# Trabalho Prático

Aprendizagem Computacional 1

202107858 Alexandre Carneiro
202106772 Rui Coelho
202107951 Sérgio Coelho

## Objetivos

- Compreender o funcionamento de um algoritmo e verificar quais são as características que mais afetam o seu desempenho.
- Avaliar o comportamento do algoritmo selecionado, documentando a sua análise e os seus resultados.
- Implementar modificações no algoritmo, de modo a torná-lo mais robusto e consistente.

## Abordagem

Verificamos para cada variação, no cálculo da distância, os valores ímpares de k num intervalo de 3 a 24, de modo a maximizar a sua precisão.

## Resumo dos Resultados

Os diferentes métodos para calcular distâncias para o KNN podem impactar significativamente o desempenho do algoritmo.

# Algoritmo KNN

O **algoritmo KNN** é uma técnica de Machine Learning usado para problemas de classificação. O KNN funciona baseado na ideia de que dados similares estão próximos uns dos outros no espaço de caracteristícas.

O primeiro passo é **escolher o valor de K,** que representa o número de vizinhos mais próximos que serão considerados na classificação.

Cálculo da Distância. Para cada ponto, o **KNN calcula a distância** entre este ponto e todos do conjunto.

Após calcular as distâncias, o algoritmo identifica os **K pontos mais próximos** do ponto de teste.

Classifica-se o novo dado com a classe mais contabilizada.

# Comportamento do KNN relativamente às características dos dados

- Tem um **bom desempenho** quando as classes do conjunto de dados têm regiões distintas e separadas no espaço de características
- Produz **bons resultados** quando o conjunto de dados não contém outliers
- Desequilíbrio nas classes, **influenciam negativamente** o algoritmo
- O seu **desempenho tende a piorar** com conjuntos de dados maiores, uma vez que a noção de vizinhança torna-se menos significativa

# Proposta

O algoritmo KNN (K-Nearest Neighbors) é amplamente utilizado devido à sua eficácia em diversos problemas de classificação.

- A **nossa motivação** parte do princípio que é possível melhorar a precisão do KNN em diferentes datasets e queremos entender o impacto das métricas de distância na eficácia do algoritmo.

- O **nosso objetivo** é explorar diferentes formas de calcular distâncias e avaliar como isso afeta a precisão dos resultados do KNN, em diferentes datasets.

# Distâncias

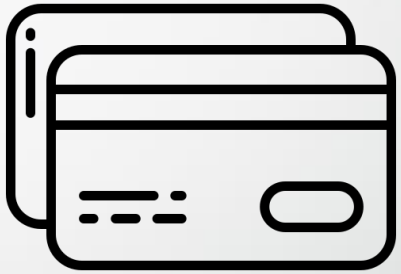
- **Euclidiana** $\sqrt{\sum_{i=0}^{n}(p_i - q_i)^2}$
  - Calculada como a raiz quadrada da soma dos quadrados das diferenças entre os componentes dos vetores
  - É frequentemente usada para dados numéricos
- **Manhattan** $P_1(x_1, y_2), P_2(x_2, y_2) \rightarrow D_{P_1P_2} = |x_1 - x_2| + |y_1 - y_2|$
  - É determinada pela **soma das diferenças absolutas entre os componentes dos vetores**
  - É uma alternativa à distância euclidiana para dados numéricos contínuos
- **Minkowski** $D(X, Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$
  - Pode ser considerada uma **generalização da distância de Manhattan e a distância Euclidiana**
- **Jaccard** $J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$
  - É como uma medida de similaridade entre conjuntos e calculada como **a diferença entre a interseção e a união de conjuntos**
  - É adequada quando os dados são categóricos binários, de vários valores ou uma combinação de ambos

- **Chebyshev** $\max(|x_1 - x_2|, |y_1 - y_2|)$
  - Calculada a partir da **maior diferença absoluta entre os componentes dos vetore**s.
  - Frequentemente utilizada para dados numéricos contínuos e quando os atributos têm escalas distinta
- **Canberra** $\sum_{i=0}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}, onde: \ P(p_1, p_2, \ldots, p_n), Q(q_1, q_2, \ldots, q_n)$
  - Calcula a **distancia através de um vetor n-dimensional**
  - É mais sensível a diferenças relativas entre as características dos pontos.
- **Cosine** $\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}}$
  - Usada para calcular a **diferença entre dois vetores de atributos**.
  - A similaridade do cosseno fornece um valor no intervalo [-1,1], porém é comum ser usada em contextos onde todos os valores são positivos, fornecendo então um valor entre [0,1].
- **Variação de Distância** $\delta(P, Q) = sup_{A \in F} |P(A) - Q(A)|$
  - Também conhecida como distância estatística, é uma medida para a probabilidade de distribuições.

# Conjuntos de dados e as suas características

Dataset nº1: **Riscos de Crédito**

- Conjunto de dados que avalia o risco de crédito
- Conjunto dividido entre valores categóricos e numéricos
- 1000 entradas x 21 features



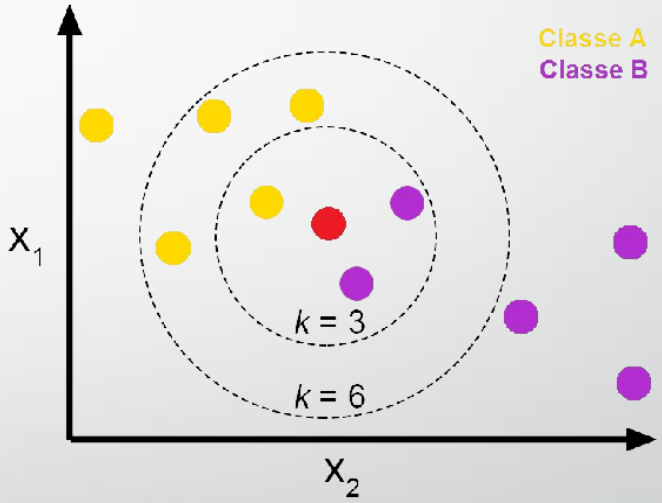
Dataset nº2: **Mobilidade Social**

- Conjunto de dados que relaciona a ocupação dos pais com a ocupação dos filhos
- Conjunto dividido entre valores categóricos e numéricos
- 1160 entradas x 6 features

# Hiperparâmetros

- No decorrer do desenvolvimento do algoritmo, utilizamos o **hiperparâmetro K,** que representa o número de vizinhos considerados na classificação.

- Tivemos o cuidado de usar um intervalo de valores entre 3 a 24, com incrementação de 2, de forma a evitar empates. Desta forma garantimos mais robustez e precisão do algoritmo.



# Método de Estimativa de Desempenho

- Para medir e analisar a performance das diferentes formas de calcular as distâncias no algoritmo KNN, utilizamos a função "accuracy score" da biblioteca "sklearn".
- Esta função mede a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões realizadas. Calcula a fração de previsões que estão corretas.

# Conclusão

Após implementarmos o algoritmo KNN de acordo com as alterações que pretendíamos fazer, comprovamos que:

- ✓ **A escolha da distância pode melhorar ligeiramente a accuracy,** com a distância de Minkowski mostrando-se particularmente eficaz.
- ✓ **O KNN enfrenta dificuldades no tratamento de datasets** com muitas features, resultando em baixa accuracy devido à "Maldição da Dimensionalidade".
- ✓ Embora a accuracy tenda a aumentar com valores maiores de k, estes podem ser enganosos, sendo recomendável usar valores mais baixos para maior confiabilidade.

# QUESTÕES