Tarea 8
Nombre: Sánchez Sauñe Cristhian Wiki

**1. En el ejemplo Party Registration, ¿cuáles son las características?. Cual es la respuesta? ¿Es este un problema de regresión o clasificación?**

Hay  tres propiedades: registro de partidos electorales (intención de voto), riqueza de votantes y una medida cuantitativa de la religiosidad electoral. Las intensiones de voto son graficados con círculos, los círculos rojos representan a los votantes republicanos, mientras que los círculos azules representan los votos demócratas.
Se quiere predecir el registro de votantes utilizando la riqueza y la religiosidad como predictores. Es un problema de clasificación pues las variables target son discretas.

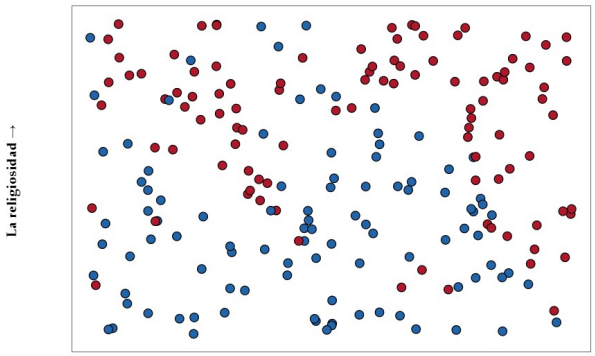
**2. Conceptualmente, ¿cómo se está aplicando KNN a este problema para hacer una predicción?**

El enfoque de KNN se usa porque los datos no presentan linealidad para poder atacarlo con una regresión logística.

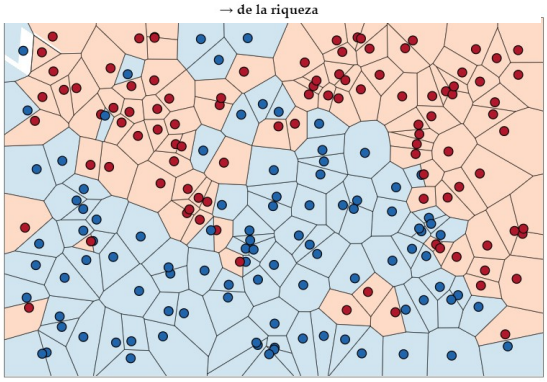
En KNN, la intención de voto de un elector dado se encontrará graficando y comparando la intención de voto de los otros votantes. Los k vecinos más cercanos a él se encontrarán utilizando una medida euclidiana de distancia, y el promedio de las intenciones de voto se utilizará para predecir la intención de voto para éste votante. Así que si el votante más cercano a él (en términos de riqueza y religiosidad) es un demócrata, también se pronosticará que es demócrata (lo mismo aplica para el republicano).

**3. ¿Cómo se relacionan entre sí las cuatro visualizaciones de la sección 3? Cambia el valor de K con el control deslizante y asegúrate de comprender qué cambió en las visualizaciones (y por qué cambió). Explica tu respuesta.**

