNTOS_NDDMG, GRUPOBMIPREG, GRUPOBMIPREG2, PESO_F, PESO_PR, BMIPV, BMIFINAL, BMIPREG, INCBMIPV. Si tornem a dibuixar la matriu de correlacions podem comprovar que han disminuït les dependències, vegeu la Figura 15.

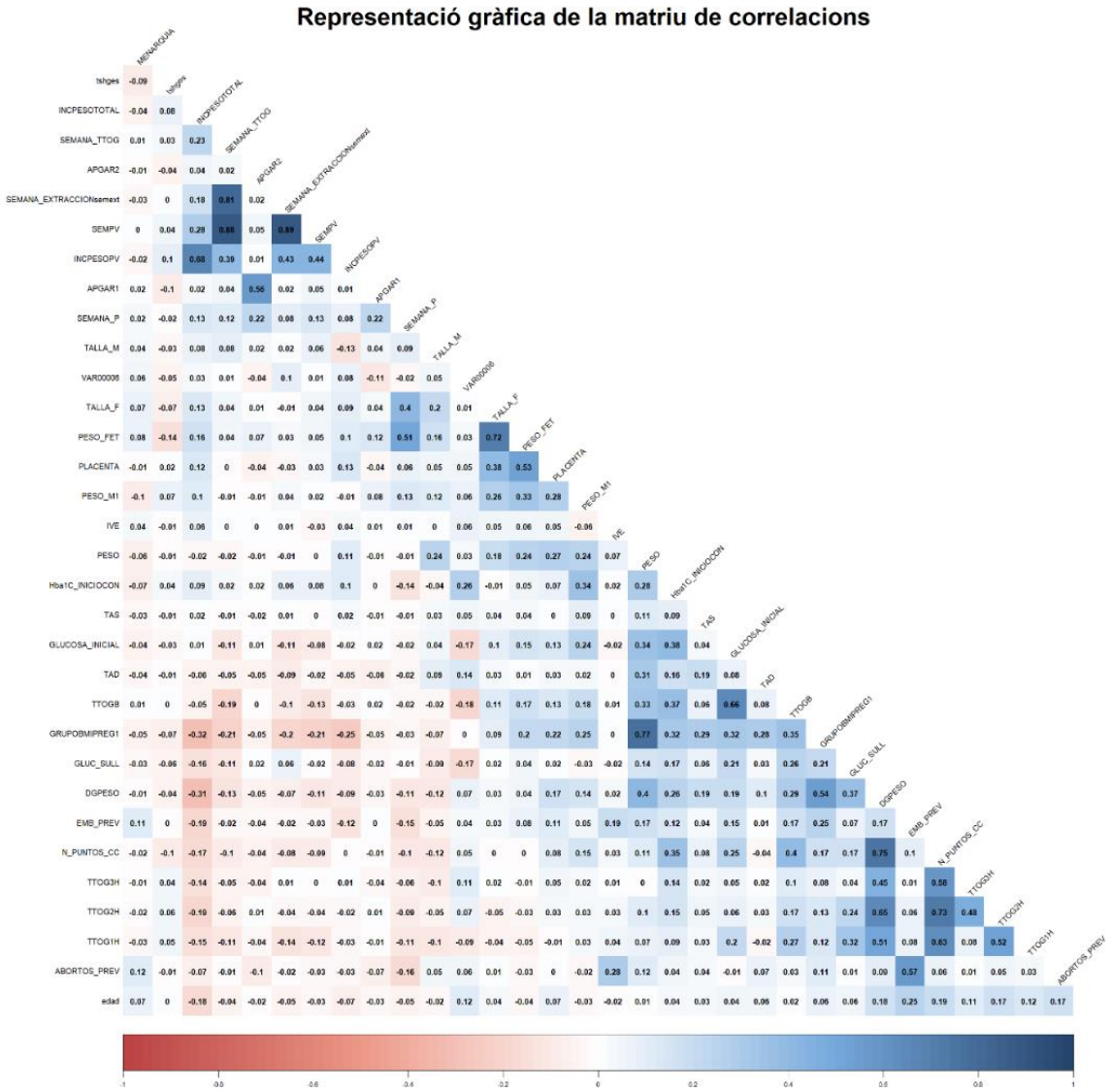

Figura 15. Representació gràfica de la matriu de correlacions un cop eliminades les variables redundants. Font pròpia.

4.2.4 Eliminació d'outliers

Un cop hem eliminat la redundància ens queda treure els individus amb valors atípics en qualsevol de les diferents variables i utilitzarem les distàncies de Mahalanobis per detectar-los. A partir de les dades, el vector de mitjanes i la matriu de covariàncies calculem les distàncies. A la Figura 16 observem que la distància de Mahalanobis de la majoria de dones es troba entre 0 i 100. Com més baixes siguin les distàncies menys outliers hi ha, per tant, fem un filtre per quedar-nos amb aquestes pacients tant en el conjunt de dades numèriques com les qualitatives que tenim guardades.

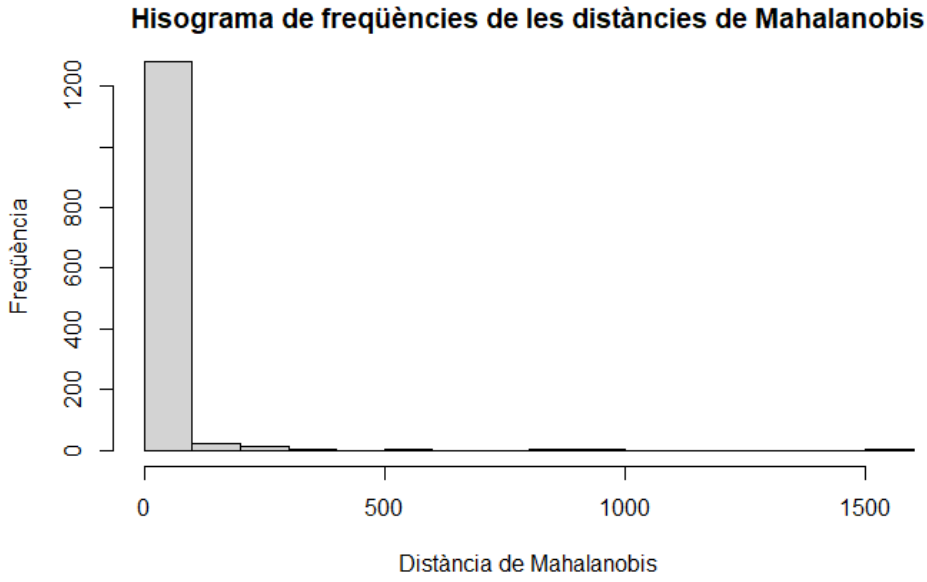

Figura 16. Histograma de freqüències de la distància de Mahalanobis. Font pròpia.

Un cop eliminats els valors atípics ens queda un dataset de 109 variables observades en 1283 pacients.

4.3 Construcció dels models de predicció

4.3.1 Dades d'aprenentatge i d'avaluació

Per comprovar l'eficàcia dels models separarem les dades en dos grups, un a partir del qual construirem el model de predicció i el segon per testejar-lo amb unes altres dades. Per les dades d'entrenament seleccionem el 80% i per les de test el 20% restant. Volem detectar si les pacients tenen la diabetis gestacional per tant afegim la variable grupo_dg del conjunt de dades qualitatives.

A l'entorn de R, carreguem les llibreries necessàries: *MASS, FactorMinerR, missMDA, nnet* i *factoextra.* Habitualment si treballem amb bases de dades que han estat omplertes per humans trobarem dades buides, com és el nostre cas. Això porta problemes a l'hora de treballar amb sistemes de càlcul informàtics, *missMDA* és una llibreria de R que permet imputar i ajustar els valors incomplerts d'un conjunt de dades a través d'un anàlisi de components principals, un anàlisi de correspondència múltiple o un anàlisi factorial múltiple. En el nostre cas utilitzarem un PCA a través de la funció *imputePCA( )* [36], indicant que s'ajusti sobre 4 components principals i s'escali les dades perquè totes tinguin mitjana 0 i desviació estàndard 1.

4.3.2 Reducció de la dimensió

Després de netejar les dades i imputar els valors nuls, reduim la dimensionalitat a través d'un PCA per buscar les direccions principals sobre el dataset d'entrenament. A la Figura 17 es mostra la variància explicada per cadascuna de les CP obtingudes, les 4 primeres representen tota la variabilitat, ja que abans hem utilitzat un algoritme d'imputació a partir d'un altre anàlisi de components principals de 4 dimensions. Finalment actualitzem les puntuacions sobre les components principals amb els dos conjunts.

