سوال ۱-

الف)

ابتدا آنتروپی را در گره والد محاسبه می‌کنیم:

$$I(parent) = -\sum_{j} P(j|t) log_2 P(j|t) = 0.4\, log_2(0.4) + 0.6\, log_2(0.6) \cong 0.97$$

حال هنگام جداسازی هرکدام از ویژگی‌ها بهره اطلاعاتی را حساب می‌کنیم:

$$I(v_{A=T}) = -\left(\frac{4}{7} log_2\left(\frac{4}{7}\right) + \frac{3}{7} log_2\left(\frac{3}{7}\right)\right) \cong 0.985$$

$$I(v_{A=F}) = -\left(\frac{0}{3} log_2\left(\frac{0}{3}\right) + \frac{3}{3} log_2\left(\frac{3}{3}\right)\right) \cong 0$$

$$\Delta_A = I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j) = 0.97 - \left(\frac{7}{10} \times 0.985 + \frac{3}{10} \times 0\right) \cong 0.28$$

$$I(v_{B=T}) = -\left(\frac{3}{4} log_2\left(\frac{3}{4}\right) + \frac{1}{4} log_2\left(\frac{1}{4}\right)\right) \cong 0.81$$

$$I(v_{B=F}) = -\left(\frac{1}{6} log_2\left(\frac{1}{6}\right) + \frac{5}{6} log_2\left(\frac{5}{6}\right)\right) \cong 0.65$$

$$\Delta_B = I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j) = 0.97 - \left(\frac{4}{10} \times 0.81 + \frac{6}{10} \times 0.65\right) \cong 0.256$$

چون بهره اطلاعاتی با جداسازی بر اساس A بیشتر است، بنابراین همین ویژگی باید انتخاب شود.

ب)

$$GINI_A = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) = \frac{7}{10}\left(1 - \left(\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right)\right) + \frac{3}{10}\left(1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right)\right) \cong 0.34$$

$$GINI_B = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) = \frac{4}{10}\left(1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right)\right) + \frac{6}{10}\left(1 - \left(\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right)\right) \cong 0.317$$

ویژگی B باید انتخاب شود زیرا GINI Index آن کوچکتر است.

ج)

بله، ممکن است ویژگی‌های مختلفی را ترجیح دهند. زیرا آنتروپی شیب بیشتری نسبت به GINI Index دارد و مقدار بیشتری برای مقادیر نزدیک 0.5 بدست می‌آورد. بنابراین معیار بهره اطلاعاتی بیشتر تمایل دارد که بعضی از گره‌های فرزند تقریبا خالص باشند. اما معیار GINI Index نگاه کلی‌تری دارد و می‌خواهد که همه فرزندان در مجموع خالص‌تر از قبل شده باشند.

سوال ۲-

الف)

$$P(A|+) = \frac{P(A,+)}{P(+)} = \frac{\frac{3}{10}}{\frac{5}{10}} = 0.6$$

$$P(B|+) = \frac{P(B,+)}{P(+)} = \frac{\frac{1}{10}}{\frac{5}{10}} = 0.2$$

$$P(C|+) = \frac{P(C,+)}{P(+)} = \frac{\frac{4}{10}}{\frac{5}{10}} = 0.8$$

$$P(A|-) = \frac{P(A,-)}{P(-)} = \frac{\frac{2}{10}}{\frac{5}{10}} = 0.4$$

$$P(B|-) = \frac{P(B,-)}{P(-)} = \frac{\frac{2}{10}}{\frac{5}{10}} = 0.4$$

$$P(C|-) = \frac{P(C,-)}{P(-)} = \frac{\frac{5}{10}}{\frac{5}{10}} = 1$$

ب)

$$P(A=0,B=1,C=0|+) = P(A=0|+)P(B=1|+)P(C=0|+) = 0.4 \times 0.2 \times 0.2 = 0.016$$

$$P(A=0,B=1,C=0|-) = P(A=0|-)P(B=1|-)P(C=0|-) = 0.6 \times 0.4 \times 0 = 0$$

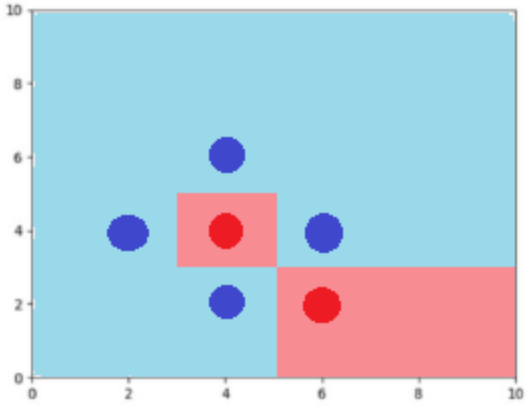
بنابراین برچسب این داده + می‌باشد.

