Anàlisi de factors de risc en diabetis amb models de classificació

Jaume Inglada, Mercè Mateu, Cristina Tuá, Melissa Vargas

17 de juny de 2025

Universitat de Barcelona - Universitat Politècnica de Catalunya
Grau en Estadística
Assignatura: Estadística per a les Biociències

Índex

1 Introducció

En l'actual era de la informació, l'anàlisi de dades s'ha consolidat com una eina essencial per a la comprensió de fenòmens complexos en les Biociències, així com en nombrosos altres àmbits del coneixement. Entre les diverses aplicacions biomèdiques, la detecció precoç de malalties constitueix un dels camps d'estudi més rellevants.

En aquest context, l'estudi actual pretén desenvolupar un model predictiu que determini si un pacient pateix diabetis, utilitzant un conjunt de dades clíniques i demogràfiques. Mitjançant tècniques d'aprenentatge automàtic, es busquen models de classificació precisos per identificar la presència d'aquesta malaltia en individus.

Per a la realització de l'anàlisi s'ha utilitzat conjunt de dades Pima Indians Diabetes proporcionat pel National Institute of Diabetes and Digestive and Kidney Diseases (Estats Units). Aquesta base de dades conté dades de pacients de sexe biològic femení majors de 21 anys, amb ascendència "Pima", un grup ètnic indígena nord-americà originari de l'estat d'Arizona.

Els resultats obtinguts podrien contribuir a millorar estratègies de diagnòstic precoç en poblacions amb risc elevat, mostrant el potencial de les aproximacions basades en dades en salut pública.

2 Materials i Mètodes

2.1 Naturalesa de les dades

El conjunt de dades analitzat conté 768 observacions i 9 variables clíniques i demogràfiques numèriques, amb les següents característiques:

- Pregnancies: Nombre d'embarassos (interval: 0-17)

- Glucose: Concentració de glucosa en plasma en prova de tolerància (0-199 mg/dL)

- BloodPressure: Pressió arterial diastòlica en mmHg (0-122)

- SkinThickness: Gruix del plec cutani tricipital en mm (0-99)

- Insulin: Concentració d'insulina en sèrum a 2 hores (0-846 µU/mL)

- BMI: Índex de massa corporal en kg/m² (0-67.1)

- DiabetesPedigreeFunction: Probabilitat genètica de diabetis basada en l'herència familiar (0.08-2.42)

- Age: Edat del pacient en anys (21-81)

- Outcome: Variable dicotòmica (1: diabètic, 0: no diabètic)

Per realitzar una primera aproximació a l'estructura i contingut del conjunt de dades, es mostren els sis primers registres de la base de dades, que inclouen les nou variables d'estudi:

```
head(diabetes)
```

```
## # A tibble: 6 x 9
##   Pregnancies Glucose BloodPressure SkinThickness Insulin   BMI
##         <dbl>   <dbl>         <dbl>         <dbl>   <dbl> <dbl>
## 1           6     148            72            35       0  33.6
## 2           1      85            66            29       0  26.6
## 3           8     183            64             0       0  23.3
## 4           1      89            66            23      94  28.1
## 5           0     137            40            35     168  43.1
## 6           5     116            74             0       0  25.6
## # i 3 more variables: DiabetesPedigreeFunction <dbl>, Age <dbl>, Outcome <fct>
```

Complementàriament, es presenta la distribució de freqüències de les observacions segons la variable resposta. Aquest desequilibri és rellevant per a la selecció de mètriques d'avaluació del model (precisió, sensibilitat, F1-score) i la interpretació dels valors predictius del model.

clase	Frecuencia
No	500
Si	268

Aquesta distribució mostra que aproximadament un terç del grup d'estudi (34,9%) té diagnòstic positiu de diabetis, mentre que els dos terços restants (65,1%) constitueixen el grup control. Com assenyalen estudis

previs Talebi Moghaddam et al. (2024), aquest desequilibri és comú. En la següent taula es troba la mitjana i desviació estàndard de totes les variables numèriques agrupades per l'estat diabètic (Outcome). Els resultats mostren diferències clau entre grups:

- Els pacients diabètics (Outcome = "Si") presenten valors mitjans més alts en variables com Glucose i BMI, coherent amb la literatura mèdica.
- Les desviacions estàndard més altes en el grup diabètic suggereixen major variabilitat biològica.

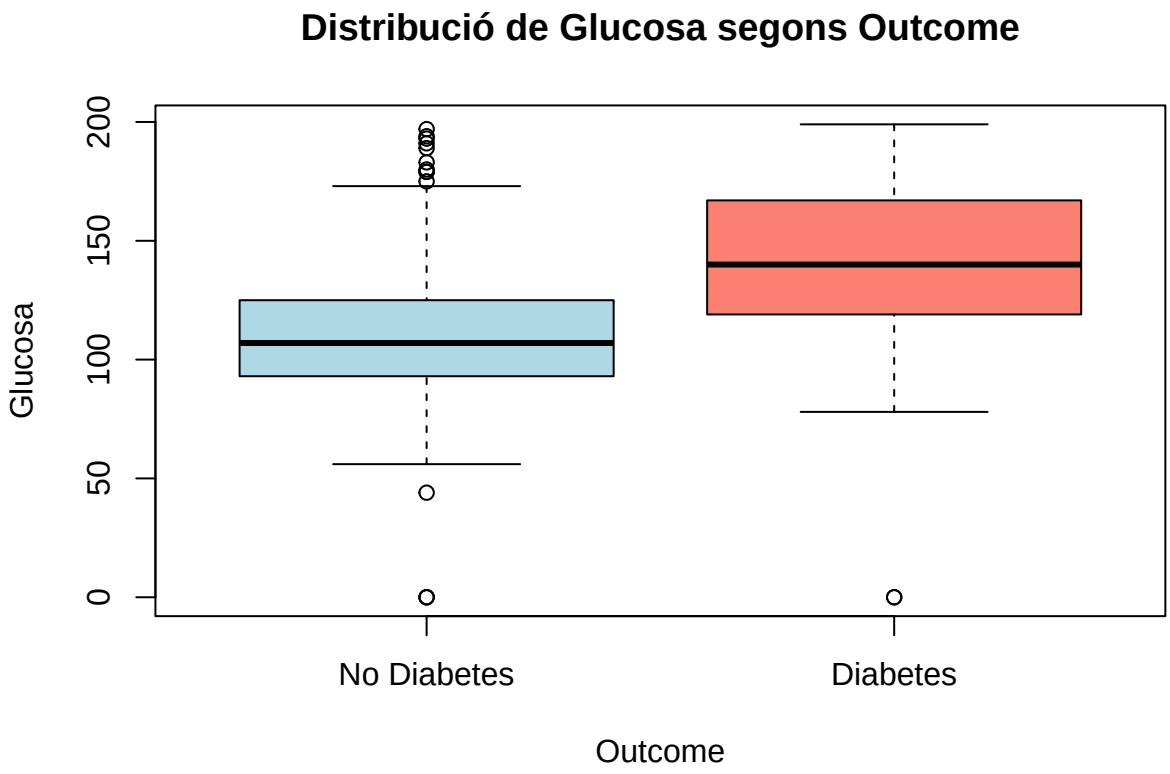
```
## # A tibble: 2 x 17
##   Outcome Pregnancies_mean Pregnancies_sd Glucose_mean Glucose_sd
##   <fct>              <dbl>          <dbl>        <dbl>      <dbl>
## 1 No                  3.30           3.02         110.       26.1
## 2 Si                  4.87           3.74         141.       31.9
## # i 12 more variables: BloodPressure_mean <dbl>, BloodPressure_sd <dbl>,
## #   SkinThickness_mean <dbl>, SkinThickness_sd <dbl>, Insulin_mean <dbl>,
## #   Insulin_sd <dbl>, BMI_mean <dbl>, BMI_sd <dbl>,
## #   DiabetesPedigreeFunction_mean <dbl>, DiabetesPedigreeFunction_sd <dbl>,
## #   Age_mean <dbl>, Age_sd <dbl>
```