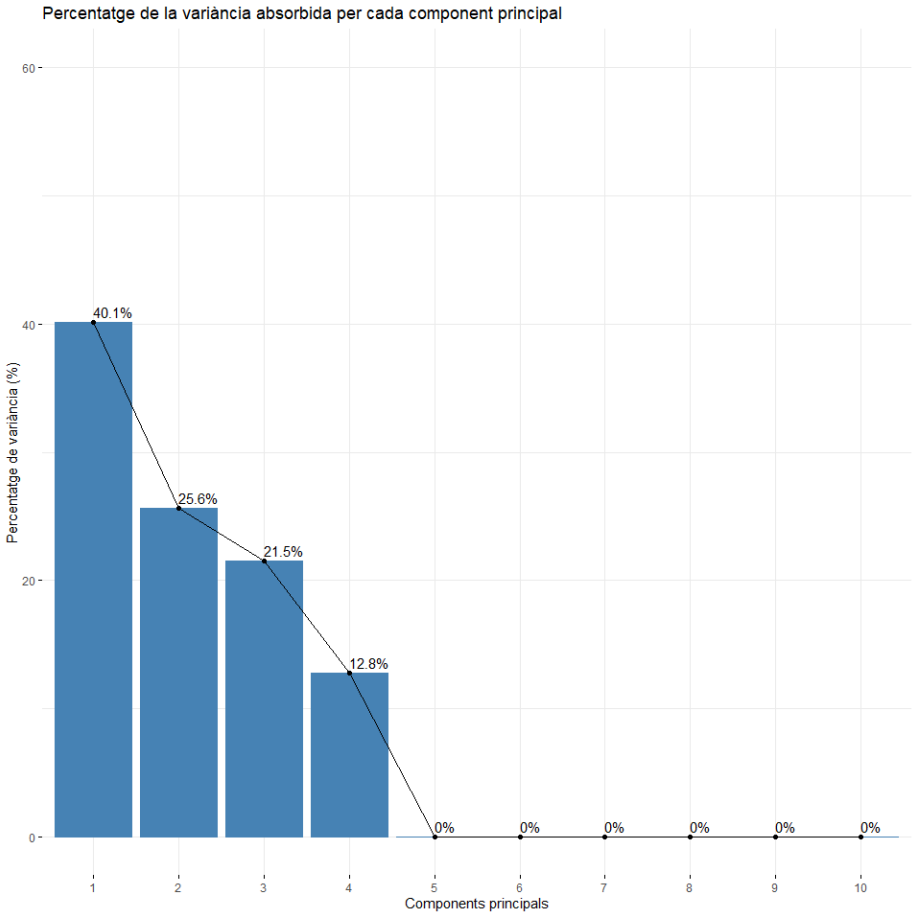

Figura 17. Representació gràfica de la variància absorbida per cada component principal. Font pròpia.

A les Figures 18 i 19 mostrem els plans formats per les direccions PC1-PC2 i PC1-PC3 seleccionant la primera component a l'eix de les x. Els vectors curts són els atributs amb menys representació, en conseqüència, els més llargs són les variables que més contribueixen. Per un altre costat, els vectors que s'agrupen, es a dir, tenen aproximadament els mateixos angles, estan correlacionats positivament, mentre que si es troben en sentits oposats significa que ho estan de manera negativa.

Els grups de dependències que es mostren coincideixen amb la matriu de correlació mostrada a la Figura 15. Si busquem diferencies entre els dos plans podem determinar que la segona component pren casi sempre valors negatius, també ho fa la primera, mentre que en la tercera trobem tant vectors positius com negatius.

Les variables que estem utilitzant les podem separar en:

- Variables relacionades amb les setmanes en que s'ha fet la primera visita i les proves.
- Resultats de les proves de cribratge: glucosa inicial, test de O'Sullivan, proves de tolerància oral a la glucosa a la primera, segona i tercera hora, el percentatge d'hemoglobina i el punts alterats del criteri de Carpenter-Coustan.
- Variables relacionades amb l'IMC: increments de pes, altura, el grup de la classificació de l'IMC.

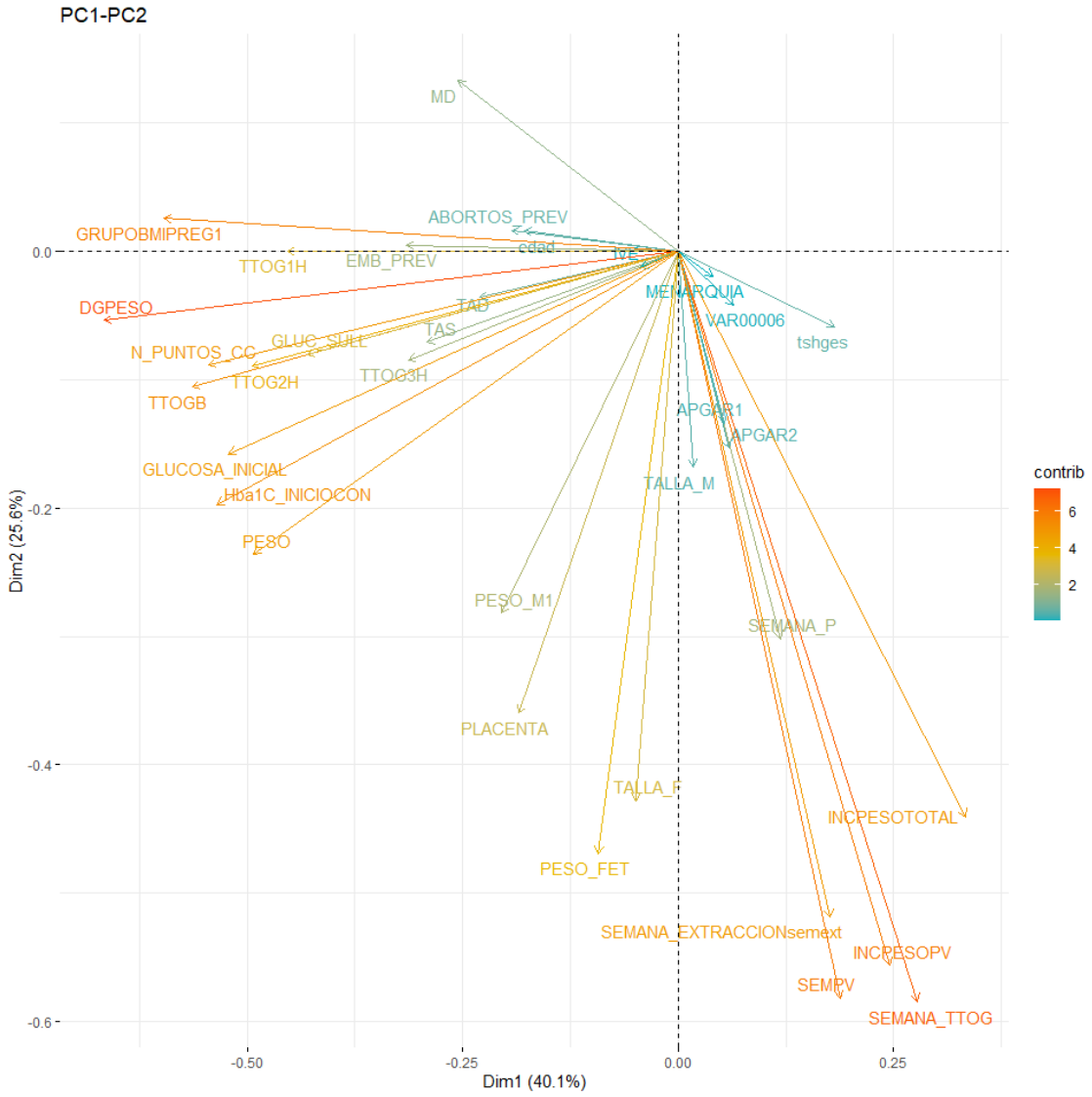

Figura 18. Representació gràfica de la contribució de les variables sobre la primera i la segona component principal. Font pròpia.

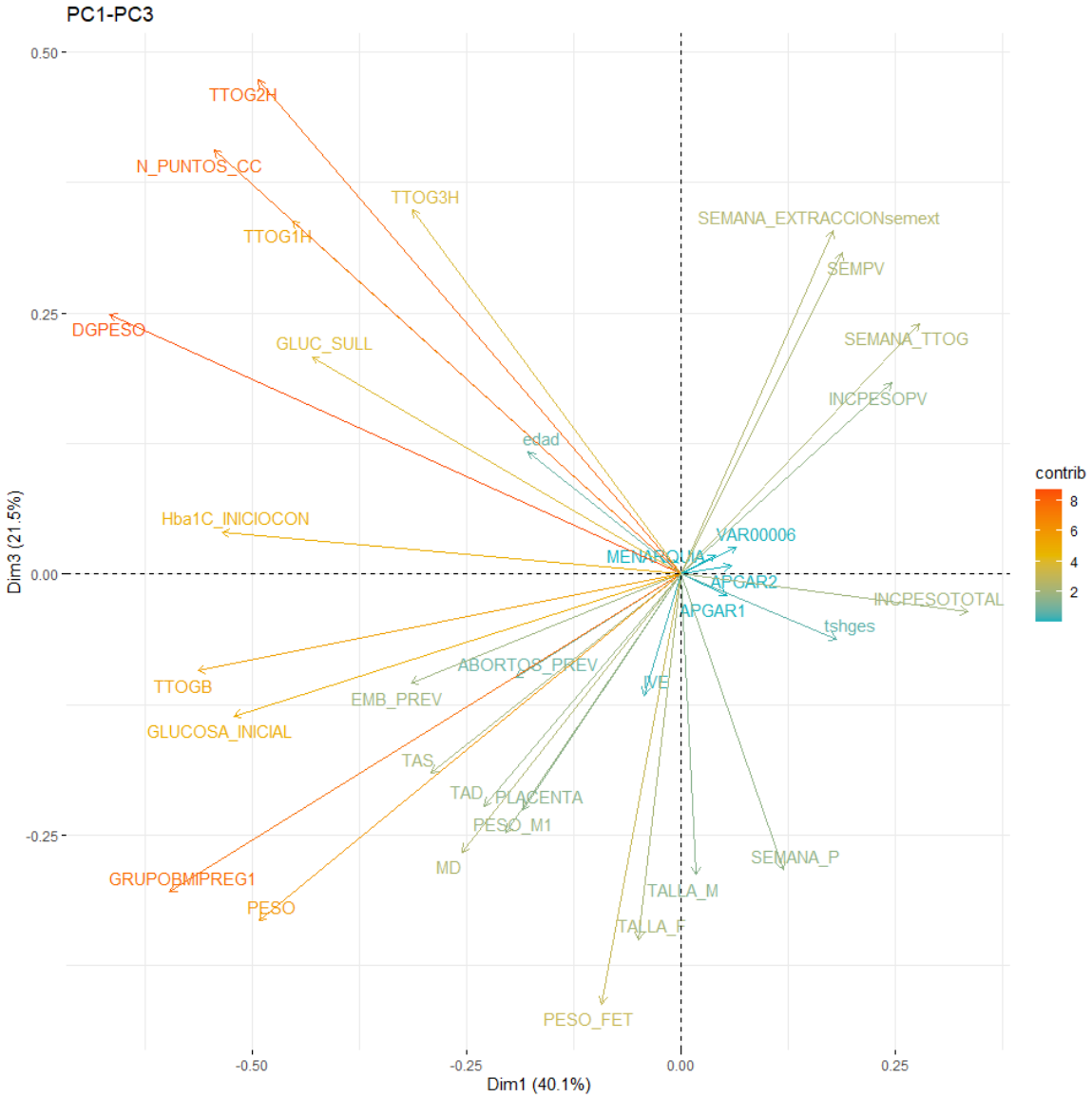

Figura 19. Representació gràfica de la contribució de les variables sobre la primera i la tercera component principal. Font pròpia.

