Informe trabajo final de Programación:

Para realizar el trabajo final de programación, nuestro grupo se decidió por determinar si un hongo es comestible o venenoso dependiendo de ciertas características principales de la estructura y morfología de este, usando como herramienta el uso de dos métodos de clasificación denominados: regresión logística y Random Forest. Para poder obtener el resultado deseado comenzamos por obtener el archivo original que contenía la información de los hongos el cual se obtuvo del repositorio UCI, id=73, este dataset contenía la información de 8124 instancias y 22 atributos lo que cumplía con los requisitos iniciales para poder proceder con el trabajo. Una vez cargado el dataset se realizaron diversos procesos como la exploración de estadísticas básicas y datos faltantes, se descartaron columnas innecesarias además de la visualización de la proporción de las clases del dataset mediante un gráfico de barras esto para poder comprender los datos que se encontraban en este, dándonos cuenta de que el conjunto de datos tenia un buen balance entre setas venenosas y comestibles.

Luego del paso anterior se procedió a ordenar el conjunto de datos y se simulo el 5% de datos faltantes todos de forma aleatoria. Posteriormente se selecciono el 80% de las primeras filas para realizar el análisis, se aplicaron procesos de codificación de etiquetas a las variables categóricas y estandarización a las características. Luego se generaron dos archivos csv, el primero de estos contenía las variables independientes y el otro contenía la variable objetivo.

Para la repartición de datos de entrenamiento y prueba se realizó una repartición del 80% y 20% respetivamente. Para el primer modelo el cual era Random Forest procedimos a explorar distintas combinaciones de número de árboles, profundidad máxima y criterios mínimos de división y hoja. l mejor modelo alcanzó una precisión de validación cruzada cercana al 99 % y obtuvo un 99.2 % de accuracy en el conjunto de prueba, con un AUC de 0.998. Con ayuda de las curvas de aprendizaje se pudo obtener que tanto el desempeño en entrenamiento y validación se mantuvo alto, además se pudo observar que las características más decisivas fueron el olor, color de esporas y tamaño de branquias.

Para el segundo modelo, regresión logística se ajustó el parámetro de regularización C y usamos penalización L2, con esto obtuvimos alrededor de 95.4 % en validación cruzada y un 95.0 % de accuracy en prueba, con un AUC de 0.96. Las curvas de aprendizaje, aunque muy efectivas nos mostraron que este modelo tenia una pequeña tendencia a subajustarse con pocos datos y al igual que el modelo anterior

las características mas importantes fueron el olor, color de esporas y forma de branquias.

Al obtener los resultados finales de los dos modelos podemos observar que Random forest resulto un poco mejor en aspectos como precisión global o estabilidad, mientras que por el otro lado el modelo de regresión logística permito mayor interpretabilidad de sus coeficientes. Para posteriores trabajos se encuentra bastante importante encontrar métodos mas precisos de imputación en casos reales de datos faltantes además de incluir diferentes variables diferentes a las actuales como información del hábitat o aspectos bioquímicos, con el fin de enriquecer la capacidad predictiva.