Es compara amb el resum descriptiu general de les vuit variables explicatives (sense agrupar per Outcome) per veure la distribució global:

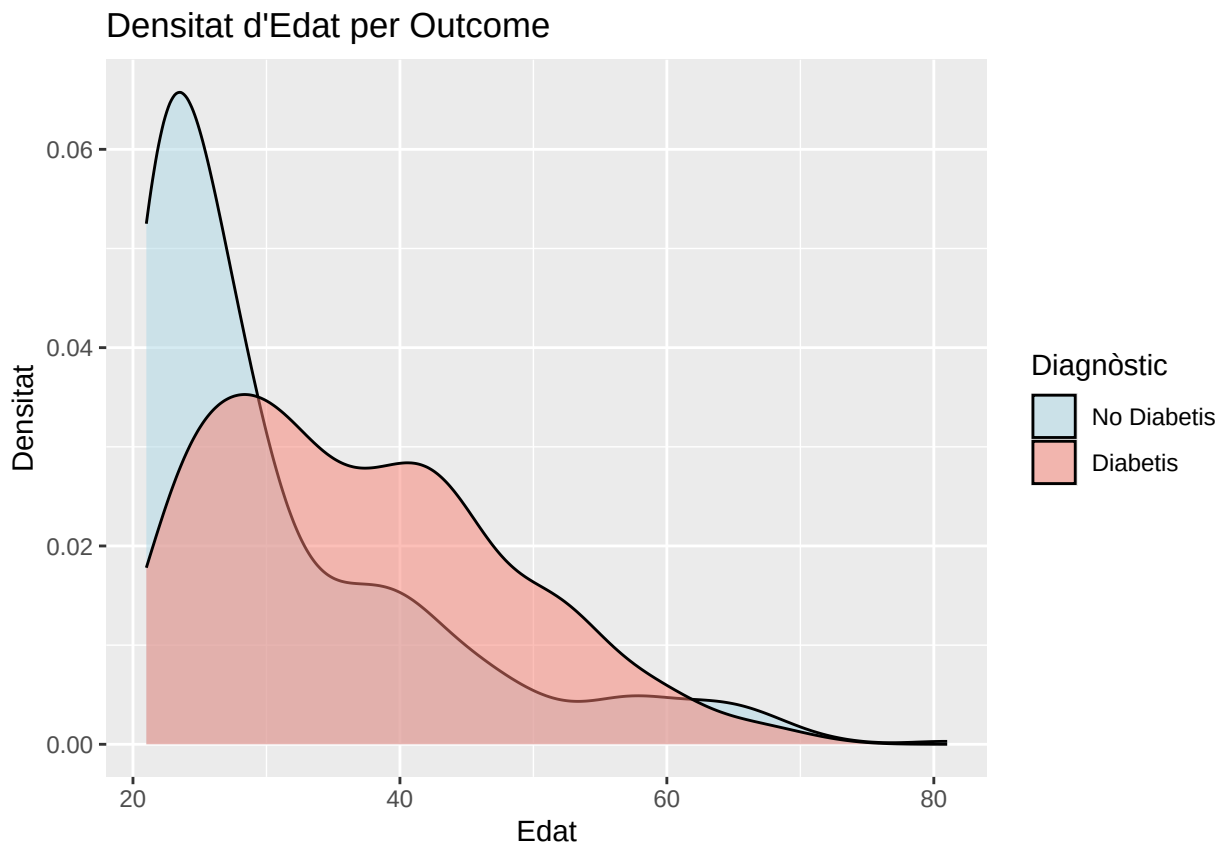
```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
```

La visualització mitjançant el següent gràfic revela diferències significatives en els nivells de glucosa entre els grups amb diabetis i sense, corroborant els resultats numèrics obtinguts:

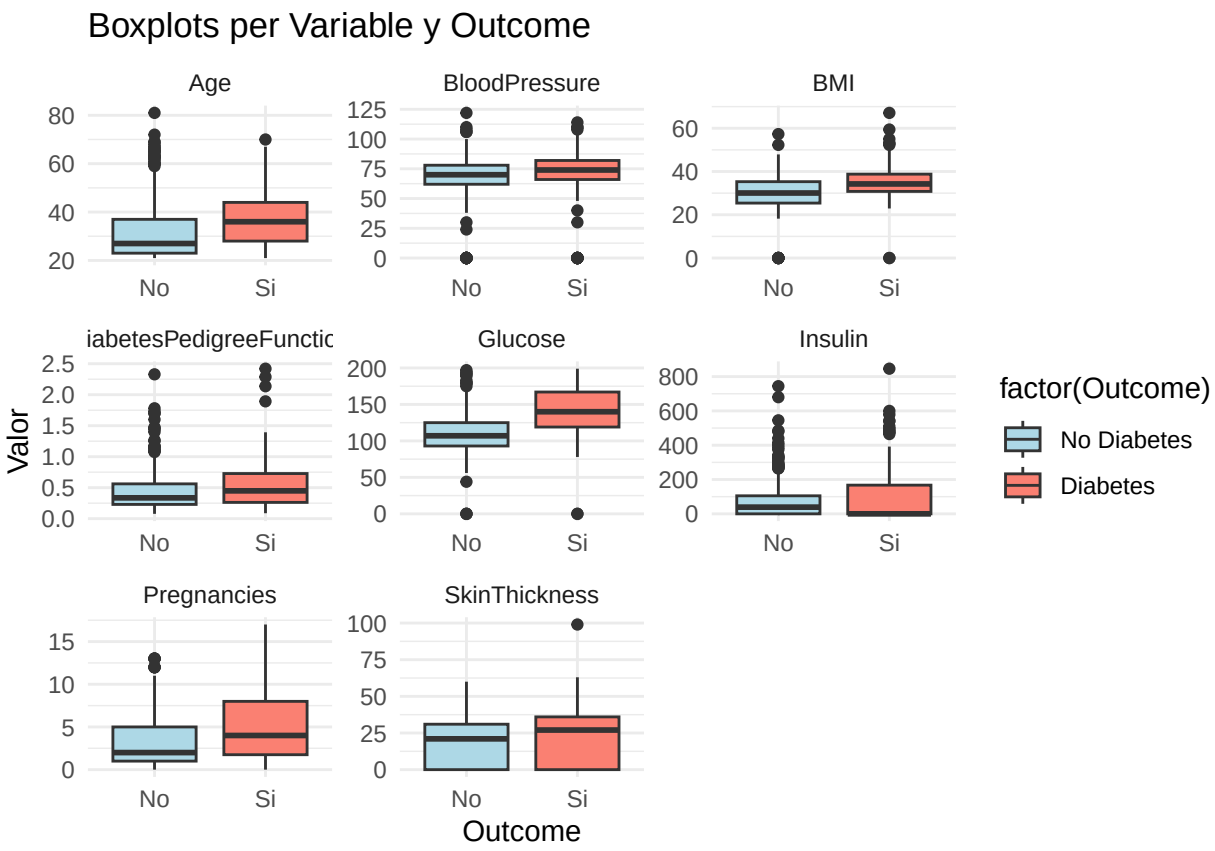
Grup diabètic (Outcome=1): - Presenta una mediana de glucosa significativament superior (140-150 mg/dL). - Major dispersió (amplada del boxplot) i alguns valors atípics. Grup no diabètic (Outcome=0): - Mediana al voltant de 100-110 mg/dL (dins dels marges considerats normals). - Menor variabilitat i més valors atípics que el grup diabètic.



El següent gràfic de densitat revela que el grup diabètic presenta una distribució d'edats més esbiaixades cap a valors superiors amb dos pics (23 i 42 anys), mentre que el grup control mostra una distribució més jove (pic 25 anys). Aquest patró coincideix amb l'evidència epidemiològica que associa la diabetis amb l'envelliment, com demostra Bierhaus et al. (1998).



A continuació el conjunt de box plots produït ofereix una comparació sistemàtica de a distribució de totes les variables numèriques en funció de l'Outcome (diabetis vs. no diabetis). Aquesta visualització és especialment útil per fer algunes observacions claus; Glucosa i BMI mostren una clara separació entre les medianes dels dos grups, corroborant la seva rellevància clínica en el diagnòstic de diabetis. Per l'altra banda, Insulina presenta una dispersió major en el grup diabètic, amb outliers evidents, possiblement indicant casos de resistència a la insulina.


Boxplots per Variable y Outcome
Age
BloodPressure
BMI
iabetesPedigreeFunctic
Glucose
Insulin
Pregnancies
SkinThickness
Valor
Outcome
No
Si
factor(Outcome)
No Diabetes
Diabetes

Mètodes utilitzats

2.2 Algorismes de classificació utilitzats

S'utilitzaran dos mètodes de classificació supervisada: k-Nearest Neighbours (k-NN) i Support Vector Machines (SVM), amb validació creuada per a l'optimització dels hiperparàmetres.

k-Nearest Neighbours (k-NN)

El mètode k-NN és un algorisme de classificació supervisada no paramètric que assigna a una nova observació la classe majoritària dels seus k veïns més propers en l'espai de característiques. La distància entre observacions es calcula mitjançant la mètrica euclidiana:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

La selecció del paràmetre k és crítica: valors petits poden conduir a sobreajustament (overfitting), mentre que valors grans poden simplificar excessivament el model. En aquest estudi, s'exploren els valors $k = \{1, 11, 21, 31\}$ mitjançant validació creuada de 3 particions (*3-fold CV*).

Support Vector Machines (SVM)

Els SVM busquen un hiperplà òptim que maximitzi el marge entre classes en un espai transformat. S'han considerat dos tipus de funcions de nucli (kernels):

Kernel lineal:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$$

Adequat per a problemes linealment separables.

Kernel radial (RBF):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right)$$

Flexible per a relacions no lineals, amb el paràmetre γ controlant la flexibilitat del model.

El paràmetre de regularització C s'ha optimitzat per a ambdós kernels, amb valors explorats en l'interval 2^{-10} a 2^{15}.

Protocol de l'estudi

Per a l'avaluació dels models, es dividirà el conjunt de dades en dos subconjunts: - Conjunt d'entrenament (train): 70% de les dades. - Conjunt de prova (test): 30% de les dades.

Aquesta partició es realitzarà de manera aleatòria fent servir la llavor 123, mantenint la proporció de les classes (Outcome) per evitar biaixos.

2.3 Obtenció de les mostres de train i test.

```
## ytrain
##  No  Si
## 334 179

## ytest
##  No  Si
## 166  89
```

3 Avaluació dels models

Model k-NN (amb caret i 3-fold CV)

Els models k-Nearest Neighbours (k-NN) s'han avaluat utilitzant el paquet caret de R, amb validació creuada de 3 particions (3-fold CV) per optimitzar el paràmetre k (nombre de veïns). S'han explorat els valors k = {1, 11, 21, 31} per determinar quin ofereix el millor rendiment en la classificació de pacients diabètics.

Taula 2: Comparativa de models k-NN amb diferents valors de k

k	Exactitud	Sensibilitat	Especificitat	Valor Predictiu Positiu	Valor Predictiu Negatiu	Kappa
1	0.682	0.777	0.506	0.746	0.549	0.288
11	0.761	0.885	0.528	0.778	0.712	0.440
21	0.749	0.874	0.517	0.771	0.687	0.414
31	0.729	0.885	0.438	0.746	0.672	0.352

k=1

Segons la matriu de confusió, el model amb k=1 presenta una precisió (accuracy) relativament baixa (68.2%) i una concordança feble (kappa=0.288). Tot i que mostra una bona sensibilitat (77.7%) per identificar casos negatius, la seva especificitat (50.6%) és insuficient per detectar casos positius. Això, juntament amb un valor predictiu negatiu baix (54.9%), indica que el model està sobre ajustat i és massa sensible a variacions petites en les dades.

k=11

El model amb k=11 millora significativament, assolint una precisió del 76.1% ($p<0.001$) i una concordança moderada (kappa=0.44). La sensibilitat és excel · lent (88.6%), però l'especificitat continua sent limitada (52.8%). Els valors predictius milloren (77.8% positiu, 71.2% negatiu), tot i mostrar un lleuger desequilibri en els errors (més falsos negatius).

k=21

Amb k=21 s'obté un equilibri òptim: precisió del 74.9%, concordança acceptable (kappa=0.414) i sensibilitat consistent (87.3%). Encara que l'especificitat (51.7%) roman baixa, els valors predictius (77.1% positiu, 68.7% negatiu) indiquen un bon rendiment global. Aquest model mostra la millor estabilitat entre tots els avaluats.

k=31

El model amb k=31 presenta una disminució en rendiment: precisió del 72.9% i concordança feble-moderada (kappa=0.352). Tot i mantenir alta sensibilitat (88.6%), l'especificitat cau fins al 43.8%, el pitjor valor de tots. Això suggereix que un k excessivament alt simplifica massa el model, perdent capacitat predictiva.

