# Universidade Tecnológica Federal do Paraná

## Curso de Engenharia de Produção


# ANÁLISE DO DESEMPENHO ACADÊMICO ENEM/ENADE COMO INSTRUMENTO DE QUALIFICAÇÃO DOS CURSOS DE ENGENHARIA DE PRODUÇÃO NO BRASIL


**Londrina - PR**

2018

**ROBERTO LEMOS ALMEIDA FILHO**

# ANÁLISE DO DESEMPENHO ACADÊMICO ENEM/ENADE COMO INSTRUMENTO DE QUALIFICAÇÃO DOS CURSOS DE ENGENHARIA DE PRODUÇÃO NO BRASIL

Trabalho de Conclusão de Curso apresentado a Universidade Tecnológica Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Engenharia de Produção.

Orientador: Prof. Dr. Rogério Tondato

Londrina - PR
2018

ROBERTO LEMOS ALMEIDA FILHO

# TERMO DE APROVAÇÃO

**ANÁLISE DO DESEMPENHO ACADÊMICO ENEM/ENADE COMO INSTRUMENTO DE QUALIFICAÇÃO DOS CURSOS DE ENGENHARIA DE PRODUÇÃO NO BRASIL**

POR

**ROBERTO LEMOS ALMEIDA FILHO**

Esta Monografia foi apresentada às 14 horas do dia 26 de Novembro de 2018 como requisito parcial para obtenção do título de bacharel em ENGENHARIA DE PRODUÇÃO, Universidade Tecnológica Federal do Paraná – Campus Londrina. O candidato foi arguido pela Banca Examinadora composta pelos professores relacionados abaixo. Após deliberação, a Banca Examinadora considerou o trabalho: **APROVADO**.

Prof. Dr. José Ângelo Ferreira
Banca Examinadora

Prof. Dra. Silvana Rodrigues Quintilhano
Banca Examinadora

Prof. Dr. Rogério Tondato
Presidente da Banca Examinadora
Orientador

ALMEIDA FILHO, Roberto L. **Análise do Desempenho Acadêmico Enem/Enade como Instrumento de Qualificação dos Cursos de Engenharia de Produção no Brasil.** 2018. 58 p. Trabalho de Conclusão de Curso (Graduação em Engenharia de Produção) - Universidade Tecnológica Federal do Paraná. Londrina - PR. 2018.

## RESUMO

O presente estudo visou mapear a possibilidade de se criar uma métrica avaliativa do desempenho acadêmico pelo perfil do estudante. Também se propôs a analisar a classificação da qualidade dos cursos de Engenharia de Produção no Brasil.Essa pesquisa buscou entender se o conceito utilizado no Enade, para classificar a qualidade dos cursos de Engenharia de Produção, realmente indica a realidade do curso, ou se esse indicador reflete outros desempenhos acadêmicos, não ligados diretamente a qualidade do curso.Para tanto, foi criada uma regressão da métrica utilizada na classificação da qualidade do curso para ser aplicada a nível aluno e, primeiramente, verificar se o resultado é relevante para identificar a qualidade do curso, e posteriormente, se a qualidade do curso está atrelada a categoria administrativa da IES, ou se está atrelada ao perfil do estudante.Concluiu-se que a proporção de alunos classificados pelo desempenho dos mesmos no Enem é mais relevante para a classificação da qualidade de um curso do que a categoria administrativa da IES, e que não há diferença de performance entre estudantes de IES de categoria administrativa Federal ou Particular, ou seja, o resultado da qualidade do curso é mais dependente do desempenho no Enem dos alunos que estão sendo avaliados no curso do que da categoria da IES. Também é possível predizer, com 72% de assertividade, a classificação dos alunos no Enade baseado apenas nas notas do Enem. Esse resultado pode ser obtido de duas formas: por regressão linear múltipla ou por *machine learning*, pelo método da Árvore de Decisão, sendo o segundo, o que melhor explica o comportamento dos dados.

Palavras-chave: Enade. Enem. Predição. Regressão linear múltipla.

ALMEIDA FILHO, Roberto L. Enem / Enade Academic Performance Analysis as a Qualification Instrument for Production Engineering Courses in Brazil. 2018. 58 p. Course Completion Work (Graduation in Production Engineering) - Federal Technological University of Paraná. Londrina - PR. 2018.

## ABSTRACT

The present study aimed at mapping the possibility of creating an evaluation metric of academic performance by the student profile. It was also proposed to analyze the quality classification of Production Engineering courses in Brazil.This research sought to understand if the concept used in Enade, to classify the quality of the courses of Production Engineering, really indicates the reality of the course, or if this indicator reflects other academic performances, not directly linked to the quality of the course.
To do so, a regression of the metric used to classify the quality of the course to be applied at the student level was created, and first, to verify if the result is relevant to identify the quality of the course, and later, if the quality of the course is linked to IES administrative category, or if it is linked to the student profile.It was concluded that the proportion of students classified by their performance in the Enem is more relevant to the classification of the quality of a course than the administrative category of the HEI, and that there is no difference in performance between students of HEIs of the Federal administrative category or Particular, that is, the quality of the course result is more dependent on the Enem performance of the students being evaluated in the course than in the HEI category.It is also possible to predict, with 72% assertiveness, the student ranking in Enade based only on the Enem grades. This result can be obtained in two ways: by multiple linear regression or by machine learning, by the Decision Tree method, the second being the one that best explains the behavior of the data.

Key-word: Enade. Enem. Prediction. Multiple linear regression.

## AGRADECIMENTOS

Primeiramente a Deus, porque por meio dele são concedidas a vida e capacidade para se fazer todas as coisas.

Aos meus pais, por me proporcionarem tudo para chegar até aqui.

À minha noiva Samara, por ter paciência, amor e por todo o apoio em todos os momentos durante essa jornada acadêmica.

Ao meu líder no trabalho, Lucas Borges Hanusch, pelo esforço, inspiração e ensinamentos que, desde 2016, mudaram toda a minha história e, com toda certeza, direcionarão meu futuro.

A todos os colegas de turma que compartilharam comigo essa jornada, e em especial à Débora, que ajudou a todos desde o princípio do curso.

A todos os colegas do trabalho pela paciência e ajuda nos momentos críticos dessa formação acadêmica.

# EPÍGRAFE

“Inovar é descobrir, imaginar, criar ou melhorar (...) prever, analisar, programar e orçamentar, depois investir e correr riscos, ou ainda convencer, motivar, organizar, negociar (...) ultrapassar os obstáculos, enfrentando as resistências mesmo psicológicas ou burocráticas, colocar em causa a ordem estabelecida, ir contra os preconceitos, contra a inércia ou a concorrência desleal, é mesmo se expor as mesquinharias; enfim, é aproveitar a vantagem que representa a introdução da novidade” (BARREYRE, 1975).

## LISTA DE FIGURAS

Figura 1 - Gráfico de árvore de decisão ..... 25
Figura 2 - Filtro do curso de Engenharia de Produção ..... 32
Figura 3 - Filtro de alunos do curso de Engenharia de Produção, Concluintes, Presentes e com todas as notas do Enem na base de IDD ..... 32
Figura 4 - Cálculo do Enem Médio ..... 33
Figura 5 - Teste de Correlação entre as Variáveis Nota Geral Bruta e Nota Geral Contínua (Faixa) ..... 33
Figura 6 - Regressão Linear entre as Variáveis Nota Geral Bruta e Nota Geral Contínua (Faixa) ..... 34
Figura 7 - Resultado do *summary(converte_faixa)* ..... 34
Figura 8 - Representação Gráfica da Regressão ..... 35
Figura 9 - Classificação da Categoria Administrativa da IES ..... 35
Figura 10 - Vinculação da Categoria Administrativa da IES na Base de Microdados IDD ..... 36
Figura 11 - Classificação Simplificada da Categoria Administrativa da IES ..... 36
Figura 12 - Explicação Quartil ..... 36
Figura 13 - Criação da Classificação dos Alunos por Quartil da Nota Média do Enem ..... 36
Figura 14 - Criação e Vinculação da Nota Geral Contínua (Faixa) ..... 37
Figura 15 - Criação da Coluna conceito_enade ..... 37
Figura 16 - Criação da coluna conceito ..... 37
Figura 17 - Criação da Função regressao_enem ..... 38
Figura 18 - Resultado do *summary(regressão_enem)* ..... 38
Figura 19 - Regressão Linear Múltipla entre as notas Enem dos alunos e Nota Geral Bruta ..... 39
Figura 20 – Criação da fx_enade_predict ..... 39
Figura 21 - Criação do conceito enade_predict ..... 39
Figura 22 - Criação do conceito_predict ..... 40
Figura 23 - Separação das bases de treino e teste ..... 40
Figura 24 - Criação do modelo de predição por Árvore de Decisão ..... 40
Figura 25 - Resultado do *summary(floresta)* ..... 41

Figura 26 - Predição do resultado dos alunos na base de teste.....41
Figura 27 - Criação do conceito_enade_ml.....41
Figura 28 - Criação do conceito_ml.....42
Figura 29 - Predição do resultado dos alunos na base completa de dados.....42
Figura 30 - Criação do conceito_enade_ super_ajustado.....43
Figura 31 - Criação do conceito_super_ajustado.....43
Figura 32 - Matriz de confusão entre resultado da regressão e resultado real.....46
Figura 33 - Matriz de confusão entre resultado do modelo aplicado na base teste e resultado real na base teste.....48
Figura 34 - Matriz de confusão entre resultado do modelo aplicado na base completa de dados e resultado real na base completa de dados.....49

## LISTA DE TABELAS

Tabela 1 – Parâmetros de Conversão do $\boldsymbol{NCj}$ em Conceito Enade .....22

## LISTA DE GRÁFICOS

Gráfico 1 - Percentual conceito satisfatório Enade no Brasil em 329 IES que ofertam curso de Engenharia de Produção ..... 29
Gráfico 2 - Percentual conceito curso satisfatório x Quantidade cursos por categoria administrativa ..... 30
Gráfico 3 - Percentual conceito curso x Média Enem dos alunos por categoria administrativa ..... 31
Gráfico 4 - Proporção de alunos por quartil Enem x Enem médio ..... 31
Gráfico 5 - Desempenho no Enade classificado em categoria administrativa e performance no Enem ..... 44
Gráfico 6 - Comportamento da regressão linear múltipla x Comportamento real dos desempenhos de Enem x Enade ..... 45
Gráfico 7 - Desempenho no Enade classificado em categoria administrativa e performance no Enem com base nos resultados reais e nos resultados obtidos da classificação por regressão linear múltipla ..... 46
Gráfico 8 - Comportamento treino da Árvore de Decisão x Comportamento real dos desempenhos de Enem x Enade na base treino ..... 47
Gráfico 9 - Comportamento teste da Árvore de Decisão x Comportamento real dos desempenhos de Enem x Enade na base teste ..... 48
Gráfico 10 - Comportamento teste da Árvore de Decisão x Comportamento real dos desempenhos de Enem x Enade na base completa de dados ..... 49
Gráfico 11 - Desempenho no Enade classificado em categoria administrativa e performance no Enem com base nos resultados reais e nos resultados obtidos da classificação por regressão linear múltipla e pela Árvore de Decisão no modelo super ajustado ..... 50

## LISTA DE EQUAÇÕES

Equação 1 ..... 21
Equação 2 ..... 23
Equação 3 ..... 24

## LISTA DE SIGLAS, ABREVIAÇÕES E SÍMBOLOS

| | | |
|---|---|---|
| Enade | - | Exame Nacional de Desempenho de Estudantes |
| Enem | - | Exame Nacional do Ensino Médio |
| Sinaes | - | Sistema Nacional de Avaliação da Educação Superior |
| Conaes | - | Comissão Nacional de Avaliação da Educação Superior |
| Inep | - | Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira |
| CPC | - | Conceito Preliminar de Curso |
| IES | - | Instituição de Ensino Superior |
| IDD | - | Indicador de Diferença entre os Desempenhos |
| MEC | - | Ministério da Educação |
| IGC | - | Índice Geral de Cursos |

**SUMÁRIO**

1. INTRODUÇÃO ..... 16
1.1. Justificativa e Caracterização do Problema ..... 16
1.2. Objetivo Geral ..... 16
1.3. Objetivos Específicos ..... 17
2. REVISÃO BIBLIOGRÁFICA ..... 18
2.1 Sistema de Avaliação Institucional ..... 18
2.2 Enade ..... 19
2.2.1 O que é o Enade? ..... 19
2.2.2 Composição da Nota ..... 20
2.2.3 Cálculo da Nota ..... 21
2.3 IDD ..... 22
2.4 Regressão Linear ..... 23
2.4.1 Regressão Linear Simples ..... 23
2.4.2 Regressão Linear Múltipla ..... 23
2.5 *Machine learning* ..... 24
3. MÉTODO DA PESQUISA ..... 27
3.1 Fases do método ..... 27
4. ANÁLISE DE DADOS ..... 28
4.1 Levantamento de Dados ..... 28
4.2 Análise Exploratória ..... 29
4.3 Conversão em Faixa Enade por aluno por Regressão Linear Simples ..... 32
4.4 Predição de Faixa Enade por aluno por Regressão Linear Múltipla ..... 38
4.5 Predição de Faixa Enade por aluno por *Machine learning* - Árvore de Decisão 40
5. RESULTADOS ..... 44
6. CONCLUSÃO ..... 53

7. BIBLIOGRAFIA ..... 54
8. ANEXOS ..... 56
8.1 ANEXO A – Dicionário de Variáveis IDD ..... 57
8.2 ANEXO B – Dicionário de Variáveis Conceito Enade ..... 59

# 1. INTRODUÇÃO

## 1.1. JUSTIFICATIVA E CARACTERIZAÇÃO DO PROBLEMA

O Exame Nacional de Desempenho de Estudantes - Enade é um dos procedimentos de avaliação do Sistema Nacional de Avaliação da Educação Superior - Sinaes, realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Inep, segundo diretrizes estabelecidas pela Comissão Nacional de Avaliação da Educação Superior - Conaes, órgão colegiado de coordenação e supervisão do Sinaes.