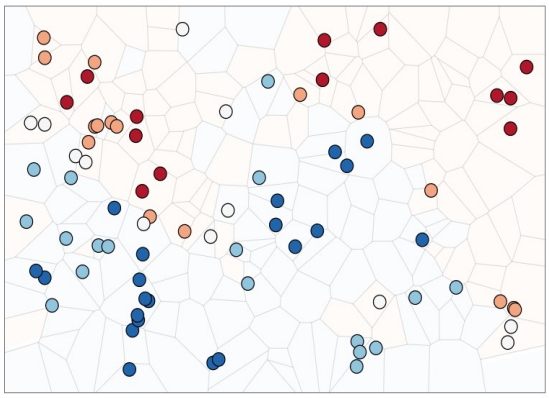
**Datos de entrenamiento:**



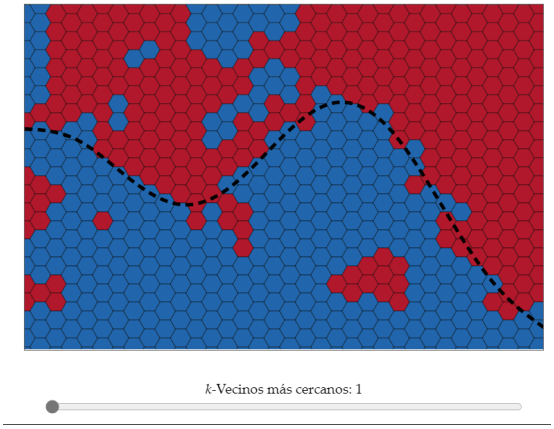

**Entrenamiento:**



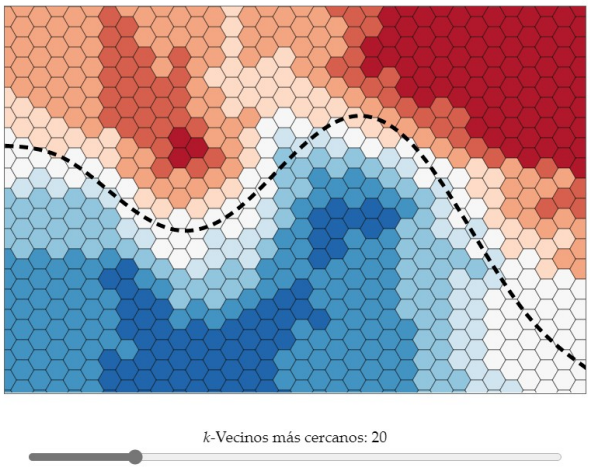
**Predicción (puntos no visto):**



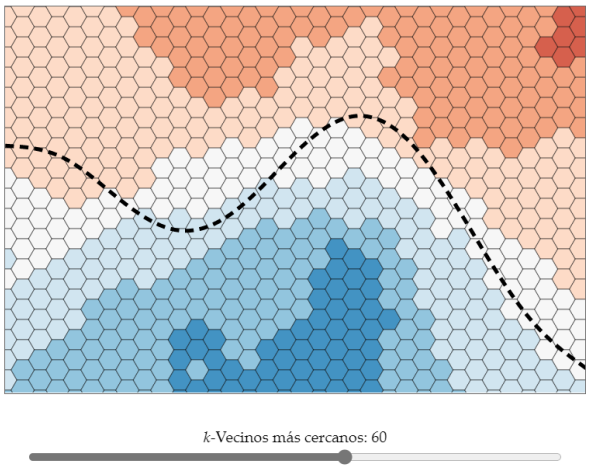
Para valores de K=1, se observa un sobreajuste, se ajusta perfectamente a cada punto ( el modelo tiene alta certidumbre), pero no logra generalizar. Usa solo 1 vecino cercano, por lo que no tendrá muchas variables con las que comparar.




Un valor de k=20, podemos observar un modelo equilibrado (no hay overfitting ni underfitting) y que podría generalizar a nuevos datos no vistos.



Un valor alto de k=60, simplemente nos da un modelo con una alta incertidumbre (alto sesgo, underfitting). Es todo lo opuesto al valor de k=1. Usa demasiados vecinos para poder obtener una etiqueta. El limite de decisión es muy difuso.



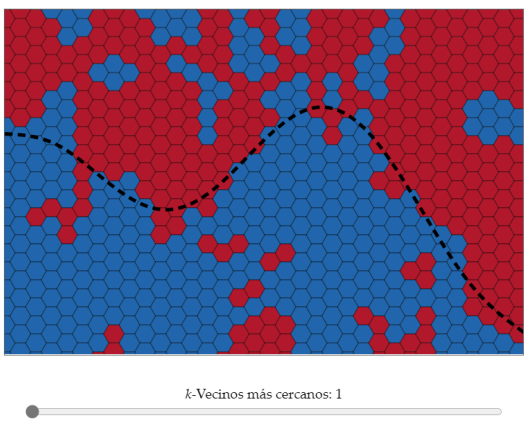

**4. En las figuras 4 y 5, ¿qué significan los colores más claros frente a los colores más oscuros? ¿Cómo se calcula esa oscuridad?.**

Un color blanco refleja una baja certidumbre del modelo (no esta seguro a que clase pertenece un dato), mientras que un color más oscuro una alta certidumbre (el modelo esta muy seguro de su clasificación.

