Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências da Computação
SCC0276 — Aprendizado de Máquina

Professor: Fernando Pereira dos Santos **PAE:** Leo Sampaio Ferraz Ribeiro

Projeto Final

Preditor da nota média da escola no ENEM baseado em informações sociodemográficas e socioeconômicas

Alunos	**NUSP**
Antônio Sebastian	10797781
Gabriell Tavares	10716400
Helbert Pinto	10716504

Introdução

A educação é um dos fatores mais importantes para sociedade. Ao longo dos anos vimos que avanços em educação propiciaram avanços na ciência e por consequência uma melhora na qualidade de vida das pessoas. Atualmente no Brasil a maior parte das pessoas tem acesso à educação por meio da rede pública.

Atualmente existem muitos meios de avaliar o desempenho escolar nas diversas regiões do Brasil, entretanto o mais importante deles é o Exame Nacional do Ensino Médio (ENEM) pois concentra em uma única avaliação anual todos os alunos da rede pública egressos do ensino médio. E além do proposito de avaliar o rendimento acadêmico serve como unificador de vestibulares para ingresso em universidades federais.

É sabido que muitos são os fatores que afetam a taxa de aprendizado por parte dos alunos, fazendo com absorvam mais ou menos os conteúdos que são abordados em salas de aula. Fatores socioeconômicos, familiares, emocionais/psicológicos, dentre muitos outros podem intensificar ou quase anular o papel que a escola tem para com a sociedade. A proposta deste projeto é descobrir a relação destes fatores com a educação dos alunos e então conseguir prever a nota média que os alunos de uma cidade obtêm ao realizar o ENEM.

O projeto está sendo armazenado no repositório do GIT:
https://github.com/helbertmoreirapinto/SCC0276-Aprendizado-de-Maquina

Fontes de dados

Para a realização do projeto, precisaremos de informações censuais sobre as cidades, como: população, renda média, média do número de filhos por família etc. Informações dedicadas exclusivamente às escolas, como: média dos salários dos docentes, quantidade de alunos por sala de aula, evasão escolar, se possui bibliotecas/laboratórios etc. E de informações relacionadas às notas médias que os alunos da escola obtiveram no ENEM, que para nosso projeto servirá como a variável resposta, ou o valor que pretendemos prever.

Os dados foram obtidos através do site do Ministério da Educação (MEC) [1]. As seguintes informações presentes em todos os conjuntos de dados:

Variável	Valores	Descrição
Ano	2019	O conjunto possui somente dados referentes a este ano
Região	[Norte, Nordeste, Sudeste, Sul, Centro-Oeste]	Região do Brasil
UF	Sigla da unidade federativa com 2 caracteres	Estado
Código Município	Inteiro	Código do município
Nome Município	Texto	Nome do município
Código Escola	Inteiro	Código da escola
Nome Escola	Texto	Nome da escola
Localização	[Rural, Urbana]	Região da cidade onde está localizada a escola
Dependência Administrativa	[Municipal, Estadual, Federal, Privada]	Órgão gestor da escola

Tabela 1 – Descrição das variáveis comuns presentes nos conjuntos de dados

Os conjuntos de dados que selecionamos são os seguintes:

Esforço docente:

Este conjunto de dados [2] nos fornece informações quanto a carga de trabalho dos professores, pois consideramos que isso pode afetar diretamente na qualidade da aula e por tabela no quanto os alunos vão conseguir aprender do conteúdo.

Grupo	Variável
Ensino Fundamental [1° ao 4° ano]	Nível 1
	Nível 2
	Nível 3
	Nível 4
	Nível 5
	Nível 6
Ensino Fundamental [5° ao 9° ano]	Nível 1
	Nível 2
	Nível 3
	Nível 4
	Nível 5
	Nível 6
Ensino Médio	Nível 1
	Nível 2
	Nível 3
	Nível 4
	Nível 5
	Nível 6

Tabela 2 – Variáveis presentes no conjunto de dados do Esforço Docente

Nível	Descrição
Nível 1	Docente que tem até 25 alunos e atua em um único turno, escola e etapa
Nível 2	Docente que tem entre 25 e 150 alunos e atua em um único turno, escola e etapa
Nível 3	Docente que tem entre 25 e 300 alunos e atua em um ou dois turnos em uma única escola e etapa
Nível 4	Docentes que tem entre 50 e 400 alunos e atua em dois turnos, em uma ou duas escolas e em duas etapas
Nível 5	Docente que tem mais de 300 alunos e atua nos três turnos, em duas ou três escolas e em duas etapas ou três etapas
Nível 6	Docente que tem mais de 400 alunos e atua nos três turnos, em duas ou três escolas e em duas etapas ou três etapas

Tabela 3 – Descrição dos níveis presentes no conjunto de dados do Esforço Docente

Alunos por turma:

Este conjunto de dados [3] informa a quantidade média de alunos por turma. Acreditamos que uma turma muito grande pode influenciar no quanto os professores podem dar de atenção às dúvidas e questionamentos dos estudantes.