De acordo com o Inep (2015a), o Enade tem por objetivo acompanhar o processo de aprendizagem e do desempenho acadêmico dos estudantes relacionado aos conteúdos programáticos previstos nas diretrizes curriculares do respectivo curso de graduação, suas habilidades para ajustamento às exigências decorrentes da evolução do conhecimento e suas competências para entender temas exteriores ao âmbito específico de sua profissão, ligados à realidade brasileira e mundial e a outras áreas do conhecimento.

De acordo com Deconto (2012, p.1)

> O cálculo do Conceito Preliminar de Curso (CPC) de desempenho de uma Unidade de Observação, que consiste no conjunto de cursos que compõe uma área de avaliação específica do Enade de uma Instituição de Educação Superior (IES) em um dado município, resulta em uma nota de 1 a 5 - sendo satisfatórias notas acima de 3.

Segundo Deconto (2012), a avaliação possui três índices em sua composição: primeiro a nota dos alunos concluintes, segundo a nota do Indicador de Diferença entre os Desempenhos (IDD) – o observado e o esperado – e, por fim, um questionário, no qual os alunos dão suas opiniões sobre a infraestrutura, a organização didático-pedagógica, professores doutores, professores mestres, professores com regime de dedicação integral ou parcial. Além disso, a partir de 2011, a nota do Enem passou a dispensar os estudantes ingressantes de fazer a prova do Enade, dado que o resultado do Enem é utilizado para o cálculo do CPC.

## 1.2. OBJETIVO GERAL

O objetivo dessa pesquisa será estabelecer uma métrica avaliativa do desempenho acadêmico pelo potencial do estudante, de acordo com o Enem, e a

classificação Enade dos cursos de Engenharia de Produção. Para tanto, será criado um modelo de predição de nota Enade baseado no Enem.

### 1.3. OBJETIVOS ESPECÍFICOS

Levantar o referencial teórico com os microdados do conceito Enade e do IDD disponibilizados pelo Inep, bem como os dicionários de variáveis.

Realizar uma análise exploratória do comportamento do curso de Engenharia de Produção no Brasil, contextualizando o cenário proposto.

Criar uma função de conversão que possibilita a conversão da nota geral bruta em nota geral contínua (Faixa) para os alunos.

Criar uma regressão linear entre a nota geral contínua (Faixa) e as notas dos alunos no Enem, identificando a tendência dos dados.

Cruzar as variáveis levantadas na pesquisa e criar uma modelo de regressão linear múltipla e modelo de *machine learning* como tentativa de predição da nota Enade.

Comparar os resultados e assertividade dos modelos.

# 2. REVISÃO BIBLIOGRÁFICA

Para desenvolvimento dessa pesquisa fez-se necessária uma abordagem teórica sobre o Sistema de Avaliação Institucional (Sinaes), o Exame Nacional do Ensino Médio (Enem) e Exame Nacional de Desempenho dos Estudantes (Enade), para compreensão do sistema e processo de avaliação e qualificação dos cursos superiores no Brasil. Foi, também, realizada uma apresentação teórica do conceito de regressão linear múltipla para aplicação prática no sistema de avaliação do Inep, conceito de aprendizado de máquina (*Machine learning*) e o método específico de árvore de decisão.

## 2.1 SISTEMA DE AVALIAÇÃO INSTITUCIONAL

Para abordar o assunto referente ao sistema de avaliação institucional dos cursos de graduação do ensino superior, inicia-se entendendo o órgão responsável por essa avaliação, o Sinaes.

Segundo o Inep (2015d), em 14 de abril de 2004 foi criado, pela Lei n° 10.861, o Sistema Nacional de Avaliação da Educação Superior (Sinaes) que é formado por três componentes principais: a avaliação das instituições, dos cursos e do desempenho dos estudantes. Este sistema faz avaliação de todos os aspectos relacionados a esses três eixos, sendo os principais o ensino, a pesquisa, a extensão, a responsabilidade social, o desempenho dos alunos, a gestão da instituição, o corpo docente e as instalações.

De acordo com o Inep (2015d), os objetivos principais da avaliação são: melhorar o mérito e o valor das instituições, áreas, cursos e programas, dentro das dimensões de ensino, pesquisa, extensão, gestão e formação, além de melhorar a qualidade da educação superior e orientar a expansão da oferta, assim como promover a responsabilidade social das IES, respeitando a identidade da instituição e a autonomia de cada organização.

O Sinaes contém uma série de instrumentos complementares como a auto avaliação, a avaliação externa, o Enade, a avaliação dos cursos de graduação e instrumentos de informação como o censo e o cadastro. A junção destes instrumentos permite que sejam dados alguns conceitos, organizados em uma escala com cinco níveis, para cada uma das dimensões e ao conjunto das dimensões avaliadas. O

Ministério da Educação disponibiliza publicamente o resultado da avaliação das instituições de ensino superior e de seus cursos (Inep, 2015d).

"A divulgação abrange tanto instrumentos de informação quanto os conceitos das avaliações para os atos de Renovação de Reconhecimento e de Recredenciamento (parte do ciclo trienal do Sinaes, com base nos cursos contemplados no Enade a cada ano)" (INEP, 2015d, p.1).

De acordo com Deconto (2012) o cálculo do Conceito Preliminar de Curso (CPC) de desempenho de uma Unidade de Observação, que se constitui no conjunto de cursos que formam uma área de avaliação específica do Enade de uma Instituição de Educação Superior (IES) em um município específico, resulta em uma nota de 1 a 5 - sendo as notas satisfatórias acima de 3.

Segundo Ministério da Educação - MEC (2015) o Índice Geral de Cursos - - IGC é estruturado com base em uma média ponderada das notas dos cursos de graduação e pós-graduação de cada instituição. Desta forma, a qualidade de todos os cursos de graduação, mestrado e doutorado da mesma instituição de ensino resume-se em um único indicador.

Segundo Inep (2015d, p.1)

> Os resultados das avaliações possibilitam traçar um panorama da qualidade dos cursos e instituições de educação superior no país. Os processos avaliativos são coordenados e supervisionados pela Comissão Nacional de Avaliação da Educação Superior (Conaes) e a operacionalização é de responsabilidade do Inep

Conforme Inep (2015d), as informações adquiridas com o Sinaes são utilizadas pelas instituições de ensino superior para auxiliar na sua eficácia operacional e efetividade acadêmica e social, pelos órgãos governamentais para destinar a elaboração de políticas públicas e pelos estudantes, pais de alunos, instituições acadêmicas e público em geral, para conduzir suas decisões quanto à realidade dos cursos e das instituições.

## 2.2 ENADE

### 2.2.1 O QUE É O ENADE?

Segundo o Inep (2015a), o Enade é um exame de conclusão do curso, que avalia como está o rendimento dos concluintes dos cursos de graduação, relacionados

aos conteúdos programáticos, habilidades e competências adquiridas em sua formação. Tal exame é obrigatório e o histórico escolar do estudante deverá conter a situação de regularidade no exame. A primeira aplicação do Enade ocorreu em 2004 e o intervalo máximo para aplicação da avaliação é a cada três anos para cada área do conhecimento.

Segundo o Inep (2015a, p.1)

> O Objetivo do Enade é o acompanhamento do processo de aprendizagem e do desempenho acadêmico dos estudantes em relação aos conteúdos programáticos previstos nas diretrizes curriculares do respectivo curso de graduação, suas habilidades para ajustamento às exigências decorrentes da evolução do conhecimento e suas competências para compreender temas exteriores ao âmbito específico de sua profissão, ligados à realidade brasileira e mundial e a outras áreas do conhecimento

De acordo com o Inep (2015a, p.1) “o Conceito Enade é calculado para cada unidade de observação, constituída pelo conjunto de cursos que compõe uma área de avaliação específica do Enade de uma mesma Instituição de Ensino Superior (IES) em um determinado município”. A partir de 2008, o Conceito Enade começou a considerar em seu cálculo somente o desempenho dos estudantes concluintes. Desta forma, todos os cálculos que serão explicados posteriormente, consideram apenas os tais estudantes, inscritos na condição de regular, que compareceram ao exame, ou seja, os estudantes concluintes participantes do Enade em 2014 (Inep, 2015).

### 2.2.2 COMPOSIÇÃO DA NOTA

“Todas as medidas originais, referentes ao Conceito Enade, são padronizadas e reescalonadas para assumirem valores de 0 (zero) a 5 (cinco), na forma de variáveis contínuas”. (Inep, 2015c, p.1)

De acordo com o Inep (2015c), o processo de padronização e reescalonamento deve passar por duas etapas: primeiro o cálculo do afastamento padronizado de cada unidade de observação, usando-se as médias e os desvios-padrão calculados por área de avaliação e posteriormente a transformação dos afastamentos padronizados em notas padronizadas que podem variar de 0 a 5.

Para iniciar o cálculo do Conceito Enade de uma unidade de observação deve-se obter o desempenho médio de seus concluintes na Formação Geral (FG) e no Componente Específico (CE) (Inep, 2015c).

Conforme Inep (2015c), a prova é constituída de duas partes para que possa atender aos objetivos a parte cabível a avaliação. A primeira parte é a Formação Geral (FG) que é composta de 10 questões, sendo 8 de múltipla escolha e 2 discursivas. A segunda parte é a de Componentes Específicos (CE) que é composta de 30 questões, sendo 27 de múltipla escolha e 3 discursivas.

"A partir das questões de Formação Geral, espera-se que os graduandos evidenciem a compreensão de temas que transcendam ao seu ambiente próprio de formação profissional específico e que sejam importantes para a realidade contemporânea" (Inep, 2015c, p.1).

De acordo com o Inep (2015c), a parte de Componente Específico abrange a particularidade de cada área e de suas possíveis modalidades, tanto no domínio dos conhecimentos, quanto nas habilidades esperadas para o perfil profissional, e investiga conteúdo do curso por meio da busca de diferentes níveis de complexidade.

### 2.2.3 CÁLCULO DA NOTA

Conforme Inep (2018, p.5) "a Nota dos Concluintes no Enade da unidade de observação j (NCj) é a média ponderada das notas padronizadas da respectiva unidade de observação em, sendo 25% o peso da Formação Geral e 75% o peso do Componente Específico da nota final", conforme **Equação 1** abaixo:

**Equação 1**

$$NC_j = 0{,}25x\ NP_{FGj} + 0{,}75x\ NP_{CEj}$$

Onde:

$NC_j$ é a nota dos concluintes no Enade da unidade de observação j;

$NP_{FGj}$ é a nota padronizada em FG da unidade de observação j; e

$NP_{CEj}$ é a nota padronizada em CE da unidade de observação j.

De acordo com o Inep (2018, p.5) "o Conceito Enade é uma variável discreta que assume valores de 1 a 5, resultante da conversão da Nota dos

Concluintes no Enade da unidade de observação j ($NC_j$)", realizada conforme definido na **Tabela 1**.

**Tabela 1 – Parâmetros de Conversão do $NC_j$ em Conceito Enade**

| **CONCEITO ENADE** (Faixa) | $NC_j$ (Valor contínuo) |
|---|---|
| **1** | $0 \leq NC_j < 0{,}945$ |
| **2** | $0{,}945 \leq NC_j < 1{,}945$ |
| **3** | $1{,}945 \leq NC_j < 2{,}945$ |
| **4** | $2{,}945 \leq NC_j < 3{,}945$ |
| **5** | $3{,}945 \leq NC_j \leq 5$ |

**Fonte:** Inep (2015)

"As unidades de observação com menos de 2 (dois) concluintes participantes no Exame não obtêm o Conceito Enade, ficando "Sem Conceito (SC)". Isso ocorre para preservar a identidade do estudante, de acordo com o exposto no § 9º do artigo 5º da Lei nº 10.861, de 14 de abril de 2004" (Inep, 2018, p.2).

## 2.3 IDD

Segundo Inep (2017, p.1)

> O IDD é um indicador de qualidade que busca mensurar o valor agregado pelo curso ao desenvolvimento dos estudantes concluintes, considerando seus desempenhos no Enade e no Enem, como medida aproximada das suas características de desenvolvimento ao ingressar no curso de graduação avaliado

O cálculo do IDD ocorre para cada indivíduo, desde 2014, que tenha participado do Enade e do Enem, encontrando-se os resultados do mesmo estudante nos dois exames a partir do número do CPF (Inep, 2017).