Model SVM

Taula 3: Comparativa dels models SVM: Kernel Lineal vs. Radial (RBF)

	Métriques de Rendiment					
Model	Exactitud	Sensibilitat	Especificitat	VPP	VPN	Kappa
SVM Lineal	**0.765**	**0.874**	**0.562**	**0.788**	**0.704**	**0.457**
SVM RBF	0.753	0.879	0.517	0.772	0.697	0.422

Kernel Lineal:

El model mostra una precisió global del 76,5%, amb una bona **sensibilitat (87,3%)** per detectar casos negatius (no diabètics), però una **especificitat moderada (56,2%)** per a casos positius. El valor Kappa (0,457) indica una concordança moderada, mentre que el test de McNemar (p=0,028) revela un desequilibri significatiu en els errors de classificació. Els valors predictius (**78,8% positiu, 70,4% negatiu**) suggereixen que el model és més fiable quan prediu "No diabètic". L'**exactitud equilibrada (71,8%)** reflecteix aquesta diferència en el rendiment entre classes.

Kernel Radial (RBF):

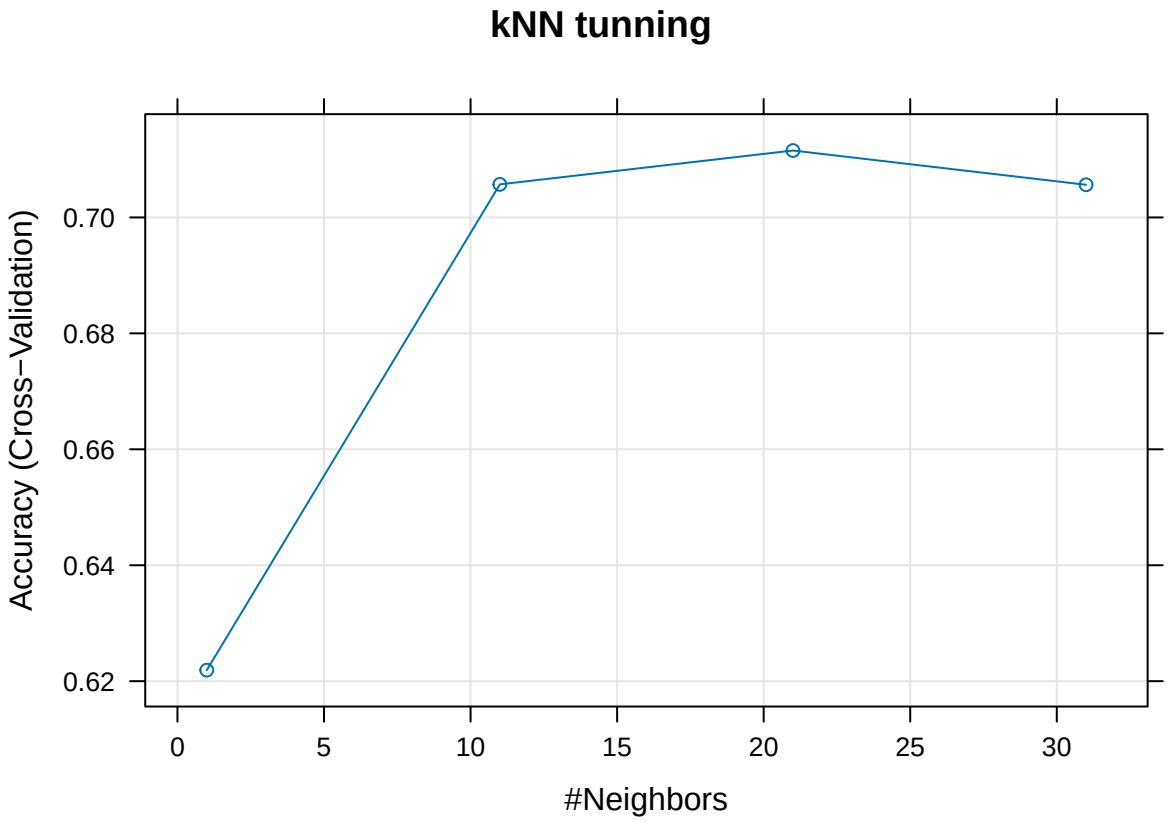
El model presenta una precisió global del 74,9%, amb una alta sensibilitat (87,9%) per identificar casos negatius ("No diabètic"), però una especificitat limitada (50,6%) per a casos positius. La concordança moderada (Kappa=0,411) i el desequilibri en els errors (test de McNemar, p=0,004) indiquen que el model tendeix a classificar incorrectament més casos diabètics. Els valors predictius (76,8% positiu, 69,2% negatiu) mostren una fiabilitat acceptable, però inferior al kernel lineal. L'exactitud equilibrada (69,3%) reflecteix la diferència de rendiment entre classes, similar al model lineal però amb una especificitat lleugerament inferior.

3.1 Optimització dels hiperparàmetres mitjançant tècniques de validació sistemàtica

Per determinar el millor model per a cada algorisme, s'han optimitzat els hiperparàmetres principals (k per k-NN, C per a SVM lineal i C amb γ per a SVM radial) mitjançant una cerca exhaustiva en graella (grid search), avaluant cada combinació amb validació creuada de 3 particions (3-fold CV).

k-NN optimitzat

```
## k-Nearest Neighbors
##
## 513 samples
##   8 predictor
##   2 classes: 'No', 'Si'
##
## No pre-processing
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 341, 343, 342
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    1  0.6219036  0.1714695
##   11  0.7057033  0.3176741
##   21  0.7115403  0.3107445
##   31  0.7056350  0.2861984
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 21.
```



El gràfic mostra l'accuracy mitjà obtingut mitjançant validació creuada (3 fold) per diferents valors de k (veïns). Es destaca que k = 21 assoleix la màxima precisió (~75%), seguint per k = 11 i k = 31.

Els valors extrems (k = 1 i k = 31) mostren pitjor rendiment, per una banda, $k = 1$: Propens a sobreajustament (overfitting), amb alta variabilitat. Per l'altra, $k = 31$: Excessivament simple, perd patrons rellevants.

El model és robust en un rang mitjà de k (11 a 21), amb k = 21 seleccionat com a òptim.