Per avaluar si hi ha diferències entre les dones amb diabetis gestacional i les control sobre les noves puntuacions, les dibuixem sobre els plans PC1-PC2 i PC2-PC3 i pintem els punts segons a qui pertanyen a les Figures 20 i 21. S'observa clarament la diferència als dos gràfics però especialment sobre el pla PC1-PC3 ja que podríem dibuixar una línia diagonal que separa casi perfectament els dos grups. En definitiva, existeixen diferències en les característiques de les pacients, per tant les variables seleccionades ens serviran per construir el model de classificació.

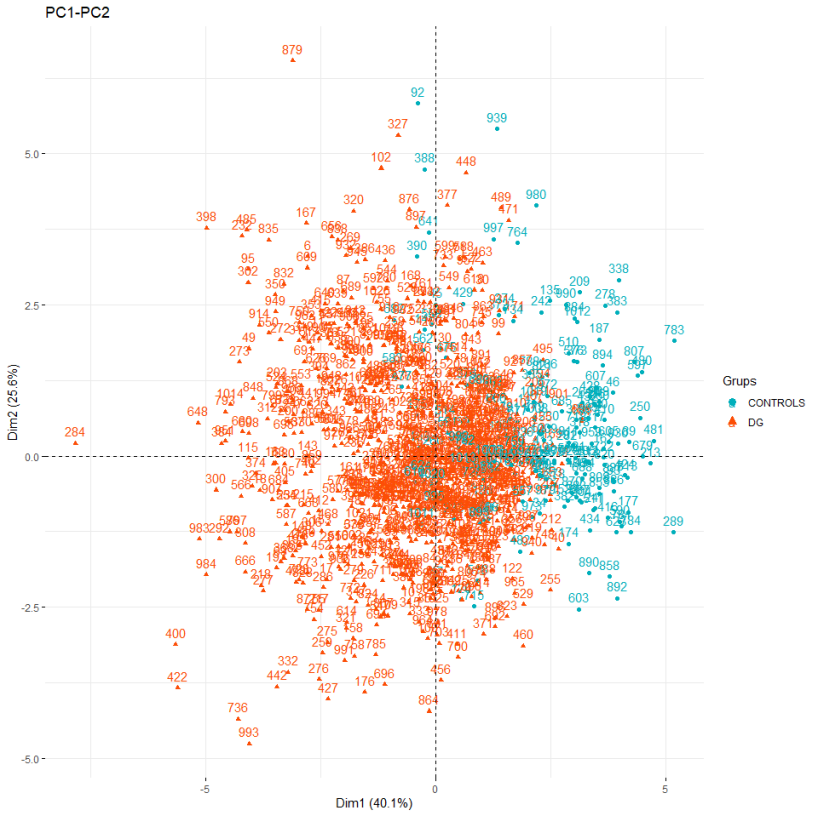

Figura 20. Representació gràfica de les puntuacions dels individus sobre les components principals 1 i 2 indicant el grup d'observació al que pertanyen. Font pròpia.

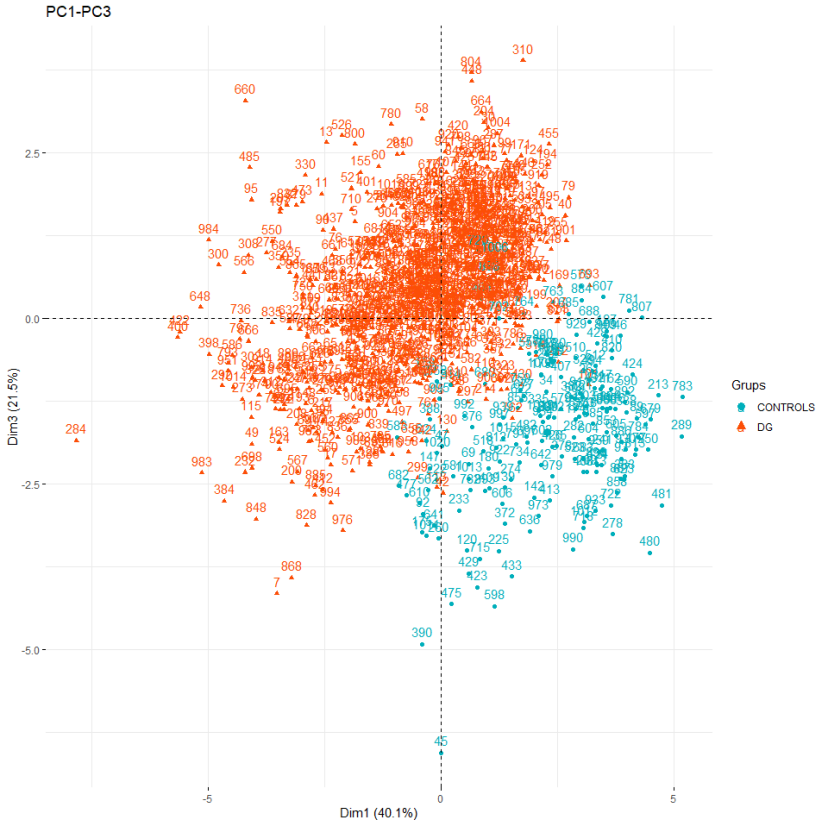

Figura 21. Representació gràfica de les puntuacions dels individus sobre les components principals 1 i 3 indicant el grup d'observació al que pertanyen. Font pròpia.

4.3.3 Construcció del model de Regressió Logística

Finalment, ara que ja hem reduït la dimensió, utilitzem la funció glm() per construir el model. Indiquem el grup d'observació com la variable factor i els 4 components com les variables explicatives i que la família és binomial.

A la Taula 9 s'indiquen les característiques de l'equació. Per un costat els coeficients de cada component, també l'error estàndard que representa una mesura de la variabilitat associada per l'estimació del coeficient, el z-value que es la divisió dels dos conceptes anteriors i finalment hi ha el p-value. A més, a l'Equació (35) s'escriu la fórmula del classificador considerant p com la probabilitat de tenir diabetis gestacional.

Taula 9. Coeficients del model de regressió logística. Font pròpia.

	Estimacions dels coeficients	Std. Error	z-value	p-value
grupo_dg	4,9714	0,4513	11,017	< 2e-16
PC1	-2,0253	0.2024	-10,006	< 2e-16
PC2	-0,5334	0.1523	-3,502	0,000462
PC3	2,5070	0.2533	9,897	< 2e-16
PC4	0,2803	0.1960	1,430	0,152632

$$Logit(p) = log\frac{p}{1-p} = 4{,}9714 - 2{,}0253PC1 - 0{,}5334\,\mathrm{PC2} + 2{,}507\mathrm{PC3} + 0{,}2803\mathrm{PC4} \quad (35)$$

El p-value de PC4 és superior a 0,05 per tant no és una variable predictora estadísticament significativa, en unes altres paraules, no podem descartar la hipòtesi nul·la i en conseqüència no podem saber si els seus canvis tenen efecte sobre el resultat del model. Per un altre costat, a partir de les estimacions dels coeficients podem determinar que:

- Per cada unitat que augmenta PC1, s'espera que el logaritme de l'odds, és a dir el logaritme de la divisió de la probabilitat que la dona tingui DMG entre la probabilitat de que no la tingui, disminueixi 2,0253 de mitjana.
- Per cada unitat que augmenta PC2, s'espera que el logaritme de l'odds de la variable del grup d'observació disminueixi 0,5334 de mitjana.
- Per cada unitat que augmenta PC3, s'espera que el logaritme de l'odds de la variable del grup d'observació augmenta 2,5070 de mitjana. Podem considerar que aproximadament PC1 i PC3 tenen un efecte simètric sobre el resultat.

4.3.4 Avaluació del model

A continuació, avaluem l'eficàcia del model per detectar la diabetis gestacional, primer utilitzem el conjunt d'entrenament: són les dades utilitzades per crear el model per tant és possible que la taxa d'encerts sigui més alta. Després classifiquem les dones del conjunt de test.

4.3.4.1 Predicció del conjunt d'entrenament

A la Figura 22 representem la corba Logit de la predicció, sobre l'eix de les x és mostra l'índex generat per cada cas i a l'eix y la probabilitat de tenir la malaltia. Els dos grups d'observació s'han separat correctament, ja que la majoria d'individus amb $p \geq 0{,}5$ són les pacients del grup DMG, pintades de color blau. Mentre que, per sota de 0,5 la majoria són les dones sense la malaltia que estan pintades de color vermell.