De acordo com o Inep (2017), o IDD mantém relação direta com o Ciclo Avaliativo do Enade, sendo os cursos avaliados segundo as áreas de avaliação a ele vinculadas. O Ciclo Avaliativo do Enade foi definido pelo art. 33. Da Portaria nº 40, de 12 de dezembro de 2007, republicada em 2010. Este ciclo compreende a avaliação

periódica dos cursos de graduação, com referência nos resultados a cada 3 anos de desempenho dos estudantes.

## 2.4 Regressão Linear

Petenate (2017, p.1) explica que

> análises de regressão tem por objetivo desvendar o comportamento entre uma variável dependente e as consideradas independentes. Modelos matemáticos são capazes de explicar essa relação por meio de uma equação que correlaciona a variável dependente com as independentes

Para o presente trabalho, será aplicada a regressão, na tentativa de explicar a relação entre as variáveis Enem e Enade, além da desconstrução da conversão entre a nota geral bruta e a nota bruta contínua (por faixa), publicadas pelo próprio Inep.

### 2.4.1 Regressão Linear Simples

Segundo Petenate (2017, p.1), "na regressão linear simples, a relação entre duas variáveis pode ser representada por uma linha reta, criando uma relação direta de causa e efeito". Desta forma é possível prever os valores de uma variável dependente com base em resultados da variável independente, como ocorre num gráfico de uma equação de primeiro grau, conforme **Equação 2** abaixo:

**Equação 2**

$$y = aX + b$$

A aplicação da regressão simples, objetiva criar a equação da reta conversora entre a nota geral bruta dos cursos para o conceito contínuo, que será aplicado a nota geral dos alunos.

### 2.4.2 Regressão Linear Múltipla

Pentenate (2017, p.1) ainda explica que, "muitas vezes uma única variável preditora não será capaz de explicar tudo a respeito da variável resposta". No caso do

presente trabalho, a performance do Enade (variável resposta) é influenciada por diversas variáveis, tais como sexo, idade, engajamento, escolaridade entre outras. Assim, será preciso realizar uma regressão linear múltipla para tentar explicar a resposta.

Dessa forma, uma equação linear pode ser representada pela **Equação 3** abaixo:

**Equação 3**

$$y = aX_0 + bX_1 + cX_2 + \cdots + nX_n$$

A aplicação da regressão múltipla, objetiva entender o comportamento das variáveis da composição da nota Enem relacionadas a nota Enade e ao comportamento da nota geral Enade para conversão nos conceitos.

### 2.5 *MACHINE LEARNING*

Um dos fatores que utilizaremos para a criação do modelo de predição é uma árvore de decisão baseada em *Machine learning*. Para contextualizar sobre o tema, segundo Matos (2015, p.1):

> *Machine learning* é um conjunto de regras e procedimentos, que permite que os computadores possam agir e tomar decisões baseados em dados ao invés de ser explicitamente programados para realizar uma determinada tarefa. Programas de *Machine learning* também são projetados para aprender e melhorar ao longo do tempo quando expostos a novos dados

Para Monaco (2017, p.1), “os algoritmos de Machine learning podem ser divididos em 3 categorias, aprendizagem supervisionada, aprendizagem não supervisionada e aprendizado por reforço”. Primeiramente, a aprendizagem supervisionada é útil quando há uma propriedade disponível para um certo conjunto de dados. Em segundo lugar, o aprendizado não supervisionado é útil quando o desafio é descobrir relacionamentos implícitos em um certo conjunto de dados não-rotulados. Por último, o aprendizado de reforço está entre estes dois extremos – há uma forma de feedback disponível para cada passo ou ação preditiva, mas sem etiqueta específica ou mensagem de erro (Monaco, 2017).

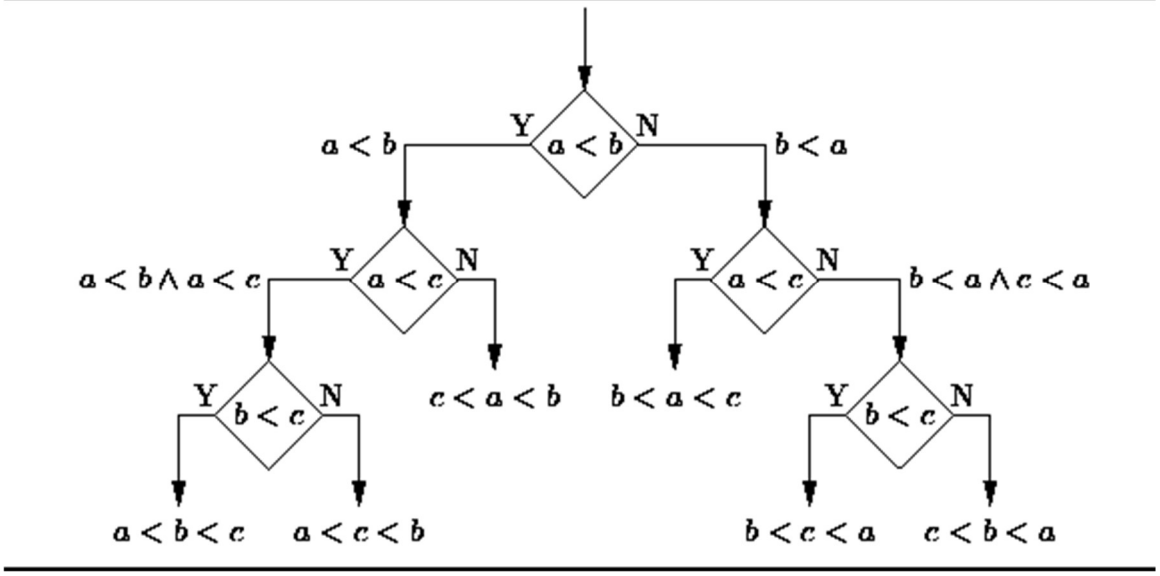
Para o presente projeto, será utilizado um algoritmo de aprendizagem supervisionada, a árvore de decisão. Esse conceito “é uma ferramenta de apoio à decisão que usa um gráfico de árvore ou modelo de decisões e suas possíveis consequências. Uma árvore de decisão é também uma maneira de representar visualmente um algoritmo” (Matos, 2015, p.1).

Segundo Pichiliani (2006, p.1), árvore de decisão funciona da seguinte forma: “com base nos registros do conjunto de treinamento, uma árvore é montada e, a partir desta árvore, pode-se classificar a amostra desconhecida sem necessariamente testar todos os valores dos seus atributos”.

A ideia geral de métodos baseados em árvores é particionar o espaço recursivamente em retângulos, nos quais um modelo simples é aprendido (Campos, 2017).

Segundo Monaco (2017, p.1), “uma árvore de decisão é uma ferramenta de apoio que utiliza um gráfico ou modelo de decisões e suas possíveis consequências, incluindo resultados de eventos fortuitos, custos de recursos e utilidade”, conforme pode ser observado na **Figura 1**.

**Figura 1 - Gráfico de árvore de decisão**




**Fonte:** Monaco, 2017

Para Monaco (2017), do ponto de vista da decisão de negócios, uma árvore de decisão é o menor número de perguntas que precisam ser respondidas para mensurar a probabilidade de tomar uma decisão certa, na maioria das vezes. Como

um método, lhe é permitido abordar o problema de uma forma estruturada e sistemática para chegar a uma conclusão lógica.

# 3. MÉTODO DA PESQUISA

## 3.1 FASES DO MÉTODO

Quanto à natureza, essa pesquisa é quantitativa, pois de acordo com Terencee Escrivão Filho (2006, p.3), “nos estudos organizacionais, a pesquisa quantitativa permite a mensuração de opiniões, reações, hábitos e atitudes em um universo, por meio de uma amostra que o represente estatisticamente”.

Como serão utilizados dados do Inep para mensurar a nota do curso em um campus, além da nota atribuída ao curso e o desempenho dos alunos na prova Enade, bem como dos dados acadêmicos de ingresso e desenvolvimento dos alunos ao longo do curso, a pesquisa é de natureza quantitativa.

Quanto ao objetivo da pesquisa, ela é exploratória, pois este tipo de pesquisa preocupa-se em proporcionar maior familiaridade com o problema, com objetivo de torná-lo mais explícito ou de construir hipóteses. Normalmente essas pesquisas envolvem levantamento bibliográfico, entrevistas com pessoas que tiveram experiências práticas com o problema pesquisado ou análise de exemplos que estimulem a compreensão (GERHARDT & SILVEIRA, 2009, p.35 apud GIL, 2007).

Como será realizado um estudo de coleta de dados divulgados pelo Inep ao final do ciclo do Enade, e também será feita análise, classificação e interpretação dos dados coletados, desta forma a pesquisa é de objetivo explicativa.

O presente trabalho será dividido em 6 etapas:

1ª Etapa: Levantar os microdados do conceito Enade e do IDD disponibilizados pelo Inep referentes ao ano de 2014, bem como os dicionários de variáveis.

2ª Etapa: Realizar uma análise exploratória do comportamento do curso de Engenharia de Produção no Brasil hoje.

3ª Etapa: Criar uma função de conversão entre a nota bruta geral do Enade em nota geral contínua (Faixa)

4ª Etapa: Criar regressão linear múltipla a partir das notas dos alunos no Enem para predição da nota geral contínua (Faixa) no Enade.

5ª Etapa: Criar modelo de *machine learning* pelo método de árvore de decisão a partir das notas dos alunos no Enem para predição da nota geral contínua (Faixa) no Enade.

6ª Etapa: Realizar análise dos resultados obtidos

# 4. ANÁLISE DE DADOS

## 4.1 LEVANTAMENTO DE DADOS

As bases de dados utilizadas para iniciar a elaboração dos cálculos, foram dos microdados do IDD fornecidos pelo Inep referentes ao ano de 2014 e a base de conceito Enade por curso também fornecida pelo Inep.

Inicialmente as tabelas de microdados do IDD e de conceito Enade foram inseridas no Excel e no R. Todas as colunas da tabela de microdados do IDD estão explicadas no Anexo A bem como as colunas da tabela de conceito Enade estão explicadas no Anexo B.

Para complementar as bases no Excel com o objetivo de conseguir iniciar os cálculos e análises, realizou-se o cálculo da média do Enem por aluno dentro da base de microdados utilizando as seguintes colunas: enem_nt_cn, enem_nt_ch, enem_nt_lc, enem_nt_mt e o valor obtido foi inserido na coluna enem_medio. Posteriormente foram inseridas 3 colunas na tabela de conceito Enade, sendo elas: Conceito Curso (regra de conceito, se ele for menor que 3 será classificado como insatisfatório e se for maior ou igual a 3 será classificado como Satisfatório), Tipo IES (definição de qual modalidade pertence a IES, entre Federal, Particular e Estadual) e Média Enem (coluna enem_medio que foi calculado na tabela de microdados).

Após a inserção, foram realizados alguns procedimentos para que a base ficasse o mais confiável possível para elaboração dos cálculos. Os alunos que não possuíam as 4 colunas referentes ao Enem preenchidas (enem_nt_cn, enem_nt_ch, enem_nt_lc, enem_nt_mt) foram removidos, pois poderiam enviesar os dados, tornando a análise menos confiável. Foram removidos, também, os alunos não concluintes e ausentes. Todos os cursos diferentes de Engenharia de Produção (co_grupo=6208) foram removidos da base, visto que as análises desse trabalho serão realizadas apenas para o curso mencionado.

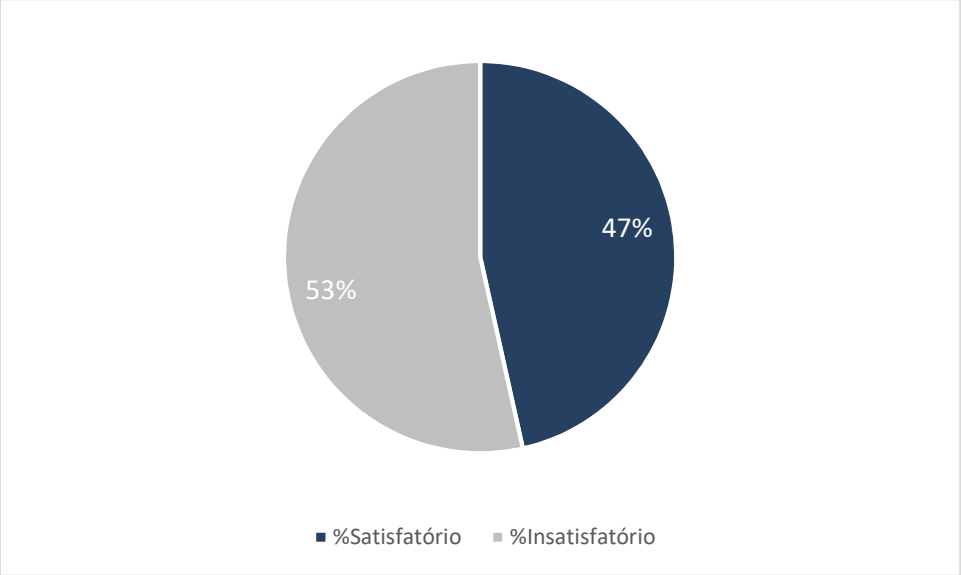
A partir dessas 2 tabelas, foram realizados os cálculos e análises exploratórios explicados no tópico 4.2, que foram feitos no Excel.

