Proposta de Projeto – Aprendizagem Automática I

1. **Membros do Grupo**
    a. Luís Freitas PG38347
    b. Luís Maia A84241
2. **Descrição do Problema - *Riiid! Answer Correctness Prediction***

"Em 2018, 260 milhões de crianças não frequentavam a escola. Ao mesmo tempo, mais da metade desses jovens alunos não atendia aos padrões mínimos de leitura e matemática. A educação já estava em uma situação difícil quando o COVID-19 forçou a maioria dos países a fechar temporariamente as escolas. Isso atrasou ainda mais as oportunidades de aprendizagem e o desenvolvimento intelectual. As lacunas de equidade em todos os países podem aumentar ainda mais. Precisamos repensar o sistema educacional atual em termos de atendimento, envolvimento e atenção individualizada.
O Riiid Labs, um provedor de soluções de IA que oferece uma inovação criativa para o mercado educacional, capacita os participantes da educação global a repensar as formas tradicionais de aprendizagem aproveitando a IA. Com uma forte crença na igualdade de oportunidades na educação, Riiid lançou um tutor de IA baseado em algoritmos de aprendizagem profunda em 2017 que atraiu mais de um milhão de estudantes sul-coreanos. Este ano, a empresa lançou o EdNet, o maior banco de dados aberto do mundo para educação em IA, contendo mais de 100 milhões de interações de alunos."
(https://www.kaggle.com/c/riiid-test-answer-prediction/overview)

O objetivo é criar algoritmos para "Rastreamento de Conhecimento", a modelagem do conhecimento do aluno ao longo do tempo. O objetivo é prever com precisão o desempenho dos alunos em futuras interações.

3. **Descrição do Conjunto de dados**
    a. Train.csv
        i. row_id: (int64) código de identificação para a linha;
        ii. timestamp: (int64) o tempo em milissegundos entre esta interação do utilizador e a conclusão do primeiro evento desse usuário;
        iii. user_id: (int32) código de identificação do utilizador;
        iv. content_id: (int16) código de ID para a interação do utilizador;
        v. content_type_id: (int8) 0 se o evento foi uma pergunta feita ao utilizador, 1 se o evento foi o utilizador assistindo a uma palestra.
        vi. task_container_id: (int16) código de identificação para o lote de perguntas ou palestras. Por exemplo, um utilizador pode ver três perguntas seguidas antes de ver as explicações de qualquer uma delas. Todos esses três compartilhariam um task_container_id.
        vii. user_answer: (int8) a resposta do utilizador à pergunta, se houver. Ler -1 como nulo, para palestras.
        viii. **answered_correctly: (int8)** se o utilizador respondeu corretamente. Leia -1 como nulo, para palestras. **(Daqui nasce a Variável dependente)**
        ix. prior_question_elapsed_time: (float32) O tempo médio em milissegundos que um utilizador levou para responder a cada pergunta no pacote de perguntas anterior, ignorando quaisquer aulas intermediárias. É nulo para o primeiro pacote de perguntas ou palestra de um usuário. Observe que o tempo é o tempo médio que um utilizador levou para resolver cada questão no pacote anterior.
        x. prior_question_had_explanation: (bool) se o utilizador viu ou não uma explicação e a (s) resposta (s) correta (s) após responder ao pacote de perguntas anterior, ignorando quaisquer palestras intermediárias. O valor é compartilhado em um único pacote de perguntas e é nulo para o primeiro pacote de perguntas ou palestra do usuário. Normalmente, as primeiras várias perguntas que um usuário vê faziam parte de um teste de diagnóstico de integração, no qual ele não obteve feedback.
    b. Questions.csv
        i. question_id: chave estrangeira para a coluna do train quando o tipo de conteúdo é question (0).
        ii. bundle_id: código para o qual as perguntas são atendidas em conjunto.
        iii. correct_answer: a resposta à pergunta. Pode ser comparado com a coluna user_answer do train para verificar se o utilizador estava certo.
        iv. part: a seção relevante do teste TOEIC .
        v. tags: um ou mais códigos de tag detalhados para a pergunta. O significado das tags não será fornecido, mas esses códigos são suficientes para agrupar as questões.
    c. Lectures.csv
        i. lecture_id: chave estrangeira para a coluna do train content_id, quando o tipo de conteúdo for lecture (1).
        ii. part: código de categoria de nível superior para a palestra.
        iii. tag: um código de tag para a palestra. O significado das tags não será fornecido, mas esses códigos são suficientes para agrupar as palestras.
        iv. type_of: breve descrição do objetivo central da palestra
4. **Supervisionado** – o modelo é construído a partir de um conjunto de dados;
5. **Classificação** - A variável Y, ou dependente é categórica, 1 quando o utilizador acerta na questão, ou 0 quando erra;
6. **Comentários**
        i. O conjunto de dados pertence a uma competição do kaggle; Numa primeira analise conseguimos perceber que vai exigir um grande esforço na parte de transformação de dados.