# Lista1 de AM, 2025.2

Prof. Evandro Costa

1) Considere a base de dados seguinte, supostamente fornecida pelo "gerente do banco", realizando nela a seguinte ampliação:
   (i) aumenta-la para que contenha 6 atributos e 30 exemplos (E15, E16, …, E30), com a adição de 16 exemplos, distribuídos entre Risco = Baixo Risco = Alto e Risco = Moderado,
   (ii) A partir da base de dados ampliada (conforme feito em (i), construa "**manualmente" e apresente**, 3 árvores de decisão, sendo uma usando o algoritmo ID3, outra com o C4.5 e uma última com o algoritmo CART.
   (iii) De cada uma das 3 árvores de decisão obtidas em (ii), extraia e apresente as respectivas bases de conhecimento com regras do tipo SE...ENTÃO…
   (iv) Compare as 3 bases de regras e selecione, justificando sua escolha, a que você julga mais apropriada para ser a base de conhecimento a ser usada na solução do problema de análise de risco de crédito, discutido em sala de aula.

- **Ex.: Conjunto de Exemplos de Avaliação de Risco de crédito:**
- **4 atributos e 3 classes** (Baseado no livro do Luger: IA)

| | Historia de Crédito | Dívida | Garantia | Renda | Risco |
|---|---|---|---|---|---|
| E1 | Ruim | Alta | Nenhuma | $0 a $15k | Alto |
| E2 | Desconhecida | Alta | Nenhuma | $15 a $35k | Alto |
| E3 | Desconhecida | Baixa | Nenhuma | $15 a $35k | Moderado |
| E4 | Desconhecida | Baixa | Nenhuma | $0 a $15k | Alto |
| E5 | Desconhecida | Baixa | Nenhuma | Acima de $35k | Baixo |
| E6 | Desconhecida | Baixa | Adequada | Acima de $35k | Baixo |
| E7 | Ruim | Baixa | Nenhuma | $0 a $15k | Alto |
| E8 | Ruim | Baixa | Adequada | Acima de $35k | Moderado |
| E9 | Boa | Baixa | Nenhuma | Acima de $35k | Baixo |
| E10 | Boa | Alta | Adequada | Acima de $35k | Baixo |
| E11 | Boa | Alta | Nenhuma | $0 a $15k | Alto |
| E12 | Boa | Alta | Nenhuma | $15 a $35k | Moderado |
| E13 | Boa | Alta | Nenhuma | Acima de $35k | baixo |
| E14 | Ruim | Alta | Nenhuma | $15 a $35k | Alto |

2) Tal como feito em 1), construa e apresente as **árvores de decisões** geradas e suas **bases de regras** correspondentes, obtidas a partir do uso dos algoritmos ID3, C4.5 e CART, só que agora usando as implementações deles disponíveis em alguma biblioteca de "Machine Learning" (ex.: Scikit Learn ou R ou Weka, …) ou se preferir, implemente os algoritmos e mostre o processo de construção e o resultado.

3) Considerando uma das 3 base de dados menionadas a seguir, que você vai localizar no Kaggle, sobre os domínios de aplicação:

   (i) Saúde (diabetes) ou
   (ii) Educação (Predict students' dropout and academic success) ou
   (iii) Gestão de Recursos Humanos (IBM HR Analytics Employee Attrition & Performance, sobre predição de employee turnover).

   Gere e apresente uma árvore de decisão e as regras correspondentes. Quando gerar a árvore, realize uma avaliação de desempenho usando a métrica de acurácia e, caso queira, veja e apresente outras métricas (apresente a matriz de confusão).

   Obs.: Alternativamente você poderá escolher uma outra base de dados, inclusive de outro repositório.

   (iv) Na base de dados escolhida em 3), rode um algoritmo que gera regras diretamente, por exemplo o algoritmo PRISM ou o Ripper. Apresente a base de dados com as regras geradas, também mostrando métricas de desempenho.

4) Descreve o que é overfitting e o que é underfitting, ilustrando sua descrição com a ajuda de exemplos usando conjuntos de treinamento, um para cada comportamento em modelos preditivos de árvore de decisão. Além disso, mostre como o algoritmo C4.5 contribui para mitigar os efeitos desses 2 comportamentos indesejáveis.

5) Sobre o algoritmo kNN, mostre e discuta como os comportamentos overfitting e underfiting podem ocorrer e quais as providências que podem ser tomadas para evitá-los ou pelo menos mitigar os seus efeitos.

6) Sobre o algorimo kNN responda às seguintes questões relacionadas ao seu funcionamento:
   a) Discuta sobre cuidados com os possíveis efeitos de overfitting e underfitting relacionados a determinadas escolhas do valor de k, considerando um determinado conjunto de treinamento. Explique e mostre com exemplo que se verifica a seguinte afirmação: normalmente k pequeno favorece o risco de overfitting, enquanto k grande favorece o risco de underfitting
   b) Descreva e explique com exemplos algumas boas estratégias que orientam a escolha de um valor adequado para k em um determinado conjuntos de dados de treinamento,
   c) Explique, mostrando um exemplo, como você poderia fazer uso de validação cruzada para escolher k, discutindo se essa estratégia se mostra efetiva,
   d) Sobre medidas de distância, apresente um exemplo de problema e conjunto de dados associado, onde a distância euclidiana não se mostra adequada e qual seria uma medida substitura para ela no problema que você indicou.
   e) Discuta situações problema, dando exemplos, nas quais o algoritmo kNN se mostra inefetivo, dando baixo desempenho.

Obs.: A pontuação das questões são: 1), 2) e 3) vale cada uma de 0 a 2 pontos, enquanto a 4 e 5 cada uma vale de 0 a 1 pontos. Já a 6, vale de 0 a 2 pontos.