As tabelas inseridas no R (microdados do IDD e conceito Enade) são as que foram utilizadas para os cálculos e gráficos da predição do Enade, os quais serão explicados nos tópicos 4.3 e 4.4.

## 4.2 ANÁLISE EXPLORATÓRIA

Para iniciar a análise exploratória, foi realizado o seguinte questionamento: como é o índice satisfatório do curso de Engenharia de Produção no Brasil hoje? Para responder ao questionamento, foi elaborado o **Gráfico 1** que se baseou na tabela de conceito Enade.

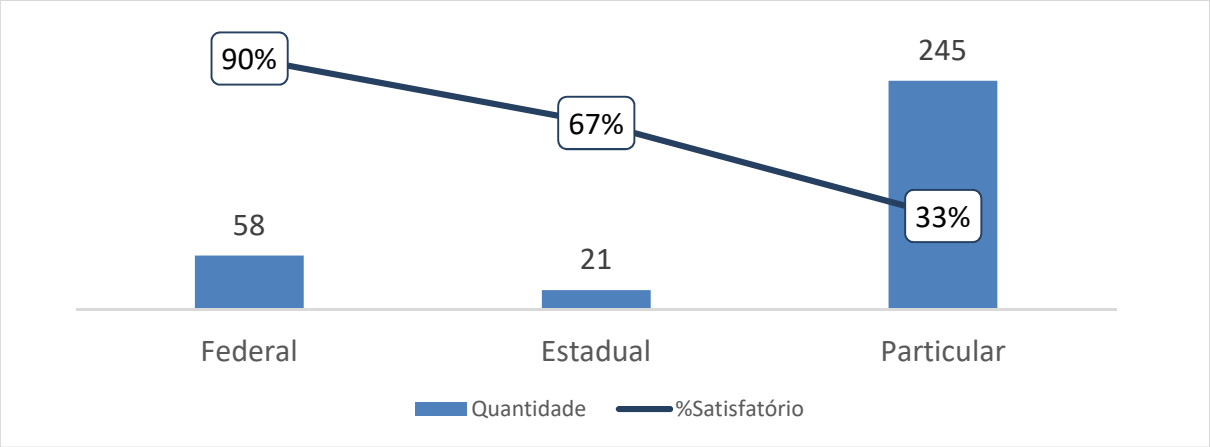
**Gráfico 1 - Percentual conceito satisfatório Enade no Brasil em 329 IES que ofertam curso de Engenharia de Produção**




**Fonte:** o próprio autor (2018)

Verificou-se que menos da metade dos cursos de Engenharia de Produção não atendem ao índice Satisfatório calculados pelo Inep. Desta forma, fez-se mais um questionamento: o índice de satisfação tem alguma relação direta com a categoria administrativa das IES? Obtém-se, então, o **Gráfico 2** para responder à esta dúvida.

**Gráfico 2 - Percentual conceito curso satisfatório x Quantidade cursos por categoria administrativa**

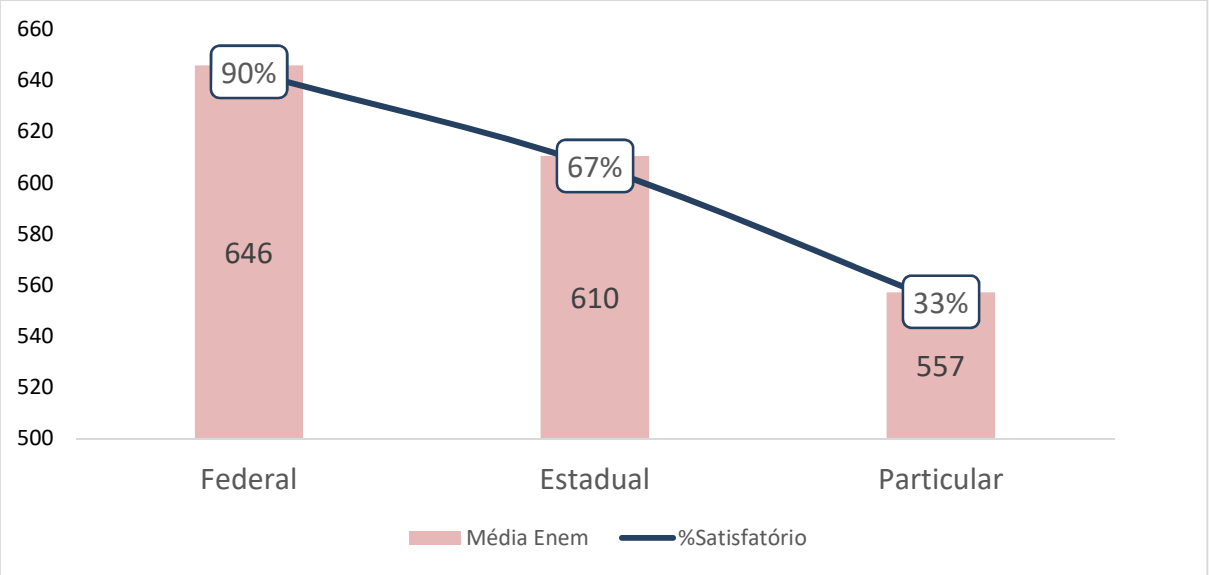



**Fonte:** o próprio autor (2018)

Analisando o **Gráfico 2**, verifica-se que, aparentemente, o problema do alto conceito insatisfatório dos cursos de Engenharia de Produção está relacionada à categoria administrativa Particular que corresponde à 75% dos cursos ofertados e apenas 33% deles atinge conceito satisfatório.

Para verificar se o resultado aparente é verdadeiro, fez-se um último questionamento: apenas a categoria administrativa influencia o conceito do curso, ou pode-se ter algum outro fator que faça com que esse conceito sofra alteração? No **Gráfico 3** foi realizada uma comparação entre a média do Enem dos alunos com o percentual de índice de conceito satisfatório para responder à esse questionamento.

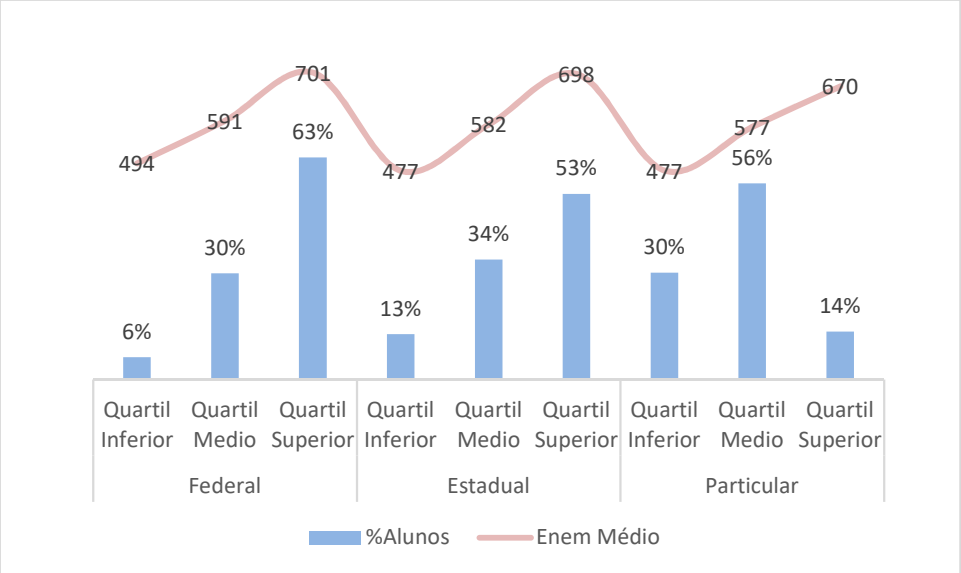
**Gráfico 3 - Percentual conceito curso x Média Enem dos alunos por categoria administrativa**




**Fonte:** o próprio autor (2018)

Observando o **Gráfico 3**, verifica-se que existe uma relação entre a média que os alunos obtém no Enem com o índice de conceito satisfatório no Enade. Entende-se que as IES federais possuem mais alunos com uma média alta no Enem e, por consequência, possuem maior índice de conceito satisfatório do que as IES particulares, pois estas possuem mais alunos com uma média inferior no Enem, conforme evidenciado no **Gráfico 4** abaixo.

**Gráfico 4 - Proporção de alunos por quartil Enem x Enem médio**



**Fonte:** o próprio autor (2018)

Devido à esta conclusão, entende-se que pode haver uma relação numérica que conseguirá prever qual será o conceito Enade dos alunos de acordo com sua nota no Enem. Segue-se então para o próximo tópico no qual será explicado algumas formas que foram elaboradas para realizar uma predição do conceito do Enade baseadas no Enem.

### 4.3 Conversão em Faixa Enade por aluno por Regressão Linear Simples

As tabelas a serem utilizadas na predição, conforme já exposto no tópico 4.2, são as tabelas dos microdados do IDD fornecidos pelo Inep referentes ao ano de 2014 e a base de conceito Enade por curso também fornecida pelo Inep. Essas tabelas foram inseridas no R para o início dos cálculos.

Todos os cursos diferentes de Engenharia de Produção (co_grupo=6208) foram removidos da base, visto que as análises desse trabalho serão realizadas apenas para o curso mencionado, conforme **Figura 2**.

**Figura 2 - Filtro do curso de Engenharia de Produção**

```
conceito <- filter(conceito, cod_area == 6208)
names(conceito)[15] <- c("nt_ger")
```

**Fonte:** o próprio autor (2018)

Após a inserção das tabelas, foram realizados alguns procedimentos para que a base ficasse o mais confiável possível para elaboração dos cálculos. Os alunos que não possuíam as 4 colunas referentes ao Enem preenchidas (enem_nt_cn, enem_nt_ch, enem_nt_lc, enem_nt_mt) foram removidos, pois poderiam enviesar os dados, tornando a análise menos confiável. Foram removidos, também, os alunos não concluintes e ausentes, conforme pode ser visto na **Figura 3**.

**Figura 3 - Filtro de alunos do curso de Engenharia de Produção, Concluintes, Presentes e com todas as notas do Enem na base de IDD**

```
base <- idd%>%
        filter(co_grupo == 6208 & tp_inscricao == 0 & tp_pres == 555 &
                 (is.na(enem_nt_cn)==FALSE & is.na(enem_nt_ch)==FALSE
                  & is.na(enem_nt_lc)==FALSE & is.na(enem_nt_mt)==FALSE))
```

**Fonte:** o próprio autor (2018)

Com o objetivo de conseguir iniciar os cálculos e análises, realizou-se o cálculo da média do Enem por aluno dentro da base de microdados utilizando as seguintes colunas: enem_nt_cn, enem_nt_ch, enem_nt_lc, enem_nt_mt e o valor obtido foi inserido na coluna enem_medio, conforme **Figura 4**.

**Figura 4 - Cálculo do Enem Médio**

```
base <- base%>%
  mutate(enem_medio = (enem_nt_cn+enem_nt_ch+enem_nt_lc+enem_nt_mt)/4)
```

**Fonte:** o próprio autor (2018)

Para iniciar a predição, é necessário obter a faixa do conceito Enade por aluno, pois na tabela de conceito Enade, existe essa faixa apenas para o curso na IES.

Para obter esse dado, foi realizado um teste de correlação baseado no coeficiente de Pearson, conforme **Figura 5** para garantir a confiabilidade dos dados. Este teste retornou um coeficiente de Pearson de 0,999466, o que significa que existe uma correlação positiva muito forte.

**Figura 5 - Teste de Correlação entre as Variáveis Nota Geral Bruta e Nota Geral Contínua (Faixa)**

```
cor(conceito$nt_ger, conceito$nota_ger_continuo)
```

**Fonte:** o próprio autor (2018)

Entendendo que este teste trouxe a confiabilidade que há correlação muito forte entre as variáveis de Nota Geral Bruta e Nota Geral Contínua (Faixa). Portanto, pode-se assumir que a utilização do processo de regressão linear para criar uma função que converta a Nota Geral Bruta na Nota Geral Contínua (Faixa) é factível e confiável estatisticamente.

Na **Figura 6** está o processo de criação da função da regressão linear que obteve a equação de conversão de Nota Geral Bruta para Nota Geral Contínua (Faixa) nomeada como converte_faixa.