Utilització del millor model k-NN

El model k-NN optimitzat (k=21) assolí una precisió global del 74.9%, classificant correctament 3 de cada 4 pacients. Presentà una bona sensibilitat (77.1%) per identificar casos no diabètics (valor predictiu positiu: 87.3%), però limitacions en especificitat (68.7%) per detectar diabètics (valor predictiu negatiu: 51.7%), cosa que el fa útil per descartar la malaltia però poc fiable per confirmar-la. El test de McNemar (p=0.009) revelà un desequilibri en els errors (més falsos negatius), probablement degut al desequilibri de classes (73.7% "No" vs 26.3% "Si") o a variables poc informatives per a la classe minoritària. Aquesta configuració seria adequada per a screening inicial, però requereix ajustos per millorar la detecció de casos positius. Comparant amb el model inicial k-NN l'optimització augmenta l'especificitat de ~50% a 68.7%.

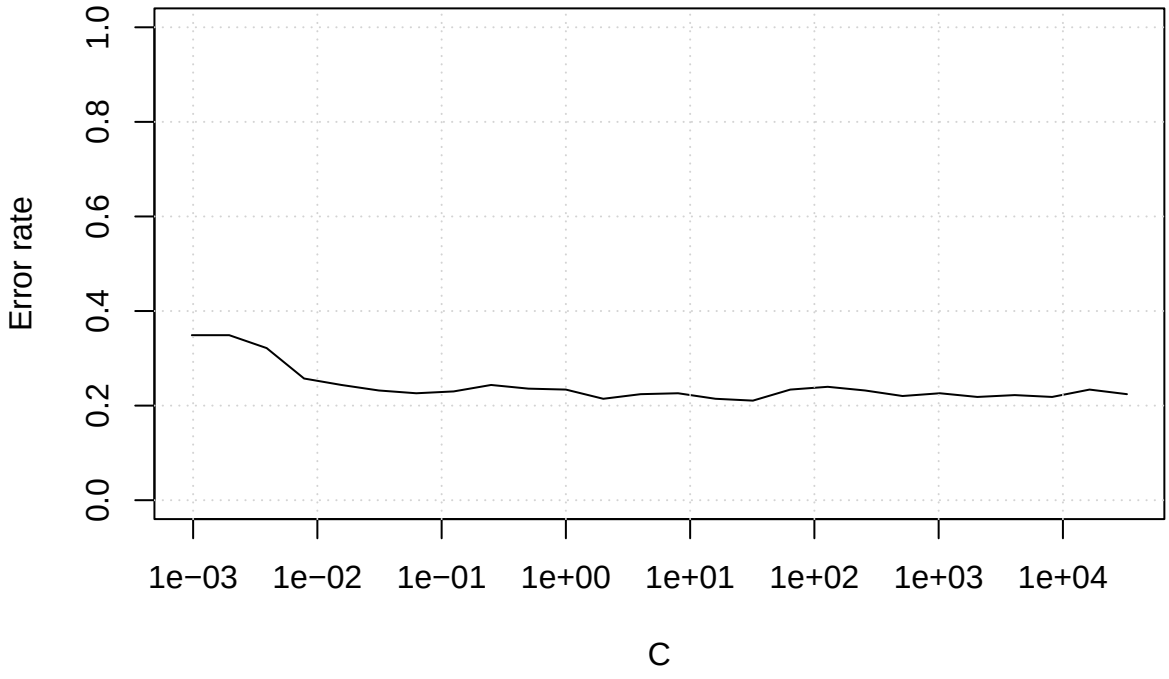
SVM

Kernel Lineal:

El procés d'optimització del paràmetre C (paràmetre de regularització) en el model SVM lineal ha avaluat valors de C en escala logarítmica (des de 2^{1} fins a 2^{1}). El gràfic d'error de classificació mostra una corba en forma de "L" típica, on l'error disminueix ràpidament a mesura que augmenta C fins a assolir un punt òptim.

```
##  Setting default kernel parameters
```

```
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
##  Setting default kernel parameters
```



Error mínim de classificació: 0.2105263 (21.64%), es considera bastant bo per a problemes mèdics Daemen et al. (2012).

Que correspon al valor òptim de C igual a 32.

El model SVM amb nucli lineal (vanilladot) i paràmetre de regularització C=8192 (determinat prèviament com a òptim) mostra els següents resultats i anàlisi d'errors:

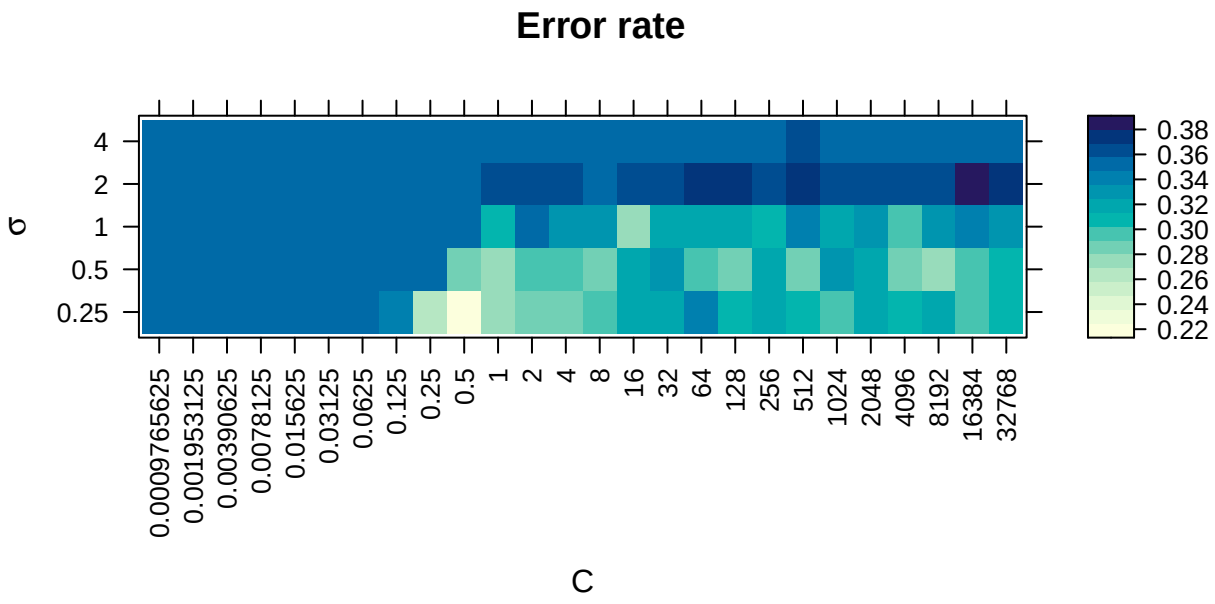
- Distribució: 37 falsos negatius (diabètics classificats com a no diabètics) vs. 23 falsos positius (p=0.093, desequilibri no significatiu).
- Prevalença: 65.1% de casos negatius (No diabètic) en la mostra.

