Universidade de Brasília
Decanato de Pesquisa e Pós-Graduação
Programa de Iniciação Científica – ProIC/UnB

ÁREA DO CONHECIMENTO**:** (x) EXATAS ( ) HUMANAS ( ) VIDA

Título do Projeto

Desenvolvimento de Framework para Construção de Sistemas de Apoio a Tomada de Decisão em Ambiente com Incerteza: Extração de Conhecimento de Base de Dados Estruturadas ou Textuais

Orientador:

Marcelo Ladeira

Unidade Acadêmica/Departamento:

Departamento de Ciência da Computação

PLANO de TRABALHO

Edital 2016 ProIC/CNPq/UnB

Título do Plano de Trabalho

Análise do Desempenho de Alunos da UnB usando um Sistema de Inteligência Artificial

Aluno

Gabriel Ferreira Silva

Matrícula

14/0140131

Plano de Trabalho **(máximo de 5 páginas)**

1. Introdução ao Plano de Trabalho

Diversas instituições disponibilizam publicamente suas informações, seja por questões de transparência ou por perceberem que o estudo sobre tais dados pode indicar oportunidades a serem capitalizadas. Uma aplicação de interesse para dados disponíveis é tentar extrair informações que expliquem a relação entre os dados e, com base em tais informações, construir modelos que façam projeções sobre o futuro ou sobre dados que ainda não dispomos.

Seguindo essa premissa, este projeto visa a construção de um software capaz de fazer projeções acerca do desempenho futuro de um aluno na UnB, e possivelmente de sua remuneração no mercado após sair da faculdade. Serão utilizados dados descaracterizados dos alunos de graduação da UnB disponibilizados por meio do sistema SIGRA. Tais dados incluem dados de perfil, o desempenho do aluno em cada matéria, se um aluno conseguiu ou não concluir a universidade, se houve participação em estágios ou PIBIC. Para análise da inserção no mercado de trabalho nacional serão utilizados dados da RAIS - Relação Anual de Informações Sociais a serem solicitados ao Ministério do Trabalho e Previdência Social.

O software previsto poderia ser usado para identificar alunos que estão com maior risco de não conseguirem completar a universidade ou quais disciplinas oferecem mais risco ou quais disciplinas influenciam mais as outras. O software a ser desenvolvido será baseado na aplicação de técnicas de Mineração de Dados, em focando, em especial, a Aprendizagem de Máquina.

A área de Inteligência Artificial estuda a construção de sistemas computacionais inteligentes [1]. Já a área de Mineração de Dados estuda a aplicação de técnicas para a descoberta de padrões a partir de um Banco de Dados [2]. Na intersecção dessas duas áreas há um campo que vem despertando muito interesse: Aprendizagem de Máquina. O objetivo é conseguir extrair um padrão a partir de uma grande quantidade de dados, e usar tal padrão para com sucesso tentar prever o que ocorrerá com um sistema no futuro ou qual a maneira mais adequada de lidar com dados futuros.

Técnicas de Aprendizagem de Máquina vêm sendo utilizadas em áreas bastante diversas, tais como jogos [3], segurança digital [4] e sistemas de recomendação [5]. Em geral, os problemas estudados por Aprendizagem de Máquina se dividem em duas áreas: classificação e regressão [6]. Para a classificação, o resultado final assume um valor discreto. Técnicas típicas incluem Máquinas de Suporte Vetorial (SVM do inglês Support Vector Machine) e Árvores de Decisão. Para a regressão, o resultado final assume um valor contínuo. Técnicas típicas incluem Regressão Linear e Redes Neurais [6].

2. Metodologia do Plano de Trabalho

Será utilizado o modelo de referência CRISP-DM para tarefas de mineração de dados com as suas seis fases: entendimento do negócio, entendimento dos dados, pré-processamento, modelagem, avaliação e colocação em uso. Antes do desenvolvimento, uma etapa de pré-processamento será realizada com os dados, para garantir que eles estejam em um estado consistente. Em seguida, como é comum na ciência, uma análise exploratória dos dados será feita. Também como etapa inicial, um estudo acerca do estado da arte das técnicas de aprendizagem de máquina será feito.