**Figura 6 - Regressão Linear entre as Variáveis Nota Geral Bruta e Nota Geral Contínua (Faixa)**

```
converte_faixa = lm(nota_ger_continuo~nt_ger, data = conceito)
summary(converte_faixa)
```

**Fonte:** o próprio autor (2018)

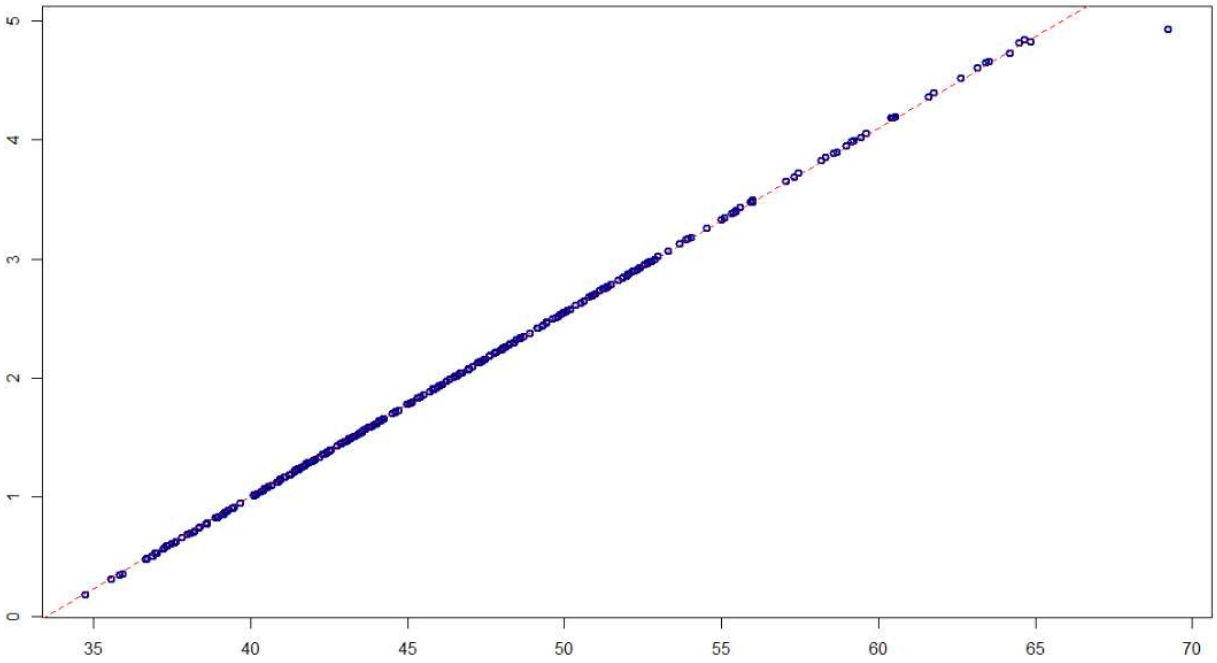
Na **Figura 7** estão os resultados das variáveis obtidas na regressão linear, indicando um $R^2$ (coeficiente de determinação) de 0,9989, o que significa que 99,89% da variável de dependente consegue ser explicada pelos regressores presentes no modelo. Para ilustrar, na **Figura 8** observa-se a representação gráfica da regressão.

**Figura 7 - Resultado do *summary(converte_faixa)***

```
Call:
lm(formula = nota_ger_continuo ~ nt_ger, data = conceito)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59270 -0.00350  0.00092  0.00629  0.02660

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.1770873  0.0131480  -393.8   <2e-16 ***
nt_ger       0.1545319  0.0002794   553.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03373 on 327 degrees of freedom
Multiple R-squared:  0.9989, Adjusted R-squared:  0.9989
F-statistic: 3.06e+05 on 1 and 327 DF,  p-value: < 2.2e-16
```

**Fonte:** o próprio autor (2018)

**Figura 8 - Representação Gráfica da Regressão**




**Fonte:** o próprio autor (2018)

Para realizar a classificação das IES de acordo com a categoria administrativa foi utilizada a **Figura 9** que foi obtida do Dicionário de Variáveis do Enade. Esta tabela foi inserida no R para que, conforme **Figura 10** e **Figura 11** vinculasse cada IES com sua categoria.

**Figura 9 - Classificação da Categoria Administrativa da IES**

| DICIONÁRIO DE VARIÁVEIS - ENADE 2014 | | | | |
|---|---|---|---|---|
| **Nome** | **Tipo** | **Tamanho** | **Descrição** | **Categorias** |
| co_catad | Numérica | 8 | Código da categoria administrativa da IES | 93 = Pessoa Jurídica de Direito Público - Federal<br>116 = Pessoa Jurídica de Direito Público - Municipal<br>118 = Pessoa Jurídica de Direito Privado - Com fins lucrativos - Sociedade Civil<br>121 = Pessoa Jurídica de Direito Privado - Sem fins lucrativos - Fundação<br>10001 = Pessoa Jurídica de Direito Público - Estadual<br>10002 = Pessoa Jurídica de Direito Público - Federal<br>10003 = Pessoa Jurídica de Direito Público - Municipal<br>10004 = Pessoa Jurídica de Direito Privado - Com fins lucrativos - Associação de Utilidade Pública<br>10005 = Privada com fins lucrativos<br>10006 = Pessoa Jurídica de Direito Privado - Com fins lucrativos - Sociedade Mercantil ou Comercial<br>10007 = Pessoa Jurídica de Direito Privado - Sem fins lucrativos - Associação de Utilidade Pública<br>10008 = Privada sem fins lucrativos<br>10009 = Pessoa Jurídica de Direito Privado - Sem fins lucrativos - Sociedade |

**Fonte:** Inep (2014)

**Figura 10 - Vinculação da Categoria Administrativa da IES na Base de Microdados IDD**

```
base <- base%>%
         left_join(tipo_ies, by = 'co_catad')
```

**Fonte:** o próprio autor (2018)

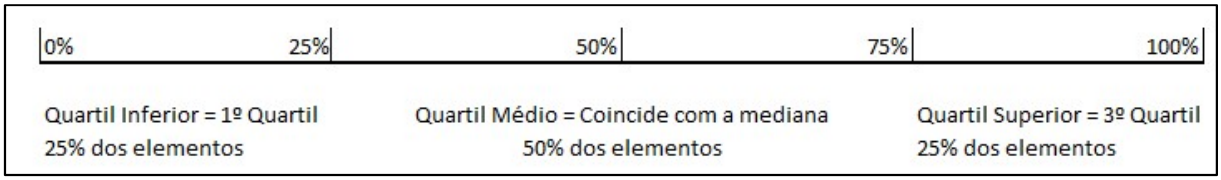
**Figura 11 - Classificação Simplificada da Categoria Administrativa da IES**

```
base <- base%>%
         mutate(perfil_ies =
            ifelse(str_detect(ds_catad,'Federal')==TRUE
                   ,'Federal',ifelse(str_detect(ds_catad,'Estadual')==TRUE
                                   | str_detect(ds_catad,'Municipal')==TRUE
                                                          ,'Estadual'
                                                          ,'Particular')))
```

**Fonte:** o próprio autor (2018)

Para simplificar análises futuras, foi feita uma simplificação da nota média no Enem, repartindo-se em 3 grupos de controle, sendo o primeiro as notas abaixo do primeiro quartil (quartil inferior), o segundo entre o primeiro e o terceiro quartil (quartil médio) e o terceiro, acima do terceiro quartil (quartil superior). Ilustra-se, de forma mais clara a explicação de quartis na **Figura 12** e a forma como os alunos foram classificados está na **Figura 13**.

**Figura 12 - Explicação Quartil**




**Fonte:** o próprio autor (2018)

**Figura 13 - Criação da Classificação dos Alunos por Quartil da Nota Média do Enem**

```
base <- base%>%
         mutate(quartil = ifelse(enem_medio<quantile(enem_medio,0.25)
                                 ,'Quartil Inferior'
                                 ,ifelse(enem_medio<quantile(enem_medio,0.75)
                                         ,'Quartil Medio'
                                         ,'Quartil Superior')))
```

**Fonte:** o próprio autor (2018)

Após a realização de todos os ajustes pontuados anteriormente, neste momento, conforme **Figura 14** é feita a criação da Nota Geral Contínua (Faixa) e sua vinculação à base de microdados do IDD utilizando a função criada na **Figura 6**.

**Figura 14 - Criação e Vinculação da Nota Geral Contínua (Faixa)**

```
fx_enade <- predict(converte_faixa, base%>%
                                      select(nt_ger))
base <- base%>%
         cbind(fx_enade)
```

**Fonte:** o próprio autor (2018)

A partir da criação e vinculação da Nota Geral Contínua (Faixa), cria-se, como pode ser visto na **Figura 15,** a coluna conceito_enade baseada na **Tabela 1**.

**Figura 15 - Criação da Coluna conceito_enade**

```
base <- base%>%
         mutate(conceito_enade =
                  ifelse(fx_enade<0.945,1
                  ,ifelse(fx_enade<1.945,2
                  ,ifelse(fx_enade<2.945,3
                  ,ifelse(fx_enade<3.945,4,5)))))
```

**Fonte:** o próprio autor (2018)

A partir da criação e vinculação da coluna conceito_enade, cria-se a coluna conceito, conforme **Figura 16**, que informará se o conceito é satisfatório ou insatisfatório, conforme citado no referencial teórico no tópico 2.1. Desta forma finaliza-se a conversão em faixa Enade por aluno por regressão linear simples a qual utilizaremos para elaborar os resultados posteriormente.

**Figura 16 - Criação da coluna conceito**

```
base <- base%>%
         mutate(conceito =
                  ifelse(conceito_enade<3
                         ,'Insatisfatorio'
                         ,'Satisfatorio'))
```

**Fonte:** o próprio autor (2018)

## 4.4 PREDIÇÃO DE FAIXA ENADE POR ALUNO POR REGRESSÃO LINEAR MÚLTIPLA

Para realizar a predição de faixa Enade por aluno por regressão linear múltipla, inicia-se criando a função de regressao_enem, conforme **Figura 17**.

**Figura 17 - Criação da Função regressao_enem**

```
regressao_enem = lm(nt_ger~enem_nt_lc+enem_nt_ch+enem_nt_cn+enem_nt_mt, data = base)
summary(regressao_enem)
```

**Fonte:** o próprio autor (2018)

**Figura 18 - Resultado do *summary(regressão_enem)***

```
Call:
lm(formula = nt_ger ~ enem_nt_lc + enem_nt_ch + enem_nt_cn +
    enem_nt_mt, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-65.941  -6.229   0.402   6.864  28.460

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.011010   1.121535 -10.709  < 2e-16 ***
enem_nt_lc    0.020836   0.002652   7.858 4.85e-15 ***
enem_nt_ch    0.025665   0.002692   9.534  < 2e-16 ***
enem_nt_cn    0.032110   0.002728  11.768  < 2e-16 ***
enem_nt_mt    0.023250   0.001768  13.148  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.22 on 4448 degrees of freedom
Multiple R-squared:  0.4008, Adjusted R-squared:  0.4003
F-statistic: 743.9 on 4 and 4448 DF,  p-value: < 2.2e-16
```

**Fonte:** o próprio autor (2018)

Após a função de regressão estar criada, é realizada uma simulação de qual deveria ser a Nota Geral Bruta baseada nas notas que os alunos obtiveram no Enem. Essa simulação está representada na **Figura 19**, e podemos ver na **Figura 18** que os valores das variáveis são estatisticamente confiáveis, pois são inferiores à 0,05 e um $R^2$ (coeficiente de determinação) de 0,4008, o que significa que 40,08% da variável de dependente consegue ser explicada pelos regressores presentes no modelo.

**Figura 19 - Regressão Linear Múltipla entre as notas Enem dos alunos e Nota Geral Bruta**

```
nt_ger_predict <- predict(regressao_enem, base%>%
                          select(enem_nt_lc,enem_nt_ch,enem_nt_cn,enem_nt_mt))
base <- base%>%
         cbind(nt_ger_predict)
```

**Fonte:** o próprio autor (2018)

A partir da Nota Geral Bruta simulada por aluno pela regressao_enem, utiliza-se novamente a função criada na **Figura 6**, converte_faixa, para simular qual seria a Nota Geral Contínua (Faixa) por aluno, conforme evidenciado na **Figura 20**, criando a coluna nomeada de fx_enade_predict.

**Figura 20 – Criação da fx_enade_predict**

```
nt_ger_predict<-as.data.frame(nt_ger_predict)
colnames(nt_ger_predict)<-'nt_ger'
fx_enade_predict <- predict(converte_faixa, nt_ger_predict)
base <- base%>%
         cbind(fx_enade_predict)
```

**Fonte:** o próprio autor (2018)

A partir da criação e vinculação da Nota Geral Contínua (Faixa) simulada, chamada fx_enade_predict conforme **Figura 20**, cria-se uma nova coluna, nomeada conceito_enade_predict, como pode ser visto na **Figura 21,** onde é classificado o resultado da fx_enade_predict pela regra exposta na **Tabela 1**.

**Figura 21 - Criação do conceito enade_predict**

```
base <- base%>%
         mutate(conceito_enade_predict =
                  ifelse(fx_enade_predict<0.945,1
                  ,ifelse(fx_enade_predict<1.945,2
                  ,ifelse(fx_enade_predict<2.945,3
                  ,ifelse(fx_enade_predict<3.945,4,5)))))
```

**Fonte:** o próprio autor (2018)

Por último cria-se a coluna, conceito_predict conforme **Figura 22**, onde classificaremos o conceito da Nota Geral Contínua simulada na coluna fx_enade_predict, em conceitos insatisfatório ou satisfatório, conforme explicado no

tópico 2.1, que utilizaremos para elaborar os resultados posteriormente. Dado essas atribuições, finalizamos o cálculo da predição de faixa Enade por aluno por regressão linear múltipla.