El model és especialment útil per descartar la diabetis (alta sensibilitat) i la capacitat per confirmar la diabetis és limitada (especificitat moderada). Per l'altra banda, la raó òptima entre precisió i capacitat de generalització (C=8192) suggereix que les classes estan relativament ben separades linealment.

Kernel gaussià:

L'anàlisi d'optimització dels hiperparàmetres C (paràmetre de regularització) i (amplada del kernel) mostra que la combinació òptima per minimitzar l'error de classificació (23.2%) s'obté amb C = 0.5 i = 0.25. El gràfic d'error revela que valors baixos de C i generen models més simples però, amb menor sobreajustament, mentre que valors alts poden produir sobreajustament (error creixent). En el conjunt de prova, el model optimitzat assolí una precisió del 73.3%, amb una alta sensibilitat (86.1%) per detectar casos no diabètics, però una especificitat limitada (49.4%) per a casos diabètics. Aquest desequilibri (test de McNemar significatiu, p=0.011) suggereix que el model tendeix a classificar incorrectament més casos diabètics com a no diabètics. Tot i això, indicant que el kernel radial captura patrons no lineals rellevants, encara que amb menys eficàcia que el kernel lineal en aquest cas concret.

```
## C= 0.0009765625.....
## C= 0.001953125.....
## C= 0.00390625.....
## C= 0.0078125.....
## C= 0.015625.....
## C= 0.03125.....
## C= 0.0625.....
## C= 0.125.....
## C= 0.25.....
## C= 0.5.....
## C= 1.....
## C= 2.....
## C= 4.....
## C= 8.....
## C= 16.....
## C= 32.....
## C= 64.....
## C= 128.....
## C= 256.....
## C= 512.....
## C= 1024.....
## C= 2048.....
## C= 4096.....
## C= 8192.....
## C= 16384.....
## C= 32768.....
```



Donats aquests resultats, es destaca que el mínim error de classificació obtingut és 0.2241715.

Obtenint els valors òptims de C i sigma: 0.5 i 0.25 respectivament.

La taula següent mostra la comparativa de rendiment entre els tres models de classificació amb els seus hiperparàmetres optimitzats. S'observa que el model SVM Lineal presenta els millors resultats globals, destacant amb la màxima exactitud (76,47%) i el valor més alt d'AUC (0,8384), indicant una gran capacitat per discriminar correctament entre les dues classes.

Tot i que el model SVM Gaussià presenta una sensibilitat molt similar (86,14% vs. 86,75%) i una AUC molt correcte(0,8003), la seva especificitat és molt baixa (49,44%) cosa que el fa menys fiable en la detecció de casos diabètics. Per la seva banda, el k-NN (amb k = 21) manté un bon equilibri entre sensibilitat i especificitat, però amb un rendiment general inferior (AUC = 0,6573).

Això suggereix que, en aquest context, el SVM Lineal és el model més robust i equilibrat, oferint una bona capacitat predictiva tant per detectar com per descartar casos de diabetis.

Taula 4: Comparativa dels models amb hiperpar
optimitzats

	Métriques de Clasificació					
Model	Exactitud	Sensibilitat	Especificitat	VPP	VPN	Kappa
k-NN (k=21)	0.7490	0.7713	0.6866	0.8735	0.5169	0.4141
SVM Lineal	**0.7647**	**0.8675**	**0.5730**	**0.7912**	**0.6986**	**0.4597**
SVM Gaussià	0.7333	0.8614	0.4944	0.7606	0.6567	0.3775

Nota:

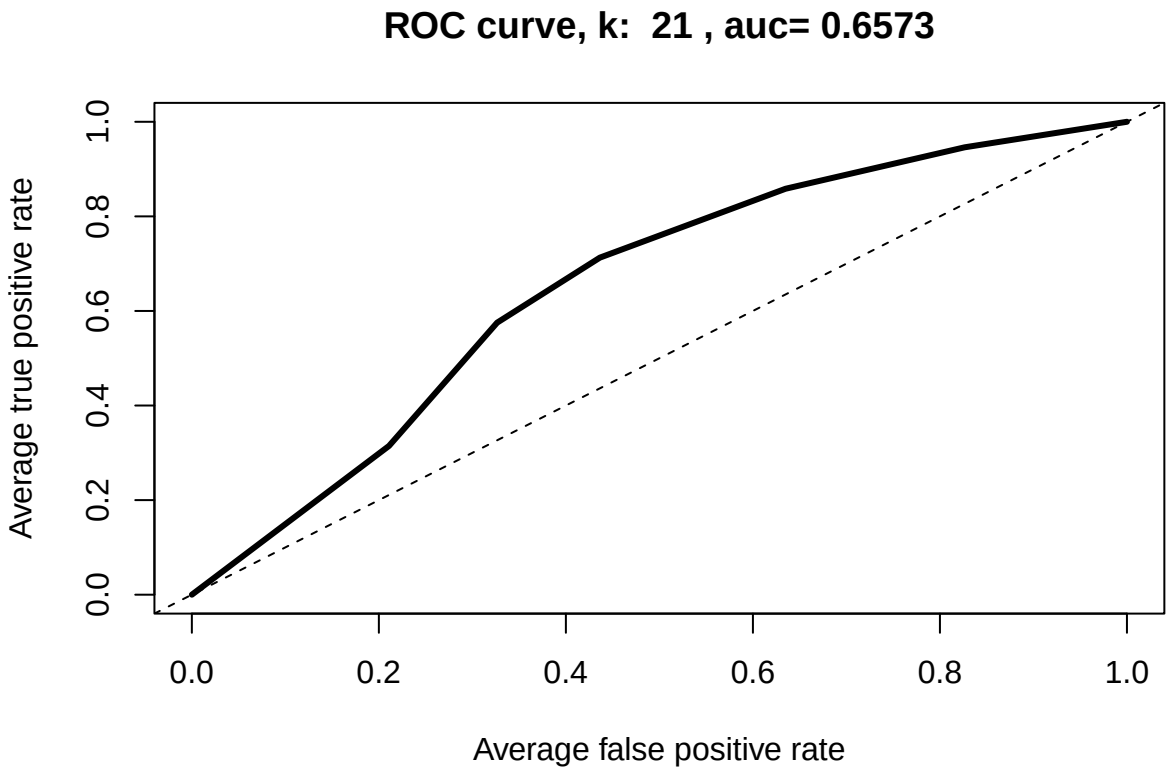
Els models han estat optimitzats mitjançant validació creuada de 3 particions (3-fold CV). k-NN (millor