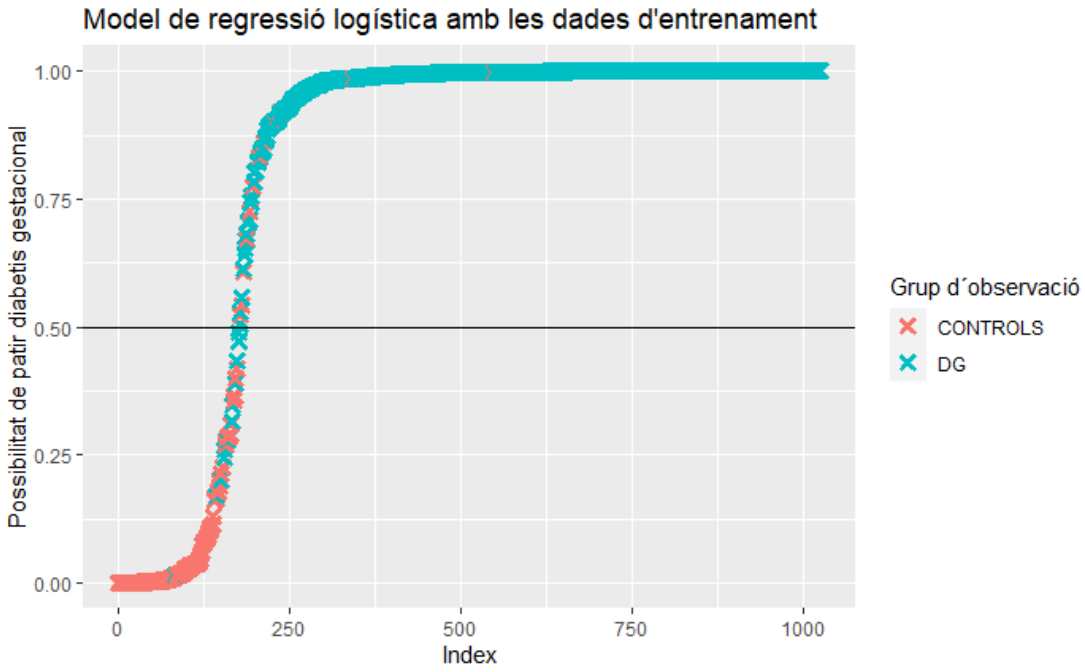

Figura 22. Model de regressió logística sobre les dades d'entrenament, la línia negra representa el límit que utilitza el model per classificar la pacient en el grup DMG si la probabilitat està per sobre de 0,5. Font pròpia.

Per un altre costat, també construïm la matriu de confusió, indicada a la Taula 10, a partir de la qual dibuixem la Corba ROC, per avaluar la predicció d'aquest conjunt. S'ha detectat 833 True Positives (TP), 165 True Negatives (TN), 12 False Negatives (FN) i 16 False Positives (FP), amb aquests valors la taxa de precisió del model és del 97,27%.

Taula 10. Matriu de confusió de les prediccions del conjunt d'entrenament. Font pròpia.

		Observacions reals	
		CONTROLS	DMG
Predicció	CONTROLS	165	12
	DMG	16	833

A més, l'àrea sota la corba ROC, vegeu la Figura 23, és del 0,989 per tant, de moment podem considerar que el model creat és molt bo però encara ens queda comprovar la seva eficàcia sobre el conjunt de test.

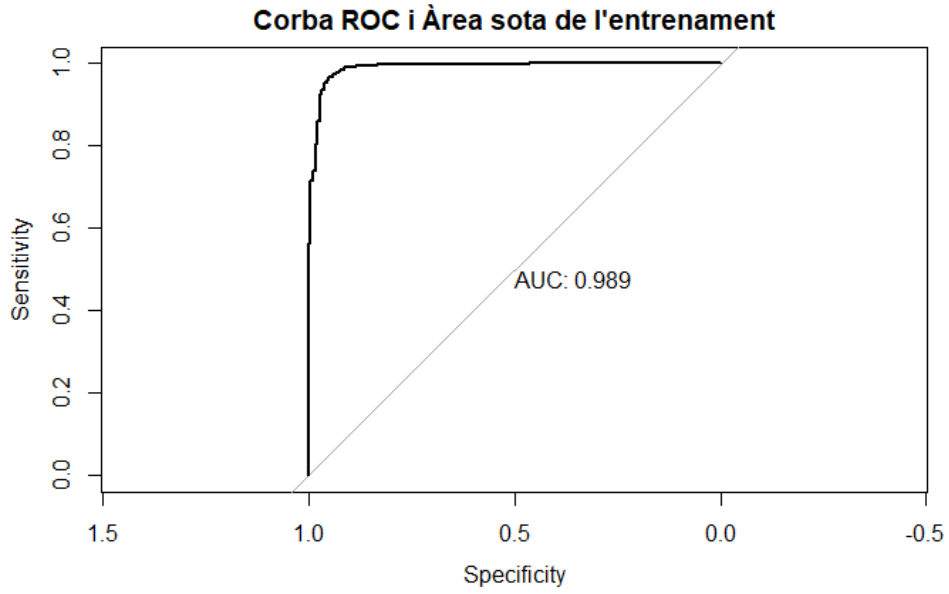

Figura 23. Corba ROC i AUC construïda a partir de la predicció del conjunt d'entrenament. Font pròpia.

4.3.4.2 Predicció del conjunt de test

Calculem la predicció i tornem a dibuixar el model Logit, vegeu la Figura 24. A primera vista, s'observa que la classificació dels dos grups ha sigut molt bona, totes les dones amb DMG estan per sobre de 0,5, això vol dir que no hi ha cap fals negatiu. Per l'altra banda, sí que hi ha alguna dona control en la que s'ha diagnosticat erròniament la malaltia.

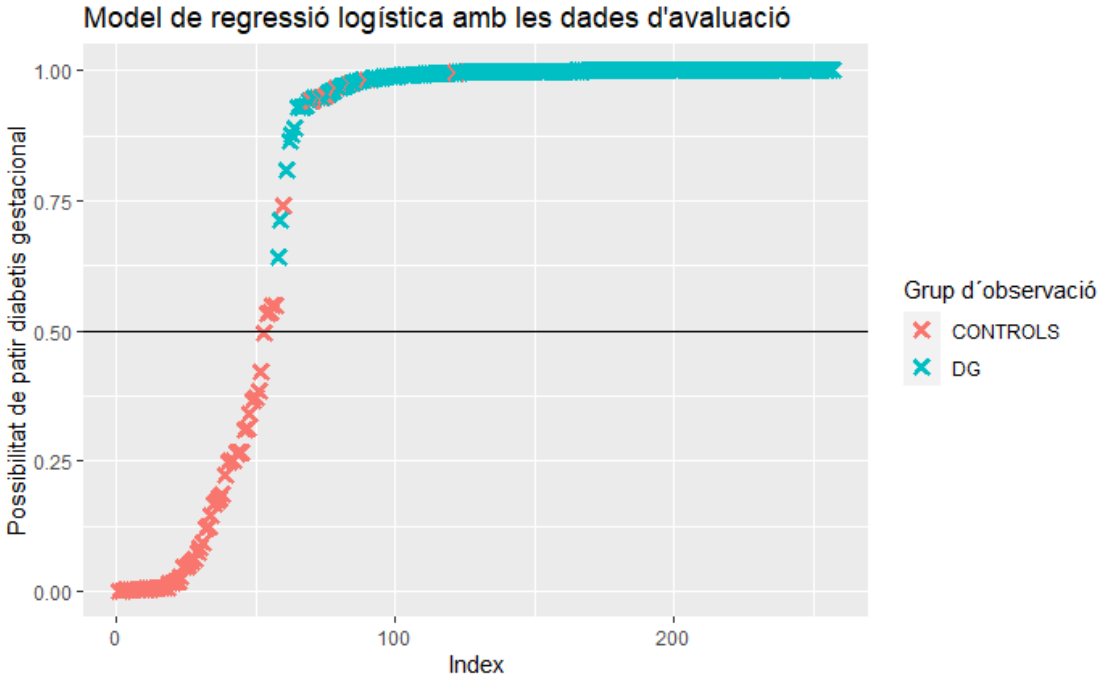

Figura 24. Model de regressió logística sobre les dades d'avaluació, la línia negra representa el límit que utilitza el model per classificar la pacient en el grup DMG si la probabilitat està per sobre de 0,5. Font pròpia.

A continuació, representem la matriu de confusió i la corba ROC. A la Taula 11 s'indica que el classificador ha fet 192 TP, 53 TN, 0 FN i 12 FP per tant la taxa de precisió és del 95.33%. Per un altre costat, l'àrea sota la corba ROC és 0,987, és a dir, amb el conjunt de dades test hem aconseguit la mateixa eficàcia que amb la predicció dels individus amb els que havíem construït el model.

S'han detectat com a pacients amb DMG a 12 dones del grup control, considerant que en la majoria de casos el tractament es basa en un canvi de l'estil de vida, incorporant exercici físic i dietes equilibrades, no és un problema molt gran aquest error. Per l'altra banda, és importat que el model sí detecti correctament la malaltia, i hem comprovat que ho fa. Tot i això el model només és un sistema de suport de decisió clínica, per tant, sempre hi ha d'haver un metge que supervisi el resultat i accepti o no la recomanació de la màquina, segons el seu criteri i les característiques clíniques de la pacient.

Taula 11. Matriu de confusió de les prediccions del conjunt d'avaluació. Font pròpia.

		Observacions reals	
		CONTROLS	DMG
Predicció	CONTROLS	53	0
	DMG	12	192

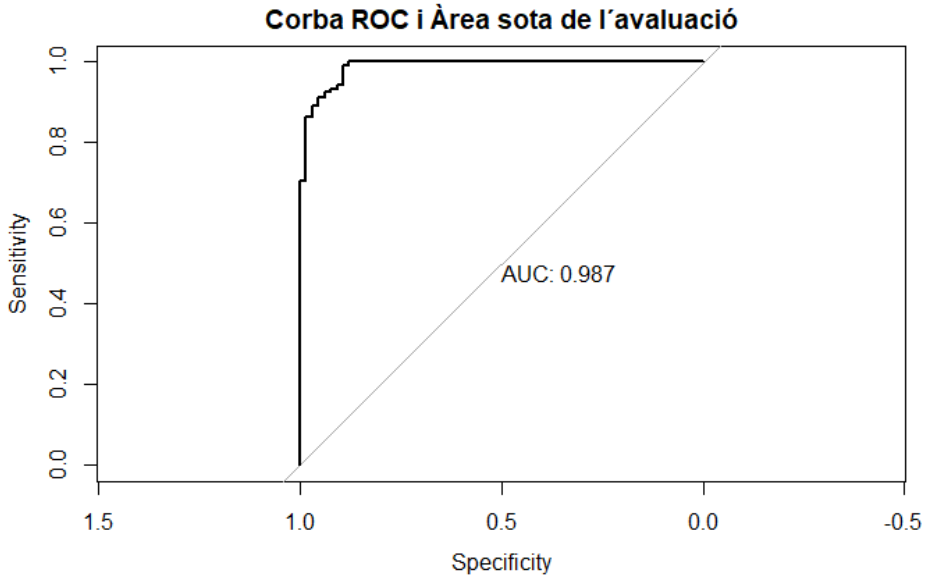

Figura 25. Corba ROC i AUC construïda a partir de la predicció del conjunt d'avaluació. Font pròpia.