**Figura 22 - Criação do conceito_predict**

```
base <- base%>%
  mutate(conceito_predict =
           ifelse(conceito_enade_predict<3
                    ,'Insatisfatorio'
                    ,'Satisfatorio'))
```

**Fonte:** o próprio autor (2018)

## 4.5 PREDIÇÃO DE FAIXA ENADE POR ALUNO POR *MACHINE LEARNING* - ÁRVORE DE DECISÃO

Para realizar a predição de faixa Enade por aluno por *Machine learning* pelo método de Árvore de Decisão, inicia-se separando uma base de treino, que conterá 75% dos dados e uma base de testes, que conterá 25% dos dados, conforme pode ser observado na **Figura 23**.

**Figura 23 - Separação das bases de treino e teste**

```
amostra <- sample(2,nrow(base),replace = TRUE,prob = c(0.75,0.25))

treino <- base[amostra==1,]
teste <- base[amostra==2,]
```

**Fonte:** o próprio autor (2018)

Após a separação da base de dados, o modelo de predição será criado utilizando as variáveis das notas do Enem dos alunos para definir qual seria a Nota Bruta Contínua (Faixa). Na **Figura 24** está ilustrada a criação do modelo utilizando a base de treino e na **Figura 25** estão os parâmetros do modelo.

**Figura 24 - Criação do modelo de predição por Árvore de Decisão**

```
floresta <- randomForest(fx_enade ~ enem_nt_cn+enem_nt_ch+enem_nt_lc+enem_nt_mt
                                   ,data = treino, ntree=2000,importance=TRUE)
print(floresta)
```

**Fonte:** o próprio autor (2018)

**Figura 25 - Resultado do *summary(floresta)***

```
Call:
  randomForest(formula = fx_enade ~ enem_nt_cn + enem_nt_ch + enem_nt_lc + enem_nt_mt
                                    , data = treino, ntree = 2000, importance = TRUE)

                                               Type of random forest: regression
                                                     Number of trees: 2000
                                          No. of variables tried at each split: 1

                                                    Mean of squared residuals: 2.477909
                                                              % Var explained: 38.63
```

**Fonte:** o próprio autor (2018)

Com o modelo finalizado, ele é utilizado na base de teste para classificar o desempenho dos alunos no exame baseado no aprendizado na base de treino, criando a coluna previsao_teste na base de treino, conforme pode ser visto na **Figura 26**.

**Figura 26 - Predição do resultado dos alunos na base de teste**

```
previsao_teste <- predict(floresta, teste)
teste <- teste%>%
          cbind(previsao_teste)
```

**Fonte:** o próprio autor (2018)

A partir da criação e vinculação da Nota Geral Contínua (Faixa) simulada, chamada previsao_teste conforme **Figura 26**, cria-se uma nova coluna, nomeada conceito_enade_ml, como pode ser visto na **Figura 27**, onde é classificado o resultado da previsao_teste pela regra exposta na **Tabela 1**.

**Figura 27 - Criação do conceito_enade_ml**

```
teste <- teste%>%
          mutate(conceito_enade_ml =
                      ifelse(previsao_teste<0.945,1
                     ,ifelse(previsao_teste<1.945,2
                     ,ifelse(previsao_teste<2.945,3
                     ,ifelse(previsao_teste<3.945,4,5)))))
```

**Fonte:** o próprio autor (2018)

Para finalizar cria-se a coluna, conceito_ml conforme **Figura 28**, onde classificaremos o conceito da Nota Geral Contínua simulada na coluna previsao_teste,

em conceitos insatisfatório ou satisfatório, conforme explicado no tópico 2.1. Dado essas atribuições, finalizamos o cálculo da predição de faixa Enade por aluno por *machine learning*.

**Figura 28 - Criação do conceito_ml**

```
teste <- teste%>%
  mutate(conceito_ml =
           ifelse(conceito_enade_ml<3
                  ,'Insatisfatorio'
                  ,'Satisfatorio'))
```

**Fonte:** o próprio autor (2018)

Com objetivo de haver uma comparação entre o modelo de predição pela regressão linear múltipla e pelo *machine learning* é necessário que seja utilizado o modelo feito com a base de treino por *machine learning*, porém replicado na base completa de dados, e não apenas na base repartida em base de teste e base de treino.

Para realizar a predição de faixa Enade por aluno por *Machine learning* pelo método de Árvore de Decisão no modelo super ajustado, ele é utilizado na base completa de dados para classificar o desempenho dos alunos no exame baseado no aprendizado na base de treino, criando a coluna previsao_super_ajustada, conforme pode ser visto na **Figura 29**.

**Figura 29 - Predição do resultado dos alunos na base completa de dados**

```
previsao_super_ajustada <- predict(floresta, base)
```

**Fonte:** o próprio autor (2018)

A partir da criação e vinculação da Nota Geral Contínua (Faixa) simulada, chamada previsao_super_ajustada conforme **Figura 29**, cria-se uma nova coluna, nomeada conceito_enade_super_ajustado, como pode ser visto na **Figura 30**, onde é classificado o resultado da previsão_super_ajustada pela regra exposta na **Tabela 1**.

**Figura 30 - Criação do conceito_enade_ super_ajustado**

```
base <- base%>%
         cbind(previsao_super_ajustada)%>%
              mutate(conceito_enade_super_ajustado =
                                ifelse(previsao_super_ajustada<0.945,1
                               ,ifelse(previsao_super_ajustada<1.945,2
                               ,ifelse(previsao_super_ajustada<2.945,3
                               ,ifelse(previsao_super_ajustada<3.945,4,5)))))
```

**Fonte:** o próprio autor (2018)

Para finalizar cria-se a coluna, conceito_super_ajustado conforme **Figura 31**, onde classificaremos o conceito da Nota Geral Contínua simulada na coluna previsao_super_ajustada, em conceitos insatisfatório ou satisfatório, conforme explicado no tópico 2.1. Dado essas atribuições, finalizamos o cálculo da predição de faixa Enade por aluno por *machine learning* no modelo super ajustado.

**Figura 31 - Criação do conceito_super_ajustado**

```
base <- base%>%
          mutate(conceito_super_ajustado =
                    ifelse(conceito_enade_super_ajustado<3
                          ,'Insatisfatorio'
                          ,'Satisfatorio'))
```

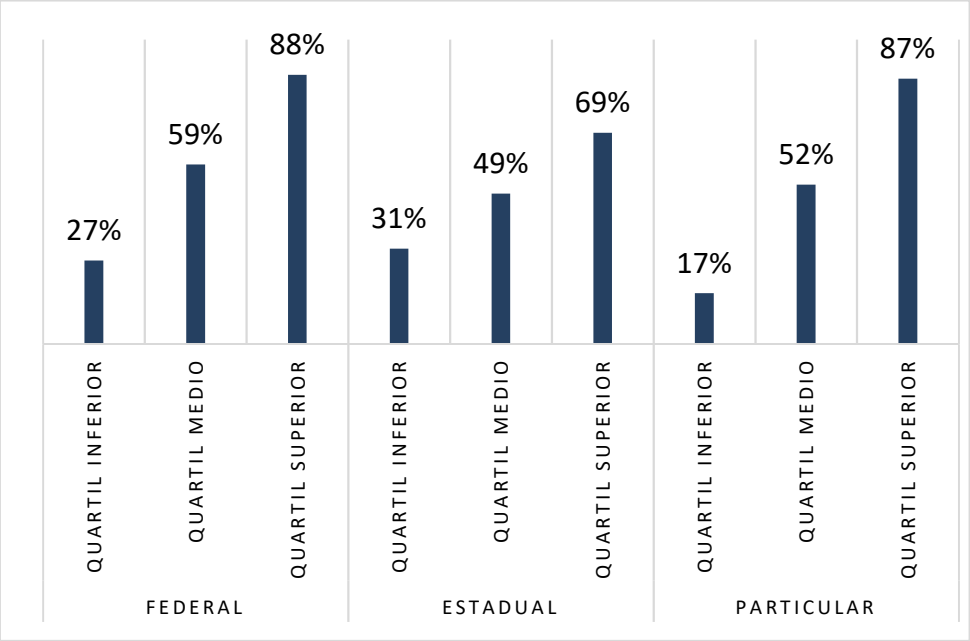
**Fonte:** o próprio autor (2018)

## 5. RESULTADOS

Em relação ao último questionamento do tópico 4.2: “apenas a categoria administrativa influencia o conceito do curso, ou pode-se ter algum outro fator que faça com que esse conceito sofra alteração?”, verificou-se que poderia haver uma relação entre a categoria administrativa da IES e o resultado satisfatório, como pode ser visto no **Gráfico 3**, mas também, poderia haver uma relação entre o perfil do egresso indicado pelo desempenho no Enem e o resultado satisfatório, conforme mostra o **Gráfico 4**.

Para que seja possível comprovar a veracidade dessas relações, foi elaborada a conversão do desempenho do aluno no exame Enade em conceitos de insatisfatório ou satisfatório, conforme demonstrado no tópico 4.3. A partir dessa conversão realizou-se uma análise do desempenho do aluno no exame Enade agora classificado pela categoria administrativa juntamente com o desempenho no Enem, conforme pode ser visto no **Gráfico 5**.

**Gráfico 5 - Desempenho no Enade classificado em categoria administrativa e performance no Enem**




**Fonte:** o próprio autor (2018)

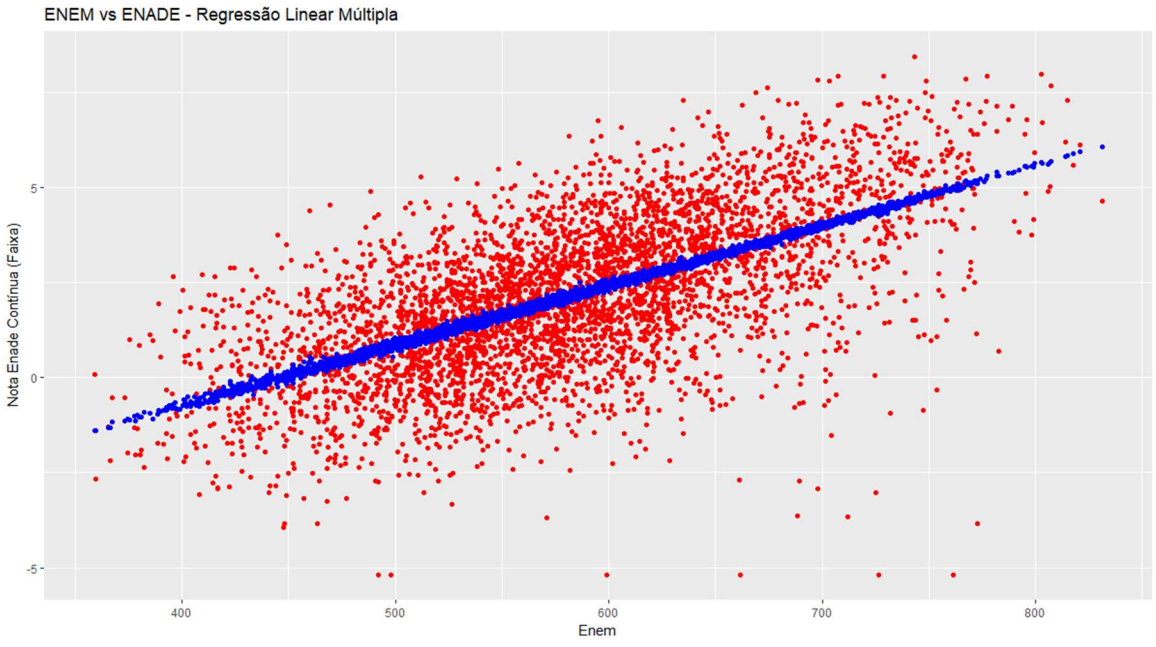
Analisando o **Gráfico 5** verifica-se que a variável mais forte para entender o resultado de um curso no Enade é o perfil do egresso, e não a categoria administrativa, pois, conforme evidenciado, a proporção de conceito satisfatório nas

três categorias segue a mesma tendência baseando-se no desempenho do Enem, sendo a maior diferença de performance final de uma IES federal e de uma IES particular a proporção dos alunos em cada quartil do Enem, conforme evidenciado no **Gráfico 4**.

Entendendo que a variável mais forte para chegar ao resultado de um curso no Enade é o perfil do egresso, faz-se o seguinte questionamento: seria possível prever o resultado de um aluno no exame Enade baseando-se em suas notas no exame Enem? Para responder à este questionamento, elabora-se então duas diferentes formas de predição da nota do desempenho conceitual no Enade com base nas notas do aluno no Enem, sendo a primeira delas a predição por regressão linear múltipla e a segunda, a predição por *machine learning*, no modelo de Árvore de Decisão.