Corba ROC

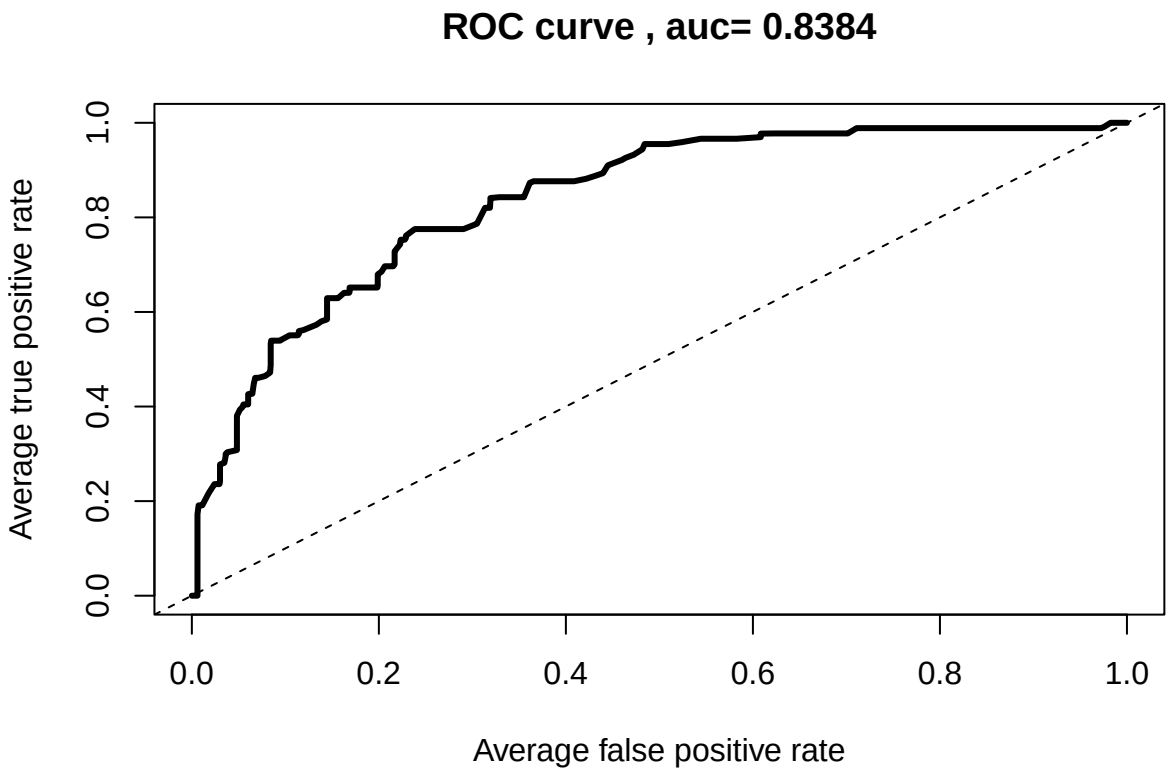
La corba ROC (Receiver Operating Characteristic) és una representació gràfica que mostra la relació entre la sensibilitat (true positive rate) i la taxa de falsos positius (false positive rate) d'un model de classificació, per diferents llindars de decisió.

Els següents gràfics mostren la corba ROC de cada un dels diferents models de decisió.