Grupo	Variável
Educação Infantil	Total
	Creche
	Pré-Escola
Ensino Fundamental	Total
	Anos Iniciais
	Anos Finais
	1° Ano
	2° Ano
	3° Ano
	4° Ano
	5° Ano
	6° Ano
	7 Ano
	8 Ano
	9 Ano
Ensino Médio	Turmas Multietapa
	Total
	1° Série
	2° Série
	3° Série
	4° Série
	Não-Seriado

Tabela 4 – Descrição das variáveis do conjunto de dados Alunos por turma

Hora-Aula diária:

Este conjunto de dados [4] nos informa qual o tempo médio diário que os alunos têm de aula por dia. Acreditamos que deve existir uma relação entre o tempo médio de estudo com a nota obtida no ENEM. As variáveis possuem os mesmos nomes que os contidos na Tabela 4.

Taxa de rendimento:

Este é um conjunto de dados [5] que informa o nível de aprovação a cada serie escolar. Temos também informações sobre taxas de abandono (evasão) que podem auxiliar na importância que a sociedade local dá à instituição.

Grupo Variável	Grupo	Variável
Taxa de Aprovação	Ensino Fundamental	Total
		Anos Iniciais
		Anos Finais
		1° Ano
		2° Ano
		3° Ano
		4° Ano
		5° Ano
		6° Ano
		7 Ano

		8 Ano
		9 Ano
	Ensino Médio	Total
		1° Série
		2° Série
		3° Série
		4° Série
		Não-Seriado
Taxa de Reprovação	Ensino Fundamental	Total
		Anos Iniciais
		Anos Finais
		1° Ano
		2° Ano
		3° Ano
		4° Ano
		5° Ano
		6° Ano
		7 Ano
		8 Ano
		9 Ano
	Ensino Médio	Total
		1° Série
		2° Série
		3° Série
		4° Série
		Não-Seriado
Taxa de Reprovação	Ensino Fundamental	Total
		Anos Iniciais
		Anos Finais
		1° Ano
		2° Ano
		3° Ano
		4° Ano
		5° Ano
		6° Ano
		7 Ano
		8 Ano
		9 Ano
	Ensino Médio	Total
		1° Série
		2° Série
		3° Série
		4° Série
		Não-Seriado

Tabela 6 – Descrição das variáveis do conjunto de dados de Taxa de rendimento

Indicadores Socioeconômicos:

Este é um conjunto de dados que nos dá informações econômicas dos alunos que frequentam a escola. Os alunos são socialmente classificados em oito grupos, sendo o grupo de Nível 1 o que representa a maior renda. Baseado no nível socioeconômicos dos alunos a escola recebe um score socioeconômico próprio.

Variável	Valores	Descrição
ID_AREA	[Capital, Interior]	Área da Escola (relacionado ao Município)
QTD_ALUNOS_INSE	Inteiro	Quantidade de alunos com INSE calculado utilizado para o cálculo das médias por escola
INSE_VALOR_ABSOLUTO	Real	Média do Indicador de Nível Socioeconômico dos alunos da escola
INSE_CLASSIFICACAO	Texto	Classificação do Nível Socioeconômico da escola
PC_NIVEL_1	Real	Percentual de alunos classificados no Nível I.
PC_NIVEL_2	Real	Percentual de alunos classificados no Nível II.
PC_NIVEL_3	Real	Percentual de alunos classificados no Nível III.
PC_NIVEL_4	Real	Percentual de alunos classificados no Nível IV
PC_NIVEL_5	Real	Percentual de alunos classificados no Nível V
PC_NIVEL_6	Real	Percentual de alunos da Escola classificados no Nível VI
PC_NIVEL_7	Real	Percentual de alunos da Escola classificados no Nível VII
PC_NIVEL_8	Real	Percentual de alunos da Escola classificados no Nível VIII

Tabela 7 – Descrição das variáveis do conjunto de dados de Indicadores socioeconômicos

Censo educação básica:

Este é um conjunto de informações completo [7], com 370 variáveis sobre as dependências da escola, como número de salas, laboratórios ou elevadores por exemplo. Temos neste conjunto de dados informações socioeconômicas básicas, como quantidade de banheiros ou se tem água potável e sua origem. Como este conjunto possui muitas variáveis, não iremos descrevê-las todas neste documento.