Durante o desenvolvimento do modelo, utilizar-se-ão várias técnicas de Aprendizagem de Máquina. A escolha das técnicas adequadas será feita com base em sua adequabilidade a situação e a quantidade de dados que se dispõe. Para cada técnica, os dados serão separados em dois conjuntos disjuntos: um de treino (para que o modelo seja capaz de aprender) e outro de teste (para avaliar quão bem o modelo aprendeu). A existência de *overfitting* será analisada comparando a diferença de performance para os dados de treino e de teste. O desempenho das técnicas também será estimado, considerando para isso sua performance nos dados de teste.

3. Resultados Esperados na Execução do Plano de Trabalho

Espera-se que ao final do trabalho, um sistema inteligente com bom desempenho previsor tenha sido desenvolvido. Embora a capacidade de previsão do software seja um fator subjetivo, os desempenhos medidos no conjunto de teste serão fornecidos.

Possibilidades de atividades a serem previstas pelo sistema incluem: se um aluno irá formar ou não, o desempenho em uma matéria que ele ainda não fez, o IRA com o qual ele acabará a universidade, entre outras. Saber quais os alunos estão com maior risco de não concluir a universidade seria útil para a UnB, pois assim um maior direcionamento poderia ser dado a tal grupo. Já prever um desempenho de um aluno em uma matéria seria útil para o estudante ter, de modo personalizado, uma estimativa de quais matérias serão mais difíceis (permitindo assim um melhor planejamento para o decorrer do semestre). Por fim, uma estimativa do IRA com o qual o estudante terminará a universidade também poderia auxiliar no planejamento dele para ingressar no mercado de trabalho.

O software desenvolvido terá código aberto e será flexível e bem documentado, permitindo assim que o seu funcionamento externo seja entendido com clareza por qualquer um e que seu funcionamento interno seja compreendido por aqueles que conhecem Aprendizagem de Máquina. Pela maneira como o software será desenvolvido, o software deve poder ser adaptado ou tomado como base para trabalhos com temas semelhantes.

4. Etapas e Cronograma de Execução do Plano de Trabalho

Etapa 1 – Detalhamento do Plano de Trabalho. Busca por outras possíveis aplicações e metodologias adequadas.

Etapa 2 – Estudo de Aprendizagem de Máquina com pesquisa em livros, artigos e MOOC's. As principais técnicas serão estudadas, assim como seu contexto de aplicação. Técnicas específicas para o contexto do trabalho serão procuradas.

Etapa 3 – Pré-processamento dos dados, de modo a identificar dados incompletos ou outros problemas. Decidir como proceder para o caso de dados inconsistentes.

Etapa 4 – Análise Exploratória dos dados, busca por informações interessantes nos dados.

Etapa 5 – Aplicação do conhecimento adquirido na etapa 2 através do desenvolvimento do modelo. Análise do desempenho via performance nos dados de teste e identificação de overfitting via disparidades na performance de treino e teste.

Etapa 6 – Preparar o relatório final e o pôster para apresentação no PIBIC

Etapa	Mês 1	Mês 2	Mês 3	Mês 4	Mês 5	Mês 6	Mês 7	Mês 8	Mês 9	Mês 10	Mês 11	Mês 12
1	◆											
2	◆	◆	◆	◆								
3		◆										
4			◆									
5			◆	◆	◆	◆	◆	◆	◆	◆	◆	
6											◆	◆

5. Referências Bibliográficas

[1] - Russell, Stuart, Peter Norvig, and Artificial Intelligence. "A modern approach."*Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* 25 (1995): 27.
[2] - Aggarwal, Charu C.*Data mining: The textbook*. Springer, 2015.
[3] - Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search."*Nature* 529.7587 (2016): 484-489.
[4] - https://people.csail.mit.edu/kalyan/AI2_Paper.pdf
[5] - Koren, Yehuda. "The bellkor solution to the netflix grand prize."*Netflix prize documentation* 81 (2009).

[6] – Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*. Berlin, Germany: AMLBook, 2012.