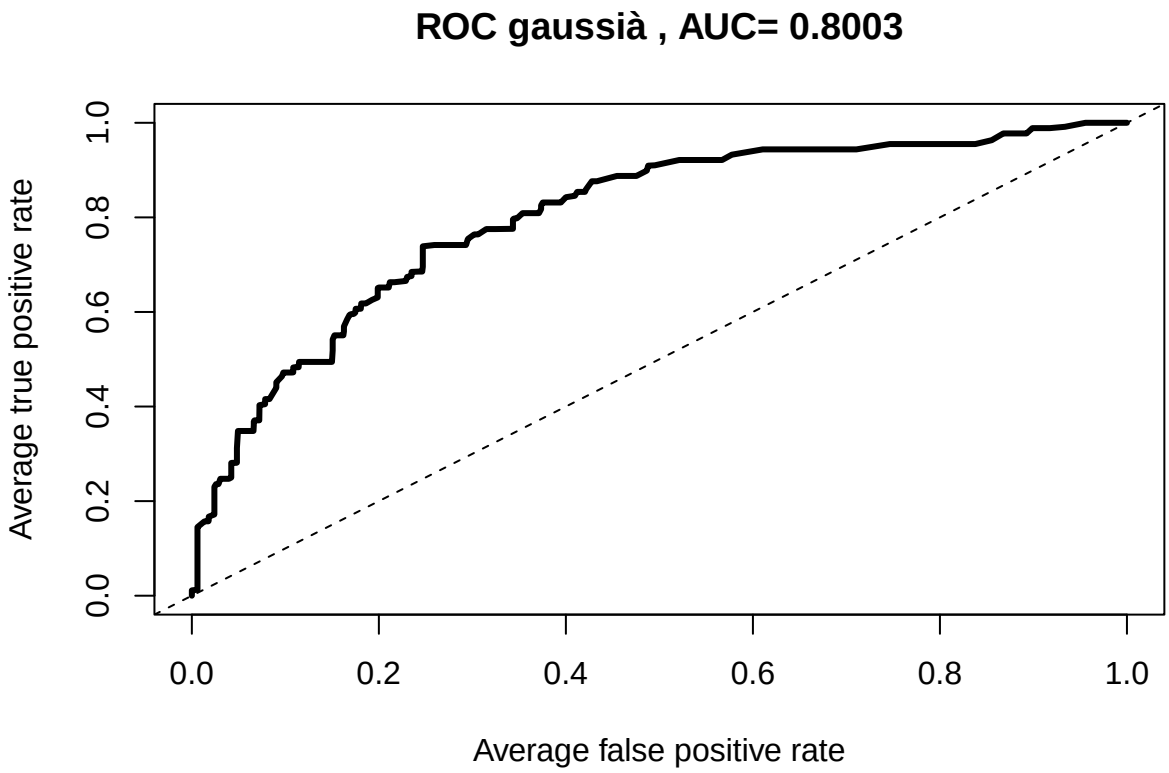
k-NN



SVM lineal

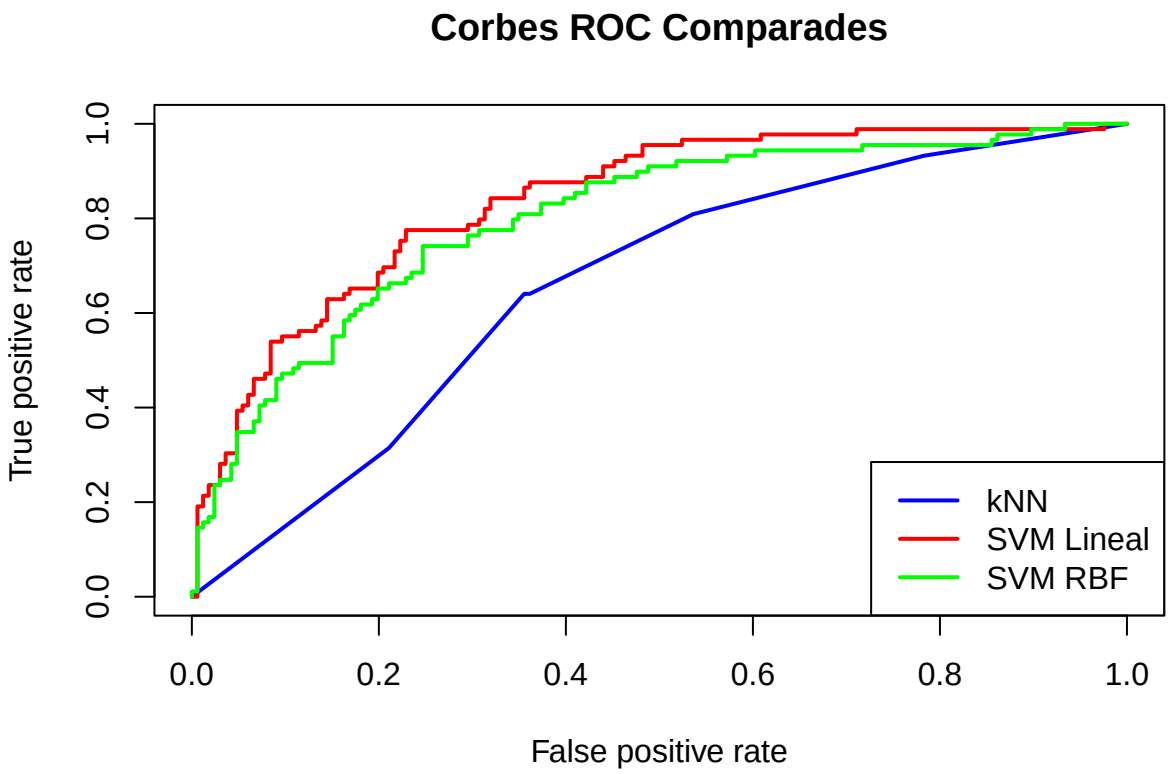

ROC curve , auc= 0.8384
Average true positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Average false positive rate
0.0
0.2
0.4
0.6
0.8
1.0

SVM RBF



En el següent gràfic es mostren les tres corbes ROC alhora. Gràficament, podem veure el que hem indicat en el comentari de la taula. El SVM lineal és el mètode més òptim. Ja que és el que queda més a prop del punt (0,1), que representa 100% de positius reals encertats i 0 % de falsos negatius.

Comparació conjunta corbes ROC



4 Discusió de resultats i conclusions

L'objectiu principal d'aquest estudi consisteix en el desenvolupament d'un mètode de classificació capaç de diagnosticar la presència de diabetis mitjançant variables mèdiques. Per a aquest fi, s'han entrenat i avaluat tres models de classificació supervisada emprant una base de dades amb informació clínica de pacients diabètics i no diabètics. Els algoritmes implementats inclouen: el mètode dels k-veïns més propers (k-NN), una màquina de vectors de suport (SVM) amb nucli lineal, i una SVM amb nucli radial basat en una funció gaussiana.

Un cop finalitzat el procés d'entrenament i optimització d'hiperparàmetres, els resultats revelen diferències significatives en el rendiment dels models. El millor model k-NN, amb $k = 21$, va assolir una exactitud del 74,9%, una sensibilitat del 77% i una especificitat del 68%. Aquestes mètriques indiquen una capacitat notable per a la detecció de pacients diabètics (alta sensibilitat), encara que amb una taxa moderada de falsos positius.

Pel que fa al model SVM amb nucli lineal, va exhibir el rendiment global més elevat, amb una exactitud del 76,5%, una sensibilitat del 86,8% i una especificitat del 57,3%. A més, l'àrea sota la corba ROC (AUC = 0,8384) va corroborar la seva robusta capacitat discriminativa. Aquests resultats suggereixen que, tot i presentar una proporció relativament elevada de falsos positius, el model és altament eficaç per a la identificació correcta de casos positius.

Quant a l'SVM amb nucli gaussià, va obtenir una exactitud lleugerament inferior (73,3%), amb una sensibilitat alta (86,1%) però una especificitat reduïda (49,4%). Aquest comportament es tradueix en una taxa elevada de falsos positius, la qual podria limitar la seva utilitat pràctica en entorns clínics reals.

En considerar el conjunt de mètriques, el model SVM amb nucli lineal destaca per oferir un equilibri òptim entre sensibilitat, exactitud i capacitat discriminativa. En el context mèdic, és prioritari maximitzar la sensibilitat per a minimitzar els falsos negatius, atès que un diagnòstic erroni en pacients diabètics pot tenir conseqüències clíniques severes. Tot i que el model k-NN presenta una especificitat lleugerament superior, l'SVM lineal resulta més adequat per a la seva implementació com a eina de suport al diagnòstic clínic de la diabetis.

Perspectives futures: En investigacions posteriors, seria rellevant explorar l'aplicació de tècniques avançades d'aprenentatge automàtic, com ara models d'assemblatge (Random Forest, Gradient Boosting) o xarxes neuronals, amb la finalitat de millorar el rendiment predictiu. Addicionalment, seria valuós analitzar les característiques dels errors de classificació per a identificar possibles patrons en els falsos positius i negatius. Una altra línia d'interès consistiria en l'ampliació de la base de dades amb mostres més extenses i diversificades, fet que podria millorar la generalització dels models. Finalment, la incorporació de variables clíniques o socioeconòmiques addicionals, juntament amb la validació dels models en entorns clínics reals, constituirien passos essencials per a avaluar-ne l'efectivitat en condicions pràctiques.

Bibliografia

Bierhaus, Angelika, Marion A Hofmann, Reinhard Ziegler, i Peter P Nawroth. 1998. «AGEs and their interaction with AGE-receptors in vascular disease and diabetes mellitus. I. The AGE concept». *Cardiovascular research* 37 (3): 586-600.

Daemen, Anneleen, Dirk Timmerman, Thierry Van den Bosch, Cecilia Bottomley, Emma Kirk, Caroline Van Holsbeke, Lil Valentin, Tom Bourne, i Bart De Moor. 2012. «Improved modeling of clinical data with kernel methods». *Artificial intelligence in medicine* 54 (2): 103-14.

Talebi Moghaddam, Maryam, Yones Jahani, Zahra Arefzadeh, Azizallah Dehghan, Mohsen Khaleghi, Mehdi Sharafi, i Ghasem Nikfar. 2024. «Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm». *BMC Medical Research Methodology* 24 (1): 220.