En definitiva, a partir de les característiques individuals de cada dona hem pogut construir un model de regressió logística que permet diagnosticar la diabetis gestacional amb una fiabilitat molt alta. Per un altre costat, només hem necessitat un conjunt de 34 variables que es poden mesurar fàcilment, ja que la majoria estan relacionades amb els increments de pes, els resultats de les proves de cribratge estàndard i les setmanes en que es van realitzar les analítiques.

5 Conclusions

Durant el desenvolupament del treball hem aprofundit en els conceptes teòrics de la diabetis gestacional. Hem vist el seu origen, els seus efectes sobre la salut de la mare i el bebè, els factors de risc que la condicionen i les propostes pel diagnòstic de diverses organitzacions. A més, també hem presentat els conceptes generals de l'AM i les dues tècniques que hem utilitzat posteriorment per construir el model, que són l'Anàlisi de Components Principals per reduir la dimensió i el mètode de regressió logística per classificar una variable qualitativa a partir d'un conjunt de variables quantitatives.

En l'anàlisi estadística primer hem preparat el dataset inicial format per 1090 variables mesurades en 3854 dones. Hem treballat només amb els casos control i els DMG comuns, amb el 10% de variables amb més dades disponibles i les dones que com a mínim tenien el 70% omplert. Seguidament, hem eliminat les variables redundants després d'una anàlisi de dependències lineals i també, hem tret els outliers a partir de la distància de Mahalanobis de cada pacient. Finalment, hem imputat els valors nuls, i els hem ajustat a 4 components principals i hem separat les dades en el conjunt d'entrenament i el d'avaluació.

Al següent pas, hem utilitzat un ACP sobre el les dades d'entrenament per trobar les 4 direccions principals i hem actualitzat les puntuacions de les dues poblacions. D'aquesta manera hem transformat les observacions de 34 variables sobre 4 components.

L'últim pas ha consistit en generar el model de regressió logística i avaluar la seva eficàcia en la classificació en les dades d'entrenament i de control, de manera que hem obtingut una taxa de precisió final del 95,33%. En definitiva, podem considerar que hem aconseguit l'objectiu del treball, ja que hem pogut classificar amb un bon percentatge de precisió la població de dones. Tot i això, el següent pas seria avaluar-lo amb un altre conjunt de dades més gran en el que es pugui veure millor l'eficàcia real del model.

Per un altre costat, podríem seguir amb l'anàlisi introduint noves tècniques que ens permetessin estudiar l'efecte de les variables qualitatives, que havíem descartat, sobre el desenvolupament de la malaltia. Entre aquestes trobem l'ètnia, si la mare és fumadora o no, si hi ha familiars de primer grau amb malalties associades per tant segurament també seran de gran utilitat.

Podem associar el bon resultat del model a diferents causes: en primer lloc, la obesitat és un dels principals causants de la DMG i això es confirma en el nostre anàlisi, és a dir, moltes de les variables utilitzades com l'altura, el pes inicial i els increments i el grup de classificació segons l'IMC, són variables que serveixen per mesurar el nivell d'obesitat de la pacient. Per un altre costat, la macrosomia és un efecte secundari de la diabetis gestacional sobre el bebè i també, està representada en la nostra equació a través del pes i la talla al néixer i el pes de la placenta. En definitiva, les variables que ens han quedat després de la neteja de la base de dades inicial recolzen les característiques de la malaltia que havíem introduït a la part teòrica del treball.

Per acabar, com hem dit anteriorment el model que hem construït només és un sistema de suport per ajudar als professionals de la salut a decidir el diagnòstic tenint en compte les condicions individuals de cada dona, el resultat d'aquest no ha de substituir l'opinió del metge només corroborar-la juntament amb les anàlisis clíniques.

6 Referències

[1] W. P. Rodas Torres, A. E. Mawyin Juez, J. L. Gómez González, C. V. Rodríguez Barzola, D. G. Serrano Vélez, D. A. Rodríguez Torres, R. E. López Pazmiño and R. D. Montes Nájera, "Diabetes gestacional: fisiopatología, diagnóstico, tratamiento y nuevas perspectivas," *Archivos Venezolanos de Farmacología y Terapéutica,* vol. 37, no. 3, pp. 218-226, 2018.

[2] H. García Alcalá, "Diagnóstico y clasificación de la diabetes," in *Manual práctico del manejo de la diabetes mellitus y sus comorbilidades*, 2 ed., vol. 1, Ciudad de México, Alfil, 2019, pp. 15-25.

[3] H. Paris J and L. Sarah L, "Physiology, Glucose.," StatPearls. Treasure Island (FL), 20 September 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK545201/. [Accessed 14 April 2022].

[4] V. Castrejón, R. Carbó and M. Martínez, "Mecanismos moleculares que intervienen en el transporte de la glucosa," *Revista de Educación Bioquímica,* vol. 26, no. 2, pp. 49-57, 2007.

[5] R. R. Preston and T. E. Wilson, "Páncreas e Hígado endocrinos," in *Fisiología*, 1 ed., Barcelona, Wolters Kluwe Health, 2013, pp. 411-420.

[6] Tech school of medicine, "Homeostasis en la glucosa," 15 Mars 2021. [Online]. Available: https://www.techtitute.com/medicina/blog/homeostasis-en-la-glucosa. [Accessed 14 April 2022].

[7] R. Zacarías Castillo and É. K. Tenorio Aguirre, "Páncreas y célula beta," in *Manual práctico del manejo de la diabetes mellitus y sus comorbilidades*, 2 ed., vol. 1, Ciudad de México, Alfil, 2019, pp. 47-59.

[8] Generalitat de Catalunya. Departament de salut, "Diabetis. Canal Salut," CatSalut, [Online]. Available: https://canalsalut.gencat.cat/ca/salut-a-z/d/diabetis/. [Accessed 13 May 2022].

[9] J. A. Rodríguez Gutiérrez, C. Ochoa Martínez and R. Violante Ortiz, "Homeostasis de la glucosa," in *Manual práctico del manejo de la diabetes mellitus y sus comorbilidades*, 2 ed., vol. 1, Ciudad de México, Alfil, 2019, pp. 73-91.

[10] C. Ortega González, T. Nava Ponce and E. Milo Suárez, "Diabetes mellitus gestacional," in *Manual práctico del manejo de la diabetes mellitus y sus comorbilidades*, 2 ed., vol. 1, Ciudad de México, Alfil, 2019, pp. 309-325.

[11] Generalitat de Catalunya. Departament de salut, "Tipus de diabetis.," CatSalut, 12 November 2020. [Online]. Available: https://canalsalut.gencat.cat/ca/salut-a-z/d/diabetis/tipus-de-diabetis/. [Accessed 13 May 2022].

[12] M. P. Antón Grández, "Actualización En El Abordaje Sanitario De La Diabetes Gestacional," *NPunto,* vol. III, no. 28, pp. 4-24, July 2020.

[13] M. Eleftheriades, I. Papastefanou, I. Lambrinoudaki, D. Kappou, D. Lavranos, A. Akalestos, A. P. Souka, P. Pervanidou, D. Hassiakos and G. P. Chrousos, "Elevated placental growth factor concentrations at 11-14 weeks of gestation to predict gestational diabetes mellitus," *Metabolism: Clinical and Experimental,* vol. 66, no. 11, pp. 1419-1425, November 2014.

[14] Grupo Español de Diabetes y Embarazo (GEDE): Sociedad Española de Diabetes (SED), Sociedad Española de Ginecología y Obstetricia (SEGO) y Asociación Española de

Pediatría (Sección de Neonatología), "Guía asistencial de diabetes," *Avances en Diabetología,* vol. 22, no. 1, pp. 73-87, 2006.

[15] M. d. C. Gómez García and L. Ávila Lachica, "Diabetes gestacional," in *Guía de actualización en diabetis mellitus tipo 2*, vol. 2, Badalona, Euromedice Vivactis, 2016, pp. 237-246.

[16] D. R. Coustan, L. P. Lowe, B. E. Metzger and A. R. Dyer, "The Hyperglycemia and Adverse Pregnancy Outcome (HAPO) study: paving the way for new diagnostic criteria for gestational diabetes mellitus," *American Journal of Obstetrics and Gynecology,* vol. 202, no. 6, 2010.

[17] G. Villanueva Hernández and J. Bayón Yusta, "Diagnóstico de la diabetes mellitus gestacional: evaluación de los nuevos criterios IADPSG," País Vasco, 2014.

[18] C. N. Fernández Pombo, M. R. Luna Cano, M. Lorenzo Carpente, E. Allegue Magaz and L. Beceiro Dopico, "Importancia de la detección de factores de riesgo para diabetes mellitus gestacional," *Índex de Enfermería,* vol. 25, no. 1-2, pp. 18-21, 2016.

[19] C. Matas Rodríguez, M. Á. Del Fresno Serrano and L. Borrego Cabezas, "Cribado de la diabetes gestacional," 11 December 2021. [Online]. Available: https://revistasanitariadeinvestigacion.com/cribado-de-la-diabetes-gestacional/. [Accessed 4 June 2022].

[20] N. Arruti, A. Rebollo, G. Mezquita, A. Alcaine and J. Andonegui, "Uso de fármacos oculares en el embarazo.," *Anales del Sistema Sanitario de Navarra,* vol. 36, no. 3, pp. 479-487, 2013.