سوال ۳-

زمانی که Error rate زیاد باشد، Accuracy معیار خوبی برای سنجش classifier نخواهد بود. به همین دلیل بهتر است از Confusion Matrix استفاده کنیم تا هم Accuracy و هم Error rate را مدنظر قرار دهیم.

سوال ۴-

الف)



همانطور که دیده می‌شود، به ازای هر داده نمونه باید کلاس نزدیک‌ترین همسایه را به آن نسبت دهیم که باعث می‌شود مرزهای تصمیم‌گیری بالا تشکیل شوند.

ب)

به کلاس آبی نسبت می‌دهیم زیرا فاصله اقلیدسی (8, 8) با نزدیک‌ترین نقطه آبی برابر با $2\sqrt{5} \cong 4.47$ و فاصله اقلیدسی آن با نزدیک‌ترین تقطه قرمز برابر با $4\sqrt{2} \cong 5.66$ می‌باشد.

ج)

بله میتوان، کافی است برای هر ویژگی یک بعد تعریف کنیم و فاصله نقاط را در فضای n=بعدی اندازه بگیریم. حال برای پیشبینی مقدار مجهول کافی است از میانگین وزن‌دار k نزدیک‌ترین همسایه استفاده کنیم.

د)

بستگی به داده ورودی دارد. در حالت کلی اگر مقدار k خیلی کوچک باشد، الگوریتم به داده‌های نویز حساس می‌شود و اگر k خیلی بزرگ باشد، ممکن است همسایه‌ها از کلاس‌های دیگر انتخاب شوند.

ه)

الگوریتم KNN با افزایش تعداد ابعاد می‌تواند دچار مشکل شود. زیرا داده‌هایی که به هم نزدیک هستند ممکن است فاصله زیادی از یکدیگر داشته باشند. به این مشکل curse of dimensionality می‌گویند.

از طرف دیگر، این الگوریتم به اصطلاح lazy learner است و نیازی نداریم که یک مدل از کل داده‌ها بسازیم (که ممکن است پیچیدگی زمانی زیادی داشته باشد)، بلکه پیشبینی با استفاده از خود نمونه‌های خام انجام می‌شود.

و)

در حالت آموزش کار خاصی را انجام نمی‌دهیم، به همین دلیل پیچیدگی زمانی نداریم.

در حالت آزمایش باید فاصله نمونه آزمایش با تمام نمونه‌های آموزش اندازه گرفته شود، سپس k نزدیک‌ترین انتخاب شوند. اگر n ویژگی و m داده آموزش داشته باشیم، آنگاه این عملیات O(mn+km) به طول خواهد انجامید.

ی)

در معیار منهتن فاصله تمام ویژگی‌ها را با هم جمع می‌کنیم اما در معیار اقلیدسی جذر مجموع مربعات فاصله ویژگی‌ها را بدست می‌آوریم یا همان نرم ۲.

سوال ۵-

Cross-Validation یک روش آماری است که با استفاده از آن دقت یک مدل آمورشی سنجیده می‌شود. از این روش هنگامی استفاده می‌شود که داده‌های نمونه ما محدود باشند. سه دسته کلی آن عبارتند از:

- Exhaustive cross-validation: در این روش تمام حالات ممکن تقسیم‌بندی داده‌های نمونه انجام می‌شود.
- Non-exhaustive cross-validation: در این روش تنها بخشی از حالات تقسیم‌بندی انجام می‌شود.
- Nested cross-validation: در این روش تقسیم‌بندی‌های تو در تو نیز انجام می‌شود.

اگر در روش K-fold cross-validation مقدار k افزایش یابد، در واقع اندازه داده‌های آموزش بیشتر شده و تعداد داده‌های تست کم می‌شود. بنابراین مقدار بایاس به دلیل داده آموزشی بیشتر کاهش یافته و مقدار واریانس نیز افزایش می‌یابد. همچنین پیچیدگی زمانی نیز افزایش خواهد یافت زیرا باید k مرتبه فرآیند cross-validation را تکرار کنیم.