**5. ¿Qué representa la línea negra en la figura 5? ¿Qué predicciones haría el mejor modelo posible de aprendizaje automático con respecto a esta línea?.**

La línea negra representa el límite de decisión ideal. Escogiendo valores de k, debemos aproximar nuestro modelo a esta línea. El mejor modelo de machine learning es el que se aproxima mejor a esa línea sin perder generalización en cuanto a sus nuevas predicciones.

**6. Escoge un valor muy pequeño de K y haz clic en el botón Generate New Training Data varias veces. ¿Observas una varianza baja o alta, y un sesgo bajo o alto?.**



Se observa una alta varianza, pues el modelo con k=1 se sobreajustó mucho al conjunto de entrenamiento y no consigue generalizar el aprendizaje. Por ello es que observamos islas azules en medio de rojas, y viceversa. En otras palabras hay una alta varianza.

**7. Repite esto con un valor muy grande de K. ¿Observas una varianza baja o alta y un sesgo bajo o alto?**

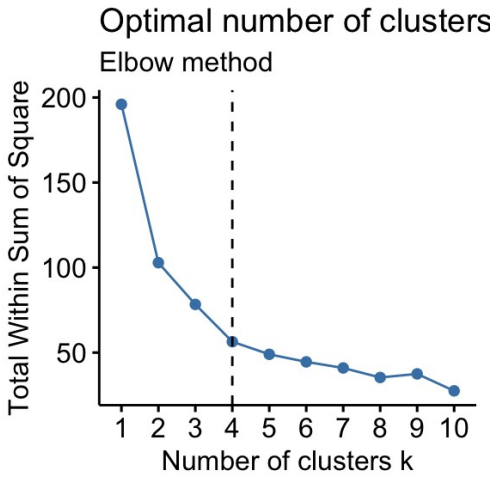
Como se ve en la última figura de la pregunta 3, un valor alto k=60 no da un modelo sustancialmente pobre y lleno de incertidumbre, en otras palabras, tiene un alto sesgo o bias.

**8. Intenta usar otros valores de K. ¿Qué valor de K cree que es mejor?. ¿Cómo defines mejor?.**

Como se ve en la penúltima figura de la pregunta 3, un valor de k=20 logra buenos resultados, tanto en el conjunto de entrenamiento, como generalizando el aprendizaje en nuevos datos no vistos. Definimos 'mejor' como la capacidad del modelo, de generalizar el conocimiento aprendido durante el entrenamiento, a datos no vistos. El mejor modelo tendrá bajo sesgo y baja varianza.

**9. ¿Un valor pequeño de K causa sobreajuste o subajuste?.**

Un valor pequeño causa sobreajuste. Para la mayoría de problemas de clasificación se suele usar valores de k>6. Además para atacar este problema de sobreajuste existen muchas técnicas entre las que se encuentra 'Elbow Method' o el método del codo, que se usa para encontrar un valor para k que de buenos resultados en el conjunto de entrenamiento y validación.



**10. ¿Por qué deberíamos preocuparnos por la varianza? ¿No deberíamos simplemente minimizar el sesgo e ignorar la varianza?.**

Un buen científico de datos debe preocuparse tanto por la varianza como por el sesgo, debido a que dependiendo del problema que estemos tratando (salud, finanzas, etc), usar un valor pequeño de sesgo solo como buen indicador dice poco de otro tipos de problemas con el que nos podemos encontrar, tales como el supuesto de que tenemos infinitos datos que compensan esa varianza, cuando en la realidad muchos problemas tiene escasez de los mismos, además que tienen mucho ruido muestral.

Dejar del lado la varianza hace que nuestro modelo pierda generalización y robustez a datos reales.

Varianza y Sesgo tienen un trace off que nos obliga a preocuparnos por ambos.