No **Gráfico 6** é possível verificar o comportamento da regressão criada no tópico 4.4 para chegar aos resultados de predição da nota do desempenho no Enade com base nas notas do aluno no Enem. Como esse é o modelo feito por regressão linear múltipla, observa-se, no **Gráfico 6**, um comportamento linear da regressão, não sendo possível absorver o comportamento real do desempenho da amostra, porém verifica-se linearidade proporcional entre o desempenho no Enem e o desempenho no Enade.

**Gráfico 6 - Comportamento da regressão linear múltipla x Comportamento real dos desempenhos de Enem x Enade**



**Fonte:** o próprio autor (2018)

Classificando a nota gerada pela regressão por conceito Enade, compara-se, pela matriz de confusão, a classificação geral e a simulada, obtendo-se uma assertividade de 72,87% conforme evidanciado na **Figura 32**.

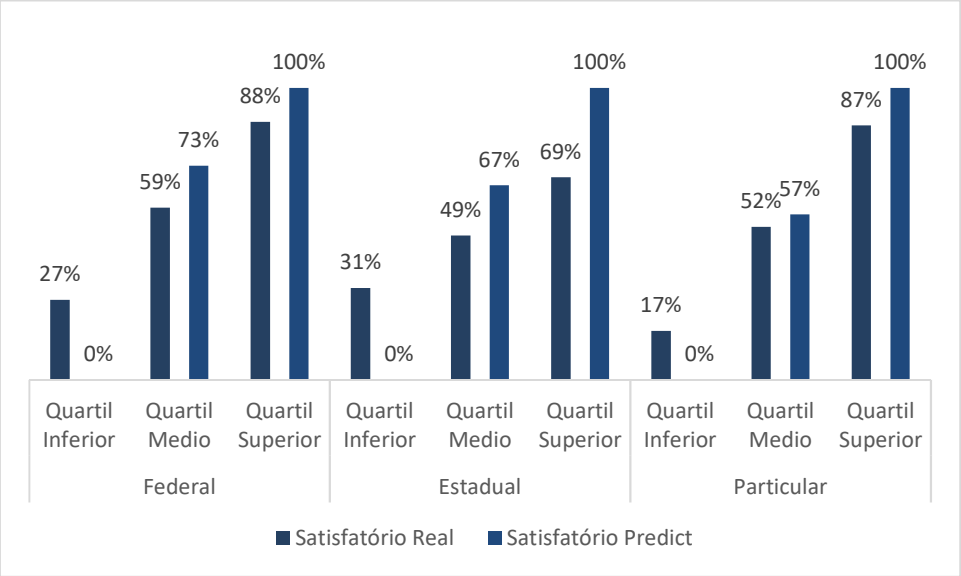
**Figura 32 - Matriz de confusão entre resultado da regressão e resultado real**

```
                Insatisfatorio Satisfatorio
  Insatisfatorio           1472          652
  Satisfatorio              556         1773
> (confusao[1]+confusao[4]) / sum(confusao)
[1] 0.7287222
```

**Fonte:** o próprio autor (2018)

No **Gráfico 7** pode ser verificada a comparação entre as porcentagens de conceito satisfatório real e satisfatório baseado na classificação pela regressão linear múltipla. A linearidade dos resultados é evidenciada quando observa-se que alunos de quartil inferior não atingiriam resultado satisfatório e todos alunos de quartil superior deveriam atingir resultado satisfatório.

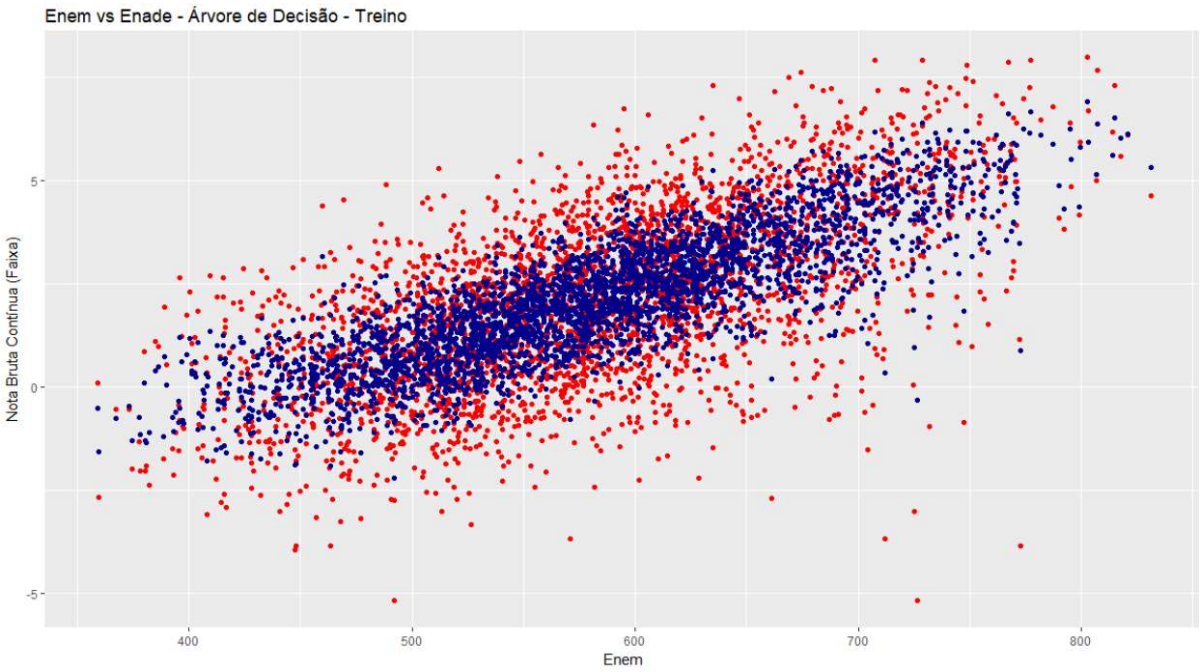
**Gráfico 7 - Desempenho no Enade classificado em categoria administrativa e performance no Enem com base nos resultados reais e nos resultados obtidos da classificação por regressão linear múltipla**




**Fonte:** o próprio autor (2018)

Conforme evidenciado nos **Gráfico 6** e **Gráfico 7**, nota-se que a regressão linear multipla não é suficinete para predizer o comportamento do resultado do aluno no Enade, pois existe uma dispersão entre os desempenhos comparados. Assim, conforme explicado no tópico 4.5, criou-se de um modelo que busca absorver o comportamento da amostra, utilizando uma metodologia de *machine learning* denominada Árvore de Decisão. No **Gráfico 8** é possível verificar o comportamento adquirido pelo modelo após a fase de treino exemplificada na **Figura 24**.

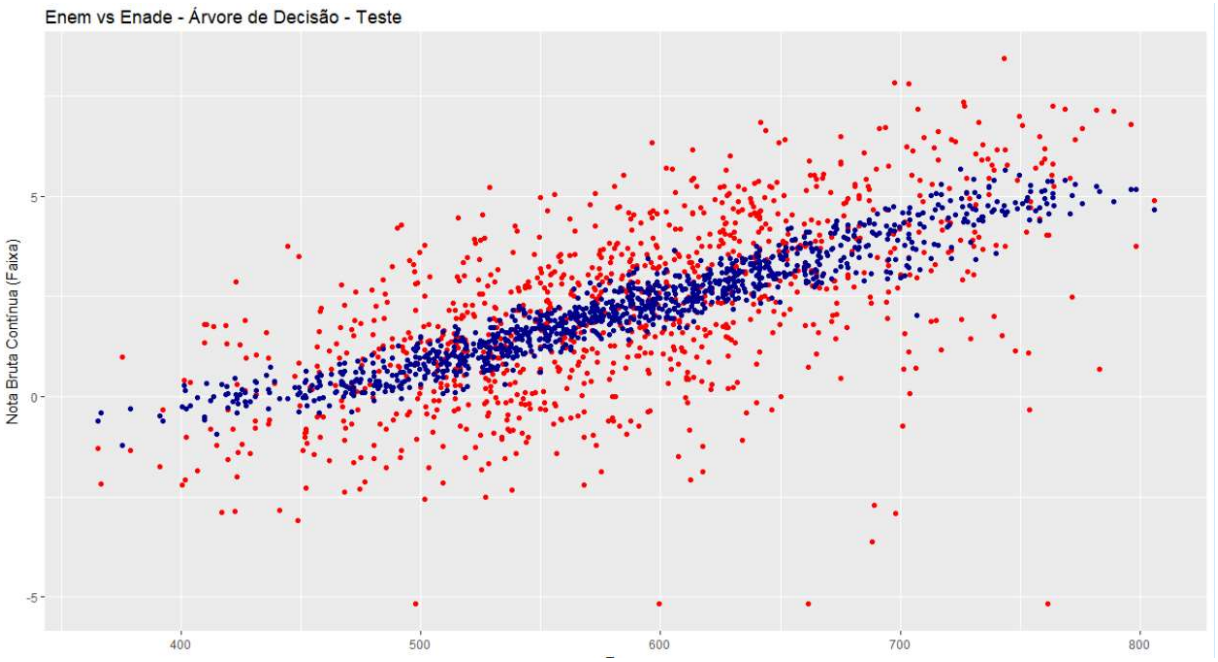
**Gráfico 8 - Comportamento treino da Árvore de Decisão x Comportamento real dos desempenhos de Enem x Enade na base treino**




**Fonte:** o próprio autor (2018)

No **Gráfico 9** é demonstrado o comportamento do modelo criado aplicado na base de teste conforme demostrado na **Figura 26**.

**Gráfico 9 - Comportamento teste da Árvore de Decisão x Comportamento real dos desempenhos de Enem x Enade na base teste**




**Fonte:** o próprio autor (2018)

Classificando a nota gerada pelo modelo aplicado na base de teste por conceito Enade, compara-se, pela matriz de confusão, a classificação geral e a simulada, obtendo-se uma assertividade de 73,02% conforme evidenciado na **Figura 33**.

**Figura 33 - Matriz de confusão entre resultado do modelo aplicado na base teste e resultado real na base teste**

```
                Insatisfatorio Satisfatorio
  Insatisfatorio            338          130
  Satisfatorio              156          436
> (confusao[1]+confusao[4]) / sum(confusao)
[1] 0.7301887
```

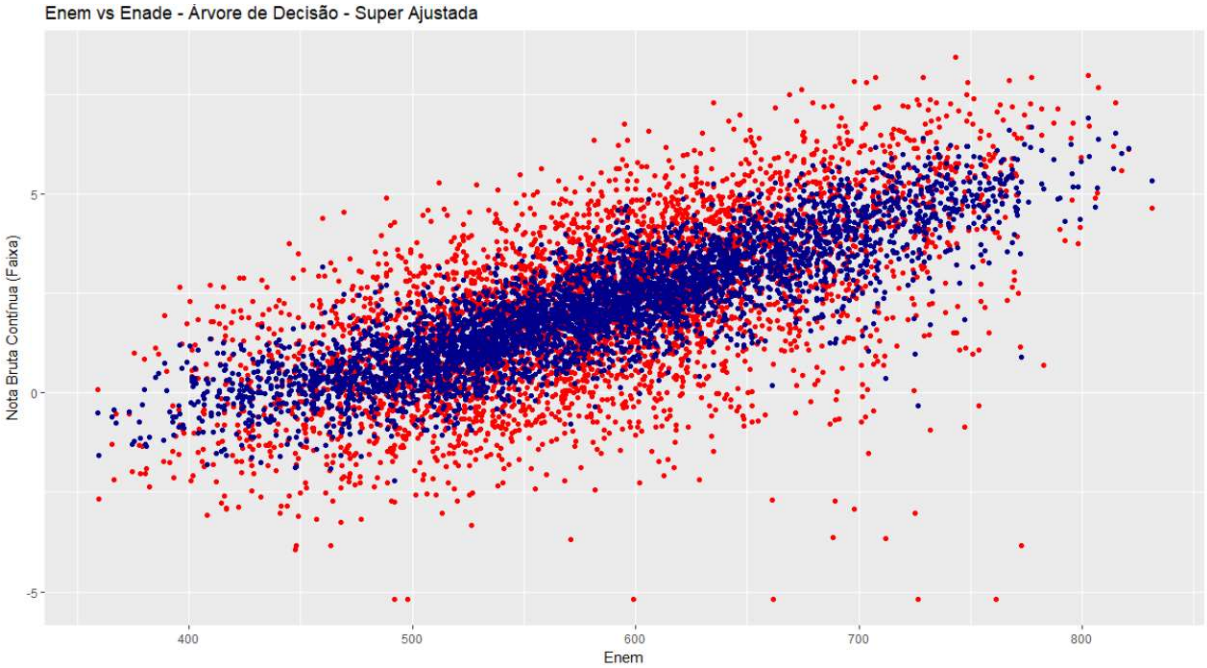
**Fonte:** o próprio autor (2018)

Ao comparar a **Figura 33** com a **Figura 32**, percebe-se que não há grande diferença na assetividade dos métodos, gerando apenas 0,15% de assertividade a mais no modelo de Árvore de Decisão, no entanto, apesar de não ter grande influência no resultando final, percebe-se, comparando o **Gráfico 9** com o **Gráfico 6** que o comportamento do modelo de Árvore de Decisão absorve mais o comportamento real do que do modelo de regressão linear múltipla.