Notas ENEM:

Para este conjunto precisamos utilizar uma técnica chamada "*Web Scrapping*" pois no site do Ministério da Educação temos apenas os dados do ENEM 2020, o que à primeira vista seria algo bom (dados mais atualizados), entretanto todos os dados presentes no conjunto são relacionados ao participante (aluno) e não tínhamos nenhuma informação que o relacionasse com a escola que estudou, inviabilizando o conceito do estudo. Encontramos as informações consolidadas do ENEM por escola apenas em sua edição 2019[8].

Neste conjunto de dados temos as notas medias discriminadas por grupos de conhecimento dos alunos que realizaram o exame e obtiveram uma nota estritamente maior que zero em todos os grupos.

Variável	Valores	Descrição
Posição	Inteiro	Posição no ranking geral
Código INEP	Inteiro	Código da escola
Escola	texto	Nome da escola
Estado	Texto	Nome do estado
Cidade	Texto	Nome da cidade
Dependência Administrativa	[Rural, Urbana]	Órgão gestor da escola

Localização	[Municipal, Estadual, Federal, Privada]	Região da cidade onde está localizada a escola
Alunos	Inteiro	Número de alunos contabilizados na média
CH	Real	Nota média de História
CN	Real	Nota média em Ciências da Natureza
LC	Real	Nota média em Linguagens
MT	Real	Nota média em Matemática
RD	Real	Nota média em redação
Média	Real	Nota média final

Tabela 8 – Descrição das variáveis presentes no conjunto de dados das Notas do ENEM 2019

Seleção de variáveis

Como vimos nos tópicos anteriores, pode parecer que estamos utilizando mais variáveis que o necessário, entretanto conseguir identificar escolas que possuem bom rendimento escolar apenas com as variáveis socioeconômicas do ano avaliado (sem contar histórico de rendimento entre outras coisas), exige informações multifacetadas. Todavia salientamos que as bases são de mesma origem (MEC) de forma que muitas variáveis estão contidas em várias bases, que após a remoção significara um número muito menor de variáveis.

Vemos que existe uma enorme quantidade de informações que são uteis e podem ser utilizadas para auxiliar no processo de identificar a nota média final das escolas, mas temos também um grande volume de variáveis que não auxiliam tanto e portanto podem ser removidas.

Deste modo optamos em remover de todos os conjuntos de dados informações referentes ao ensino infantil e fundamental, pois somente alunos do ensino médio podem participar do ENEM.

Nos conjuntos de dados que possuem discriminação por ano do ensino médio, decidimos utilizar os valores consolidados totais, pois é comum ver alunos que não estão em condições de terminar o ensino médio realizando a prova do ENEM. Neste ponto podemos considerar também que a escola onde os alunos cursam as series do ensino médio são muito relevantes para a nota final que o aluno obtém no ENEM.

No conjunto de dados de Censo da Educação Básica, temos algumas variáveis categóricas que transforemos em variáveis *dummy*.

Propostas de desenvolvimento

Como nossos dados são rotulados, temos em mãos um problema de Aprendizado Supervisionado. Como a variável resposta é um valor real, não vemos como utilizar algoritmos de classificação, como KNN ou Arvores de Decisão. Temos como objetivo testar alguns algoritmos de regressão, como uma Regressão Logística por exemplo. E por fim podemos comparar os resultados e tempo de processamento com uma Rede Neural Multicamada.

Referências:

[1] Site do Ministério da Educação - https://www.gov.br/mec/pt-br

[2] Conjunto de dados de Índice de Esforço Docente - https://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2019/IED_2019_ESCOLAS.zip

[3] Conjunto de dados de Alunos por Turma - https://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2019/ATU_2019_ESCOLAS.zip

[4] Conjunto de dados de Horas-Aula diárias - https://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2019/HAD_2019_ESCOLAS.zip

[5] Conjunto de dados de Taxa de rendimento - https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/taxas-de-rendimento

[6] Conjunto de dados de Indicadores Socioeconômicos - https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/nivel-socioeconomico

[7] Conjunto de dados de Censo da Educação Básica - https://download.inep.gov.br/dados_abertos/microdados_censo_escolar_2019.zip

[8] Site do sistema Evolucional - https://enem2019.evolucional.com.br/RD.html