[21] D. Monroy, L. Guillermo, M. Rivera and M. Alfonso, "Conceptos preliminares," in *Análisis estadístico de datos multivariados*, Bogotá, Universidad Nacional de Colombia, 2012, pp. 3-44.

[22] M. R. Martínez Arias, "Introducción y conceptos básicos," in *El Análisis multivariante en la investigación científica*, La Muralla, 1999.

[23] D. Peña, "Introducción," in *Análisis de los datos multivariantes*, 2 ed., Madrid, McGraw-Hill, 2013, pp. 11-12.

[24] M. C. Ximénez and R. San Martín, "Introducción," in *Fundamentos de las técnicas multivariantes*, Madrid, UNED. Universidad Nacional de Educacion a Distancia, 2013, pp. 1-12.

[25] D. Peña, "Descripción de datos multivariantes," in *Análisis de los datos multivariantes*, 2 ed., Madrid, McGraw-Hill, 2013, pp. 61-101.

[26] J. A. Muñoz García and I. Amón Uribe, "Técnicas para detección de outliers multivariantes," *Revista en Telecomunicaciones e Informática,* vol. 3, no. 5, pp. 11-25, 2014.

[27] M. C. Ximénez and R. San Martín, "Nociones básicas de álgebra de matrices," in *Fundamentos de las técnicas multivariantes*, Madrid, UNED. Universidad Nacional de Educacion a Distancia, 2013, pp. 13-35.

[28] M. C. Ximénez and R. San Martín, "Análisis de Componentes Principales," in *Fundamentos de las técnicas multivariantes*, Madrid, UNED. Universidad Nacional de educación a distancia, 2013, pp. 86-99.

[29] Coding Club, "Introduction to ordination. Finding patterns in your data," [Online]. Available: https://ourcodingclub.github.io/tutorials/ordination/. [Accessed 5 August 2022].

[30] J. L. Vicente Villardón, "Teoría del Análisis de Componentes Principales," [Video] Youtube, 2013. [Online]. Available: https://www.youtube.com/watch?v=Dru4gDLFRyI&ab_channel=JoseLuisVicenteVillardon. [Accessed 5 August 2022].

[31] D. Peña, "Componentes Principales," in *Análisis de datos multivariantes*, 2 ed., Madrid, McGraw-Hill España, 2013, pp. 133-170.

[32] I. Moral Peláez, "Modelos de regresión: lineal simple y regresión logística," in *Métodos estadísticos para enfermería nefrológica*, SEDEN, 2006, pp. 195-214.

[33] D. Peña, "Discriminación logística y otros métodos de clasificación," in *Análisis de datos multivariantes*, 2 ed., Madrid, McGraw-Hill, 2013, pp. 429-455.

[34] Departamento de Educación Médica Facultad de Medicina Universidad de la República de Uruguay, "Regresión Logística Parte II," [Video] Youtube, 31 March 2020. [Online]. Available: https://www.youtube.com/watch?v=EizLkX1E-Qk&t=511s&ab_channel=dem. [Accessed 5 August 2022].

[35] E. Ortega Paéz, C. Ochoa Sangrador and M. Molina Arias, "Regresión logística binaria simple," *Evidencias en pediatría,* vol. 18, no. 1, March 2022.

[36] F. Husson and J. Josse, "Handling Missing Values with Multivariate Data Analysis," 9 December 2020. [Online]. Available: https://cran.r-project.org/web/packages/missMDA/missMDA.pdf. [Accessed 5 August 2022].

7 Índex de figures

8 Índex de taules

Annexos

A continuació, adjuntem el codi utilitzat per l'anàlisi multivariant i també, un enllaç que condueix a un fitxer .html que es pot descarregar per visualitzar el codi en format Rmarkdown des d'un navegador.

Enllaç a l'informe RMarkdown

Enllaç: https://drive.google.com/file/d/1vcqTtOB2VC2Te88qgdvPj7KjAdUEkvpK/view?usp=sharing

