INFORME AMB ELS RESULTATS

Preprocessing

El primer pas és realitzar el preprocessing de la base de dades. En el nostre cas no ha sigut necessària ninguna discretització ja que totes les variables són de tipus factor. A més a més, no tenim cap valor que falta (NA) ni tampoc cap valor incorrecte.

Com que la base de dades té moltes variables amb molts nivells cadascuna, hem decidit eliminar i recodificar les següents:

Agrupacions:

- cap-color: g,n,e i a.
- stalk-root: Desconegut, b i a.
- stalk-surface-above-ring: k,s i a.
- stalk-color-above-ring: p,w i a.
- ring-type: e, p i a.
- spore-print-color: d (dark), l (light).
- population: s, v, y i a.
- habitat: d,g, p i a.

Variables eliminades:

- cap-surface: s'elimina perquè té una categoria molt minoritaria i no es indispensable per per les prediccions.
- odor: s'elimina perquè la gran majoria dels bolets comestibles no tenen olor (81%), mentre que els venenosos si que en tenen. Considerem que és una variable massa significativa a l'hora de fer les prediccions.
- gill-color: Veiem que si el color és b classifica tots els bolets com verinosos.
- gill-attachment: s'elimina perquè la gran majoria dels bolets estan dintre d'una categoria (97.4%), per tant no ens aporta informació.
- stalk-surface-below-ring: l'eliminem ja que la seva distribució és molt semblant a stalk-surface-above-ring i per tant podrien estar correlacionades
- stalk-color-below-ring: l'eliminem seguint el mateix criteri que la variable anterior.
- veil-type: s'elimina ja que només té una categoria, no ens dóna ninguna informació.
- veil-color: s'elimina perquè la gran majoria de bolets están dintre d'una categoria (97.5%).
- ring-number: s'elimina perquè la majoria estan en una sola categoria.

Quan intentavem fer l'Augmented Naive en format gRain amb tots els casos i les variables ja recodificades ens sortia un error. Per tant, vam decidir ajuntar més variables i treure'n algunes més, però continuava donant el mateix error.

També vam provar de fer una xarxa TAN (Tree Augmented Naive), la qual no ens donava cap error, però finalment vam provar de treure alguns casos i es quan ens va compilar. Vam decidir quedar-nos aproximadament amb les primeres 4000 observacions, i sí que ens va funcionar. El data set queda 3309 (82.7%) e i 691(17.3%) p.
(L'elecció de les observacions no va ser aleatoria perquè es repetia l'error)

Procés de validació de les dades

Per fer el procés de validació hem fet un k-fold cross validation agafant k=10. Aquest mètode consisteix en separar el conjunt de dades de la xarxa seleccionada en 10 subconjunts. I entrenar-los deixant un per a realitzar la validació. Finalment hi ha 10 estimacions dels paràmetres que escollim de les quals es fa la mitjana. En el nostre treball hem escollit: accuracy, recall (exhaustivitat), especificitat, precisió i per últim valor F.

Escollir model

Al comparar les mesures anteriors dels dos models, del Naive Bayes i l'Augmented Naive ens hem fixat principalment en el valor de l'accuracy, que és 0.8378 i 0.9571 respectivament. Hem tingut en comte totes les altres mesures però especialment el valor de l'especificitat ja que considerem més important encertar en els veritables verinosos que en els comestibles.

Totes les estimacions menys la precisió del model Augmented Naive són més altes que les del Naive Bayes, per tant ens quedarem amb el model augmentat.

Construcció del model final i prediccions pels casos triats

Finalment hem tornat a entrenar la xarxa de tipus Augmented Naive però amb totes les variables del dataset per a realitzar les prediccions.

Prediccions:

1. Xampinyó:

   Hem introduït les seves característiques per a veure si el nostre model predia correctament que el xampinyó es un bolet comestible. Hem indicat que el color del barret no és ni gris ni marró ni vermell. No té taques blaves. Les làmines estan molt juntes. L'arrel de la tija és bulbosa, el color de la tija sobre l'anell és rosa. No té un anell evanescent ni de penjoll, el color de les seves espores és fosc. No és un bolet amb poblacions disperses ni diverses i tampoc és solitari. I creix a l'herba.

   Una vegada introduïdes totes les evidències anteriors les probabilitats obtingudes són:

e (comestible)	p (verinós)
100%	0%

   Ens ha donat una probabilitat de ser verinós de 0%, això és degut a la precisió de les evidències. Per això hem refet les evidències i només hem seleccionat les relacionades amb el color, ara les probabilitats són:

e (comestible)	p (verinós)
82.72%	17.27%

2. Chanterelle:
   En el següent bolet les seves característiques són les següents. La forma del barret és convexa i el seu color no és ni gris ni marró ni vermell. No té taques, i té les làmines estretes i properes entre elles. L'arrel de la tija és bulbosa, la superfície d'aquesta no és ni sedosa ni suau, i el seu color no és ni rosa ni blanc. El color de les seves espores és blanc i sòl habitar en zones pastoses.

e (comestible)	p (verinós)
90.4%	9.6%

   El model ha classificat correctament aquest cas ja que és un bolet comestible.

3. Matamoscas:
   Aquest bolet és verinós i té les següents característiques. La forma del barret és convexa i de color vermell. Les làmines són estretes i són properes entre elles. L'anell no és ni pendent ni evanescent. I les espores són de colors clars.

e (comestible)	p (verinós)
82.7%	12.7%

   Podem veure que la predicció no és correcta, ja que diu que un bolet amb aquestes característiques hauria de ser comestible i no ho és.

4. Cortinarius Orellanus:
   El següent bolet té un barret de forma convexa, i de un color que ni és ni gris ni marró ni vermell i sense taques. Les làmines son estretes i properes entre si. La tija és gran i la seva arrel és bulbosa. A més la superfície de la tija per sobre de l'anell és sedosa. També té unes espores de colors clars. Aquest tipus de bolets és solitari.

e (comestible)	p (verinós)
37.8%	62.7%

   Com aquest bolet és verinós la predicció està força bé ajustada.

Un cop fetes les prediccions hem pogut veure l'eficàcia real del nostre model. El que hem observat és que prediu molt bé els bolets comestibles, amb probabilitats altes. Però no passa el mateix pels bolets verinosos. Això és degut al conjunt final de dades, ja que més del 80% són comestibles.