Codi del fitxer .R

```
##-----------------------------------------------------------------------------------------------------------------
   # TÍTOL: Anàlisi de la diabetis gestacional i els seus factors de risc mitjançant models matemàtics
   # ALUMNA: MIRANDA SILVERIA MANZANARES
   # DIRECTOR: Dr. AGUSTÍ SOLANAS
   # GRAU D'ENGINYERIA BIOMÈDICA URV 2021/2022
   # L'objectiu d'aquest codi és preparar les dades i implementar un anàlisi multivariant per construir
   # un model de regressió logística que detecti la diabetis gestacional.

## 1. IMPORTACIÓ DE LES DADES-------------------------------------------------------------------------------

  setwd("C:/Users/34668/OneDrive/Escritorio/TFG")

  #Importem l'arxiu .sav en un dataframe:

  library(foreign)
  rawdata<-
as.data.frame(read.spss("C:/Users/34668/OneDrive/Escritorio/TFG/DiabetisGestacional_Abril_2021/DMG_AS.sav"))

## 2. NETEJA DE LES DADES-----------------------------------------------------------------------------------

### 2.1. SELECCIÓ DELS GRUPS D'OBSERVACIÓ-----------------------------------------------------------------

  #Dimensió de la base de dades:
  dim(rawdata)

  #Grups d'observació:
  summary(rawdata$grupo_dg)

  #Ens quedem només amb pacients controls i amb diabetis gestacional comú.

  data<- rawdata[(rawdata$grupo_dg=="controles")|(rawdata$grupo_dg=="DMG"),]

  data<-data[!is.na(data$grupo_dg),]

  #Eliminem els nivells que ja no necessitem:

  data$grupo_dg <- droplevels(data$grupo_dg)

  #Canviem el nom de l'etiqueta controles a CONTROLS

  data$grupo_dg<- factor(data$grupo_dg,levels=c("controles","DMG"),labels=c("CONTROLS","DMG"))
```

```
summary(data$grupo_dg)

### 2.2. SELECCIÓ DE VARIABLES AMB MENYS VALORS PERDUTS----------------------------------------

#CÀLCUL DEL PERCENTATGE DE NULS DE CADA VARIABLE:

# Calculem el número de valors nuls que té cada variable:

NA_V<-as.data.frame(apply(X = is.na(data), MARGIN = 2, FUN = sum),col.names=c("nuls"))

#Calculem el percentatge de nuls de cada variable:

perc<-vector()

for(i in 1:1090){
  perc[i]<-(NA_V[i,1]/3194)*100
}

#Guardem el nom de les variables:

nom_var<-colnames(data)

#Unim els valors en un dataframe ordenem les variables de menys a més percentatge de nuls

df_NA_VAR<-cbind(nom_var,NA_V,perc)

df_NA_VAR<-df_NA_VAR[order(df_NA_VAR[,2],decreasing = FALSE), ]

#Afegim la posició de la llista i el nom de les columnes

df_NA_VAR<-cbind(df_NA_VAR,num=c(1:1090))

colnames(df_NA_VAR)<-c("Variable","NA","Percentatge de NA","Posició llista")

head(df_NA_VAR)


#IMPRIMIM LES CADENES I APUNTEM ELS ESPAIS

print(data$OTRASCOMPL[1])# "              "
print(data$PAQUETES[1])#" "
print(data$OTROSTX[1])#"                      "
print(data$ENFERMASOC[1])#"                "
print(data$TIPO[1])#"                  "
print(data$PESO_M3[1])#"         "
print(data$TIPOMALF[1])#"                          "
print(data$TIPOCOMPL[1])#"                                    "
print(data$HIJO[1])#"         "
print(data$ttos[1])#"                       "
print(data$MED_2A[1])#"                   "
print(data$TX_3A[1])#"       "
print(data$MEDICACION[1])#"     "
print(data$VAR00011[2])#"                "
print(data$VAR00018[2])#"                "

#TRANSFORMEM ELS ESPAIS EN NA:
```

```
for (i in 1:3854){

  if(isTRUE(data$OTRASCOMPL[i]=="               ")==TRUE){
    data$OTRASCOMPL[i]<-NA
  }

  if(isTRUE(data$PAQUETES[i]==" ")==TRUE){
    data$PAQUETES[i]<-NA
  }

  if(isTRUE(data$OTROSTX[i]=="                          ")==TRUE){
    data$OTROSTX[i]<-NA
  }

  if(isTRUE(data$ENFERMASOC[i]=="                   ")==TRUE){
    data$ENFERMASOC[i]<-NA
  }

  if(isTRUE(data$TIPO[i]=="                  ")==TRUE){
    data$TIPO[i]<-NA
  }

  if(isTRUE(data$PESO_M3[i]==""         ")==TRUE){
    data$PESO_M3[i]<-NA
  }

  if(isTRUE(data$TIPOMALF[i]=="                           ")==TRUE){
    data$TIPOMALF[i]<-NA
  }

  if(isTRUE(data$TIPOCOMPL[i]=="                                    ")==TRUE){
    data$TIPOCOMPL[i]<-NA
  }

  if(isTRUE(data$HIJO[i]==""        ")==TRUE){
    data$HIJO[i]<-NA
  }

  if(isTRUE(data$ttos[i]="                          ")==TRUE){
    data$ttos[i]<-NA
  }

  if(isTRUE(data$MED_2A[i]=="                  ")==TRUE){
    data$MED_2A[i]<-NA
  }

  if(isTRUE(data$TX_3A[i]==""       ")==TRUE){
    data$TX_3A[i]<-NA
  }

  if(isTRUE(data$MEDICACION[i]==""      ")==TRUE){
    data$MEDICACION[i]<-NA
  }

  if(isTRUE(data$VAR00011[i]="                  ")==TRUE){
    data$VAR00011[i]<-NA
  }
```

```
  if(isTRUE(data$VAR00018[i]=="                  ")==TRUE){
    data$VAR00018[i]<-NA
  }

}

# COMPROVEM QUE EL CANVI S'HA FET CORRECTAMENT:
#OTRASCOMPL:
nuls<-as.numeric(sum(is.na(data$OTRASCOMPL)))
NOnuls<-as.numeric(sum(!is.na(data$OTRASCOMPL)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#PAQUETES:
nuls<-as.numeric(sum(is.na(data$PAQUETES)))
NOnuls<-as.numeric(sum(!is.na(data$PAQUETES)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#OTROSTX:
nuls<-as.numeric(sum(is.na(data$OTROSTX)))
NOnuls<-as.numeric(sum(!is.na(data$OTROSTX)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#ENFERMASOC:
nuls<-as.numeric(sum(is.na(data$ENFERMASOC)))
NOnuls<-as.numeric(sum(!is.na(data$ENFERMASOC)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#TIPO:
nuls<-as.numeric(sum(is.na(data$TIPO)))
NOnuls<-as.numeric(sum(!is.na(data$TIPO)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#PESO_M3:
nuls<-as.numeric(sum(is.na(data$PESO_M3)))
NOnuls<-as.numeric(sum(!is.na(data$PESO_M3)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
```

```
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#TIPOMALF:
nuls<-as.numeric(sum(is.na(data$TIPOMALF)))
NOnuls<-as.numeric(sum(!is.na(data$TIPOMALF)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#TIPOCOMPL:
nuls<-as.numeric(sum(is.na(data$TIPOCOMPL)))
NOnuls<-as.numeric(sum(!is.na(data$TIPOCOMPL)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#HIJO:
nuls<-as.numeric(sum(is.na(data$HIJO)))
NOnuls<-as.numeric(sum(!is.na(data$HIJO)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#ttos:
nuls<-as.numeric(sum(is.na(data$ttos)))
NOnuls<-as.numeric(sum(!is.na(data$ttos)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#MED_2A:
nuls<-as.numeric(sum(is.na(data$MED_2A)))
NOnuls<-as.numeric(sum(!is.na(data$MED_2A)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#TX_3A:
nuls<-as.numeric(sum(is.na(data$TX_3A)))
NOnuls<-as.numeric(sum(!is.na(data$TX_3A)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#MEDICACION:
```

```
nuls<-as.numeric(sum(is.na(data$MEDICACION)))
NOnuls<-as.numeric(sum(!is.na(data$MEDICACION)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#VAR00011:
nuls<-as.numeric(sum(is.na(data$VAR00011)))
NOnuls<-as.numeric(sum(!is.na(data$VAR00011)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")

#VAR00018:
nuls<-as.numeric(sum(is.na(data$VAR00018)))
NOnuls<-as.numeric(sum(!is.na(data$VAR00018)))
TOTAL<-nuls + NOnuls
sprintf("NULS: %i", nuls)
sprintf("NO NULS: %i", NOnuls)
sprintf("TOTAL: %i", TOTAL)
sprintf("--------------")


#SELECCIÓ DE VARIABLES AMB MENYS VALORS PERDUTS:

# Calculem el número de valors nuls que té cada variable:

NA_V<-as.data.frame(apply(X = is.na(data), MARGIN = 2, FUN = sum),col.names=c("nuls"))

#Calculem el percentatge de nuls de cada variable:

perc<-vector()

for(i in 1:1090){
  perc[i]<-(NA_V[i,1]/3194)*100
}

#Guardem el nom de les variables:

nom_var<-colnames(data)

#Unim els valors en un dataframe ordenem les variables de menys a més percentatge de nuls

df_NA_VAR<-cbind(nom_var,NA_V,perc)

df_NA_VAR<-df_NA_VAR[order(df_NA_VAR[,2],decreasing = FALSE), ]

#Afegim la posició de la llista i el nom de les columnes

df_NA_VAR<-cbind(df_NA_VAR,num=c(1:1090))

colnames(df_NA_VAR)<-c("Variable","NA","Percentatge de NA","Posició llista")

print(head(df_NA_VAR))
```

```
  #Seleccionem el 10% amb més valors disponibles, es a dir, les 109 primeres variables de la llista.
  data<-subset(data, select = df_NA_VAR[1:109,1])#Nova data frame

### 2.3. SELECCIÓ DE PACIENTS AMB MENYS VALORS PERDUTS----------------------------------------

  #Calculem el número de NAs de cada pacient:
  NA_P<-as.data.frame(apply(X = is.na(data), MARGIN = 1, FUN = sum))

  percent<-vector()

  #Calculem el percentatge de nuls de cada pacient:

  for(i in 1:3194){
    percent[i]<-(NA_P[i,1]/109)*100
  }

  #Construim la taula de dades

  df_NA_PAC<-cbind(data[,"CODIGO_ID"],NA_P,percent)

  df_NA_PAC<-df_NA_PAC[order(df_NA_PAC[,2],decreasing = FALSE), ]

  df_NA_PAC<-cbind(df_NA_PAC,num=c(1:3194))#Posició de la llista

  colnames(df_NA_PAC)<-c("CODI_ID","NA","Percentatge_NA","num")


  #Ens quedem amb les pacients que tenen com a mínim el 70% de les dades disponibles, per tant
tallem a la 1818 de la llista:
  CODIS<-df_NA_PAC[1:1818,1]#CODIS ID de les 1818 primeres pacients
  data1<- data[CODIS,]

### 2.4. ANÀLISI DE DEPENDENCIES LINEALS-------------------------------------------------------------
  numeriques<-data1[ , unlist(lapply(data1, is.numeric))]
  qualitatives<-data1[ , (unlist(lapply(data1, is.factor)))]

  #Arreglem el format de les dates:
  numeriques$FECHA_TTOG<-as.Date(numeriques$FECHA_TTOG/86400 ,origin = "1582-10-14")
  numeriques$FECHA_ANAL_INICIAL<-as.Date(numeriques$FECHA_ANAL_INICIAL/86400 , origin =
"1582-10-14")
  numeriques$FECHAPV<-as.Date(numeriques$FECHAPV/86400 , origin = "1582-10-14")
  numeriques$FUR<-as.Date(numeriques$FUR/86400 , origin = "1582-10-14")
  numeriques$FPARTO<-as.Date(numeriques$FPARTO/86400 , origin = "1582-10-14")

  #Treiem les dates del conjunt de dades numèriques:
  quantitatives<-numeriques[ , unlist(lapply(numeriques, is.numeric))]

  #Treiem les variables referents a dates i afegim el codi ID del pacient:
  dates<-cbind("CODI_ID"=quantitatives[,1],numeriques[ , !unlist(lapply(numeriques, is.numeric))])

  #Eliminem el codi identificatiu de les pacients de les variables quantitatives ja que no es aporta res:
  quantitatives$CODIGO_ID<-NULL

  #Si utilitzem les variables source01 i DMGTTOG hi ha NA a la matriu i ens dona problemes en la
representació i també les eliminem.
  quantitatives$source01<-NULL
  quantitatives$DMGTTOG<-NULL
```

```
#### 2.4.1. MATRIU DE CORRELACIÓ: Dependències lineals per parells de variables----------

  correlacions<-data.frame()

  #Càlcul de la matriu sense tenir en compte els NA:
  correlacions<-cor(quantitatives,use="pairwise.complete.obs")

  #Arrodonim a 3 decimals:
  round(correlacions, digits=3)


  library(corrplot)
  color<-colorRampPalette(c("#BB4444","#EE9988","#FFFFFF","#77AADD","#26456E"))

  #REPRESENTACIÓ GRÀFICA DE LA MATRIU DE CORRELACIONS:

  corrplot(correlacions,main = "\n\n\n Representació gràfica de la matriu de
correlacions",cex.main=3,tl.col="black",
           tl.srt=45,shade.col=NA,addCoef.col =
"black",order="AOE",type="lower",diag=F,method="shade",col=color(100))


#### 2.4.2. COEFICIENTS DE CORRELACIÓ MÚLTIPLE-----------------------------------------------------------


  #CÀLCUL DEL QUADRAT DEL COEFICIENT DE CORRELACIÓ MÚLTIPLE

  covariancies<-cov(quantitatives,use="pairwise.complete.obs")#Covariàncies
  invcov<-solve(covariancies)#Inversa de la matriu de covariàncies

  #Multipliquem els valors de les diagonals de la matriu de variancies i de la seva inversa

  v<-diag(covariancies)*diag(invcov)
  v2<-vector()

  for (i in 1:46){

    v2[i]<-solve(v[i])#Calculem la inversa de cada resultat

  }

  coefs<-vector()

  for (i in 1:46){

    coefs[i]<-(1-solve(v[i]))#Li restem a 1 el valor de la inversa de cada element

  }

  #Contruim la taula

  coefs<-as.data.frame(coefs)

  noms<-colnames(quantitatives)

  coefs<-cbind(noms,coefs)
```

```
  #ordenem de més a menys
  coefs<-coefs[order(coefs[,2],decreasing = TRUE), ]

  head(coefs)


#### 2.4.3. ELIMINACIÓ DE LES VARIABLES REDUNDANTS----------------------------------------------

  # ELIMINACICÓ DE LES VARIABLES REDUNDANTS:

  quantitatives$TALLA<-NULL
  quantitatives$TALLA2<-NULL
  quantitatives$AUMENTOP<-NULL
  quantitatives$BMIFINAL<-NULL
  quantitatives$BMIPV<-NULL
  quantitatives$BMIPREG<-NULL
  quantitatives$INCBMIPV<-NULL
  quantitatives$PESO_F<-NULL
  quantitatives$PESO_PR<-NULL
  quantitatives$INCBMITOTAL<-NULL
  quantitatives$N_PUNTOS_NDDMG<-NULL
  quantitatives$GRUPOBMIPREG2<-NULL
  quantitatives$GRUPOBMIPREG<-NULL


#### 2.4.4. AVALUACIÓ DEL RESULTAT DE L'ELIMINACIÓ DE REDUNDÀNCIA----------------------------

  # ANÀLISI DE LES DEPENDENCIES LINEALS POSTERIOR A L'ELIMINACIÓ DE LES VARIABLES
REDUNDANTS:

  #Matriu de correlació:
  correlacions2<-cor(quantitatives,use="pairwise.complete.obs")#No es tenen en compte els NA
  round(correlacions2, digits=3)

  #Representació gràfica:
  corrplot(correlacions2,main = "\n\n\n Representació gràfica de la matriu de
correlacions",cex.main=3,tl.col="black",
           tl.srt=45,shade.col=NA,addCoef.col =
"black",order="AOE",type="lower",diag=F,method="shade",col=color(100))

  #Coeficients de correlació múltiple al quadrat:

  covariancies2<-cov(quantitatives,use="pairwise.complete.obs")#Covariancies
  invcov2<-solve(covariancies2)#Inversa de la matriu

  v<-diag(covariancies2)*diag(invcov2)#Multipliquem els valors de les diagonals de la matriu de
variancies i la seva inversa

  v2<-vector()


  for (i in 1:33){

    v2[i]<-solve(v[i])#Calculem la inversa de cada resultat

  }

  coefs2<-vector()
```

```
  for (i in 1:33){

    coefs2[i]<-(1-solve(v[i]))#Li restem a 1 el valor de la inversa de cada element

  }


  coefs2<-as.data.frame(coefs2)

  noms<-colnames(quantitatives)

  coefs2<-cbind(noms,coefs2)

  #ordenem de més a menys
  coefs2<-coefs2[order(coefs2[,2],decreasing = TRUE), ]

  head(coefs2)


#### 2.5. ELIMINACIÓ DE VALORS ATÍPICS----------------------------------------------------------------------

  #DETECCIÓ D'OUTLIERS AMB LA DISTÀNCIA DE MAHALANOBIS:

  library(mvoutlier)
  library(modi)

  vectMean<-vector()

  #Vector de mitjanes:
  for (i in 1:length(quantitatives)){
    vectMean[i]<-mean(quantitatives[,i], na.rm = TRUE)
  }

  #Càlcul de les distàncies de Mahalanobis a partir del conjunt de dades, el vector de mitjanes i la
matriu de covariàncies:

  MD<-MDmiss(quantitatives, vectMean, covariancies2)

  #Afegim el resultat a la dataframe:
  quantitatives2<-cbind(quantitatives,MD)

  #Dibuixem l'histograma de freqüències:
 hist(MD,main = "Hisograma de freqüències de les distàncies de Mahalanobis",xlab = "Distància de
Mahalanobis",ylab="Freqüència")

  #ELIMINACIÓ DELS CASOS EN QUE LA DISTÀNCIA DE MAHALANOBIS ÉS SUPERIOR A 100

  #Juntem per eliminar també del conjunt de quantitatives:
  df<-cbind(qualitatives,quantitatives2)

  #Selecció:
  df<-subset(df,(df[,87]<100))

  #Tornem a separar els datasets:

  quant<-df[ , unlist(lapply(df, is.numeric))]
  qual<-df[ , (unlist(lapply(df, is.factor)))]
```

```
  varName<-colnames(quant)

## 3. CONSTRUCCIÓ DEL MODEL----------------------------------------------------------------------------------

### 3.1. CONJUNT D'ENTRENAMENT I EL D'AVALUACIÓ-------------------------------------------------------

  #GENEREM EL CONJUNT D'ENTRENAMENT I EL D'AVALUACIÓ:

  #Creem el training set amb el 80% dels individus i afegim el grup:
  training_set<-cbind("grupo_dg"=qual[1:1026,1],quant[1:1026,])#80%

  #Creem el testing set amb el 20% dels individus restants i afegim el grup:
  testing_set<-cbind("grupo_dg"=qual[1027:1283,1],quant[1027:1283,])#20%


### 3.2. ANÀLISI DE COMPONENTS PRINCIPALS-------------------------------------------------------------

  #ANÀLISI DE COMPONENTS PRINCIPALS DEL TRAINING DATASET:

  #Cridem les llibreries que necessitem

  library(MASS)
  library(FactoMineR)
  library(missMDA)
  library(nnet)
  library(factoextra)

  #Completem les dades amb imputePC, escalem els valors i ajustem a 4 components principals

  trainingpc<-imputePCA(scale(training_set[2:35]),ncp=4)
  training<-trainingpc$fittedX

  testingpc<-imputePCA(scale(testing_set[2:35]),ncp=4)
  testing<-testingpc$fittedX

  colnames(training)<-varName
  colnames(testing)<-varName


  #Calculem els components principals del dataset d'aprenentatge
  pcs<-prcomp(training)

  #Actualitzem les puntuacions del conjunt d'entrenament i testeig sobre les components principals
generades:

  #Training set:

  trg<-predict(pcs,training)
  trg<-data.frame(trg,"grupo_dg"=training_set[,1])


  #Testing set:

  tst<-predict(pcs,testing)
  tst<-data.frame(tst,"grupo_dg"=testing_set[,1])
```

```
#Gràfica dels percentatges de variància absorbida per cada CP:

fviz_eig(pcs, addlabels = TRUE, ylim = c(0, 60),
        title="Percentatge de la variància absorbida per cada component principal",
        ylab="Percentatge de variància (%)",
        xlab= "Components principals")

#Gràfica de les contribucions de les variables sobre CP1 i CP2:
fviz_pca_var(pcs,
           col.var = "contrib", # Pintem per les contribucions a la CP
           gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
           repel = TRUE,
           title="PC1-PC2"
)

groups <- as.factor(training_set[,1])

#Gràfica de les puntuacions dels individus sobre CP1 i CP2:
fviz_pca_ind(pcs,
           col.ind = groups, # Pintem els grups
           palette = c("#00AFBB",  "#FC4E07"),
           ellipse.type = "confidence",
           legend.title = "Grups",
           title="PC1-PC2"
)


#Gràfica de les contribucions de les variables sobre CP1 i CP3:
fviz_pca_var(pcs,
           axes = c(1, 3),
           col.var = "contrib", # Pintem per les contribucions a la CP
           gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
           repel = TRUE,
           title="PC1-PC3"
)


#Gràfica de les puntuacions dels individus sobre CP1 i CP3:
fviz_pca_ind(pcs,
           axes = c(1, 3),
           col.ind = groups, # Pintem els grups
           palette = c("#00AFBB",  "#FC4E07"),
           ellipse.type = "confidence",
           legend.title = "Grups",
           title="PC1-PC3"
)



### 3.3. CONSTRUCCIÓ DEL MODEL DE REGRESSIÓ LOGÍSTICA----------------------------------------

trg$grupo_dg<-relevel(trg$grupo_dg,ref = "CONTROLS")

#Construcció del model de classificació de grupo_dg amb les 4 components principals:
logistic<-glm(grupo_dg~PC1+PC2+PC3+PC4,data=trg,family = "binomial")

#Característiques:
summary(logistic)
```

```
### 3.4. AVALUACIÓ DEL MODEL-------------------------------------------------------------------------------

  #llibreries:
  library(caret)
  library(pROC)

#### 3.4.1. AVALUACIÓ DE LA PREDICCIÓ AMB LES DADES D'ENTRENAMENT--------------------------

  # Guardem les predicions del model sobre les dades training.
  pdata1 <- predict(logistic, type = "response")


  #Representació gràfica:
  pred <- data.frame(probDMG=pdata1,DMG=trg$grupo_dg)
  pred<- pred[order(pred$probDMG, decreasing=FALSE),]
  pred$rank <- 1:nrow(pred)

  library(ggplot2)
  ggplot(data=pred, aes(x=rank, y=probDMG)) +
    geom_point(aes(color=DMG), alpha=1, shape=4, stroke=2) +
    xlab("Index") +
    ylab("Possibilitat de patir diabetis gestacional")+
    ggtitle("Model de regressió logística amb les dades d'entrenament")+
    labs(color='Grup d´observació')+
    geom_hline(yintercept = 0.5)

  #Matriu de confusió:
  trg$grupo_dg<-relevel(trg$grupo_dg,ref = "CONTROLS")
  tab=table(pdata1>0.5,trg$grupo_dg)#Es selecciona el grup DMG si la probabilitat >0.5

  tab
  #FALSE = 1 = CONTROLS
  #TRUE = 2 =DMG

  #Calculem l'eficàcia del model
  sprintf("Taxa de precisió: %f",(sum(diag(tab))/sum(tab))*100)

  #CORBA ROC:
  res.roc1 <- roc(as.numeric(trg$grupo_dg), pdata1)
  plot.roc(res.roc1, print.auc = TRUE,main="Corba ROC i Àrea sota la corba amb les dades
d'entrenament")

#### 3.4.2. AVALUACIÓ DE LA PREDICCIÓ AMB LES DADES TESTING-------------------------------------

  # Guardem les predicions del model sobre les noves dades:
  pdata2 <- predict(logistic, newdata = tst, type = "response")

  #Representació gràfica:

  pre <- data.frame(probDMG=pdata2,DMG=tst$grupo_dg)
  pre<- pre[order(pre$probDMG, decreasing=FALSE),]
  pre$rank <- 1:nrow(pre)

  library(ggplot2)
  ggplot(data=pre, aes(x=rank, y=probDMG)) +
    geom_point(aes(color=DMG), alpha=1, shape=4, stroke=2) +
    xlab("Index") +
```

```
    ylab("Possibilitat de patir diabetis gestacional")+
    ggtitle("Model de regressió logística amb les dades d'avaluació")+
    labs(color='Grup d´observació')+
    geom_hline(yintercept = 0.5)


  #Matriu de confusió:
  tst$grupo_dg<-relevel(tst$grupo_dg,ref = "CONTROLS")
  tab2=table(pdata2>0.5,tst$grupo_dg)#Es selecciona el grup DMG si la probabilitat >0.5

  tab2
  #FALSE = 1 = CONTROLS
  #TRUE = 2 =DMG


  #Calculem l'eficàcia del model:
  sprintf("Taxa de precisió: %f",(sum(diag(tab2))/sum(tab2))*100)

  #Corba ROC:
  res.roc2 <- roc(as.numeric(tst$grupo_dg), pdata2)
  plot.roc(res.roc2, print.auc = TRUE,main="Corba ROC i Àrea sota la corba amb les dades
d´avaluació")
```