Para fins de comparação aplicou-se, na base completa de dados, o modelo criado na base treino, criando um comportamento chamado de super ajustado.

Conforme esperado, oberva-se no **Gráfico 10** que a assertividade, neste método, é superior ao método da base teste, bem como da regressão linear múltipla, uma vez que 75% dos dados foi utilizado para treino do modelo.

**Gráfico 10 - Comportamento teste da Árvore de Decisão x Comportamento real dos desempenhos de Enem x Enade na base completa de dados**




**Fonte:** o próprio autor (2018)

Classificando a nota gerada pelo modelo super ajustado aplicado na base completa de dados por conceito Enade, compara-se, pela matriz de confusão, a classificação geral e a simulada, obtendo-se uma assertividade de 85,60% conforme evidanciado na **Figura 34**Figura 33.

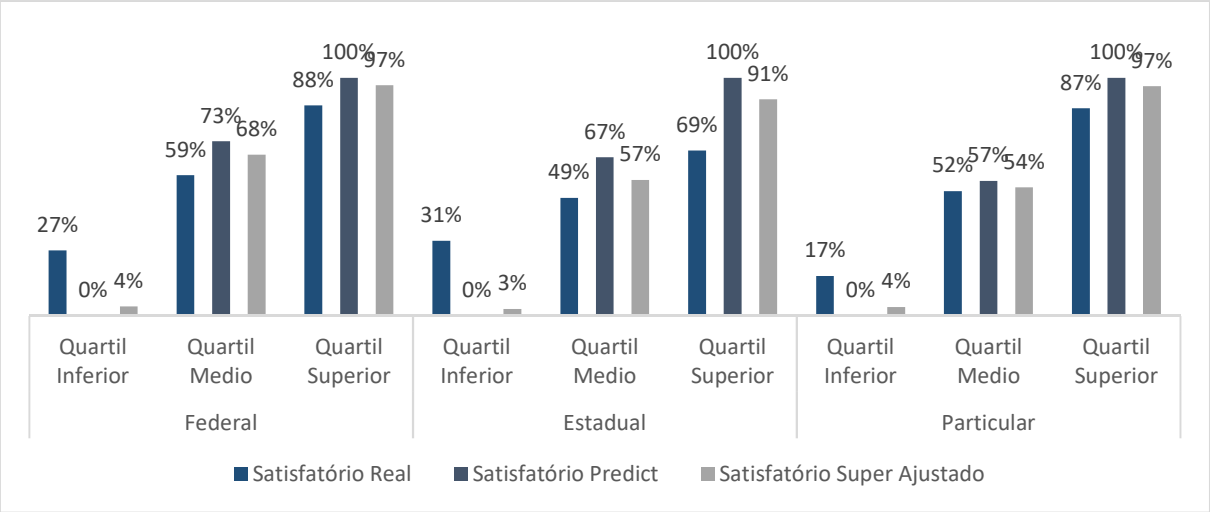
**Figura 34 - Matriz de confusão entre resultado do modelo aplicado na base completa de dados e resultado real na base completa de dados**

```
                Insatisfatorio Satisfatorio
  Insatisfatorio           1794          330
  Satisfatorio              311         2018
> (confusao[1]+confusao[4]) / sum(confusao)
[1] 0.8560521
```

**Fonte:** o próprio autor (2018)

Ao comparar os resultados da **Figura 33** com a **Figura 34**, verifica-se uma assertividade maior, conforme esperado, uma vez que 75% dos dados foi utilizado para treino do modelo. No entanto, realizou-se este teste para exemplificar no **Gráfico 11** qual a diferença que há na prática do desempenho dos dois modelos.

**Gráfico 11 - Desempenho no Enade classificado em categoria administrativa e performance no Enem com base nos resultados reais e nos resultados obtidos da classificação por regressão linear múltipla e pela Árvore de Decisão no modelo super ajustado**




**Fonte:** o próprio autor (2018)

O objetivo dessa pesquisa era estabelecer uma métrica avaliativa do desempenho acadêmico pelo potencial do estudante, de acordo com o Enem, e a classificação Enade dos cursos de Engenharia de Produção, tentando responder a seguinte problemática: o conceito Enade traz um real significado relacionado à avaliação do curso ou esse cálculo reflete outros desempenhos acadêmicos não ligados diretamente a qualidade do curso?

Para contextualizar sobre o cenário do curso de Engenharia de Produção no Brasil, buscou-se realizar uma análise exploratória, disposta no tópico 4.2, que revelou, que em 2014, apenas 47% dos cursos em ofertas tiveram conceito satisfatório pelo Enade. Além disso, 3 em cada 4 cursos são ofertados em IES de categoria administrativa particular, que, no total, tem apenas 33% dos cursos de Engenharia de Produção com conceito satisfatório.

A princípio, a conclusão é que, a categoria administrativa de uma IES impacta na qualidade do curso. No entanto, buscou-se confrontar essa visão, no

intuito de identificar se há outras variáveis que podem ser identificadas nos dados coletados, que também poderiam impactar esse índice.

Entende-se, com base nos dados dispostos no **Gráfico 3,** que existe uma relação de satisfação dos cursos com o perfil do egresso, sendo classificado pelo Enem médio, e, analisando o resultado em conjunto com os dados no **Gráfico 4,** que a proporção dos alunos por quartil classificados por Enem médio seria o responsável pela satisfação do curso.

Partindo desse pressuposto, buscou-se aprofundar essa visão, tentando entender se o desempenho do aluno no Enem teria relação com o desempenho do aluno no Enade.

Portanto, no tópico 4.3, busca-se criar uma função para conversão da nota geral bruta do aluno no Enade para uma nota geral contínua (Faixa), baseado na regressão dos resultados das notas dos cursos de Engenharia de Produção.

A partir dessa regressão, criou-se o desempenho do aluno por nota geral contínua (Faixa) e, posteriormente, classificação conceitual, para analisar o comportamento dos alunos por categoria administrativa, quartil Enem e conceito satisfatório no Enade, evidenciado no **Gráfico 5.**

Fica evidente que, entre categoria administrativa e Enem, a variável mais relevante para o desempenho no Enade de um aluno é o Enem, complementando ainda, que alunos de categoria administrativa particular tem o desempenho muito semelhante que a categoria administrativa federal, sendo inclusive, superior ao de categoria administrativa estadual.

Assim, conclui-se que a proporção de alunos por quartil do Enem é mais mandatório para o conceito no Enade do que a categoria administrativa.

No entanto, sendo as notas do Enem variáveis fortes para classificar o conceito de um curso no Enade, criou-se dois modelos de predição baseado nas notas Enem dos alunos, sendo o primeiro por regressão linear múltipla, descrito no tópico 4.4, e o segundo por *machine learning* pelo método de árvore de decisão, descrito no tópico 4.5.

Comparando os resultados dos modelos propostos, evidenciou-se que a assertividade dos dois modelos é praticamente a mesma, sendo o modelo por *machine learning* 0,15% mais assertivo que o modelo por regressão linear múltipla, conforme evidenciado na comparação entre a **Figura 33** e a **Figura 32**.

No entanto, apesar de não ter grande influência no resultando final, comparando o **Gráfico 9** com o **Gráfico 6**, observa-se que o modelo de *machine learning* absorve grande parte do comportamento de dispersão do desempenho dos alunos Enade e desempenho Enem, enquanto que a regressão linear múltipla absorve somente a tendência linear do proporção positivia do desempenho no Enade e no Enem.

Ainda realizou-se a aplicação do modelo por *machine learning* na base completa de dados, alcançando a assertividade 85,60%, que, por ser um modelo super ajustado, esperava-se uma assertividade maior, mas que foi realizada com a finalidade de comparar os três resultados, ou seja, real, por regressão liner múltipla e por modelo de *machine learning*.

## 6. CONCLUSÃO

Conclui-se que o Enem é uma variavel extremamente forte para se obter um resultado satisfatório no Enade, e ainda, que seria possível manipular o resultado conceitual de um curso, controlando a proporção de alunos egressos pelo quartil Enem. Fica em aberto o quão assertivo poderia ser o modelo de *machine learning* proposto, se houvesse maior volume de dados, e ainda, se fosse possível ter fatores sociais semelhantes, como o engajamento nos dois exame, pois espera-se um engajamento maior no Enem, devido ao resultado para o aluno ser a porta para entrar em grande parte das universidades, contra o Enade, que não tem grande influência em sua carreira, seja acadêmica ou profissional.

A partir da conclusõe citada e buscando atingir o objetivo dessa pesquisa, que é, estabelecer uma métrica avaliativa do desempenho acadêmico pelo potencial do estudante de acordo com o Enem, e a classificação Enade dos cursos de Engenharia de Produção, sugere-se então, que o indicador de qualidade seja um índice baseado no modelo de predição por regressão linear apresentado nesse trabalho. As IES deveriam ser pontuadas por conseguirem manter seus alunos dentro do desempenho esperado ou superior, e penalizadas por não conseguirem manter o o desempenho do aluno, sendo uma pontuação ponderada de forma descendente para acima do esperado, dentro do esperado e abaixo do esperado.

## 7. BIBLIOGRAFIA


CAMPOS, R. **Árvores de Decisão**. Disponível em: <https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69>. Acesso em 30 de abr. 2018.

DECONTO, E. **Entenda como funciona o conceito Enade de Avaliação.** Disponível em: <http://portal.eusoufamecos.net/entenda-como-funciona-o-conceito-Enade-de-avaliacao/>. Acesso em 02 de nov. 2017.

GERHARDT, T. E., SILVEIRA, D. T. **Métodos de pesquisa.** Coordenado pela Universidade Aberta do Brasil - UAB/UFRGS e pelo Curso de Graduação Tecnológica - Planejamento e Gestão para o Desenvolvimento Rural da SEAD/UFRGS. Porto Alegre: Editora da UFRGS, 2009.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP), **Enade**. 2015a. Disponível em: <http://portal.inep.gov.br/Enade>. Acesso em: 30 de abril. 2018.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP), **Indicador de Diferença entre os Desempenhos Observado e Esperado (IDD)**. 2017. Disponível em: <http://portal.inep.gov.br/educacao-superior/indicadores-de-qualidade/indicador-de-diferenca-entre-os-desempenhos-observado-e-esperado-idd>. Acesso em: 02 de nov. 2017.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP) C, **Índice Geral De Cursos (IGC)**. 2015b. Disponível em: <http://portal.inep.gov.br/indice-geral-de-cursos-igc->. Acesso em: 02 de nov. 2017.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). **Nota Técnica Daes/Inep Nº 57/2015.** 2018. Disponível em: < http://portal.inep.gov.br/documentos-e-legislacao12>. Acesso em 02 de nov. 2017.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP), **Perguntas Frequentes**. 2015c. Disponível em: < http://portal.inep.gov.br/perguntas-frequentes4>. Acesso em: 02 de nov. 2017.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP), **Sinaes.** 2015d. Disponível em: <http://portal.inep.gov.br/sinaes>. Acesso em: 30 de abril. 2018.

MATOS, D. **Conceitos fundamentais de *machine learning*.** Disponível em: <http://www.cienciaedados.com/conceitos-fundamentais-de-machine-learning/> Acesso em: 30 de abril. 2018.

MINISTÉRIO DA EDUCAÇÃO (MEC), **Avaliações da aprendizagem**. Dica de Leitura. Disponível em:
<http://portal.mec.gov.br/educacao-quilombola-/190-secretarias-112877938/setec-1749372213/18843-avaliacoes-da-aprendizagem>. Acesso em: 02 de nov. 2017.

MONACO, J. **10 algoritmos de *machine learning* que você precisa conhecer**. Disponível em:
<http://www.semantix.com.br/10-algoritmos-de-machine-learning/>. Acesso em: 30 de abr. 2018.

PETENATE, M. **Regressão Linear simples e múltipla, entenda as diferenças!** Disponível em: <https://www.escolaedti.com.br/regressao-linear-e-multipla-entenda-as-diferencas>. Acesso em: 21 de jun. 2018.

PICHILIANI, M. **Data mining na prática: árvores de decisão**. Disponível em: <https://imasters.com.br/artigo/5130/sql_server/data_mining_na_pratica_arvores_de_decisao>. Acesso em: 30 de abr. 2018.

# 8. ANEXOS

## 8.1 ANEXO A – DICIONÁRIO DE VARIÁVEIS IDD

| DICIONÁRIO DE VARIÁVEIS - ENADE 2014 | | | | |
|---|---|---|---|---|
| **Nome** | **Tipo** | **Tamanho** | **Descrição** | **Categorias** |
| nu_ano | Numérica | 8 | Ano de realização do exame | - |
| co_grupo | Numérica | 8 | Código da Área de enquadramento do curso no Enade | 21 = ARQUITETURA E URBANISMO<br>72 = TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS<br>73 = TECNOLOGIA EM AUTOMAÇÃO INDUSTRIAL<br>76 = TECNOLOGIA EM GESTÃO DA PRODUÇÃO INDUSTRIAL<br>79 = TECNOLOGIA EM REDES DE COMPUTADORES<br>701 = MATEMÁTICA (BACHARELADO)<br>702 = MATEMÁTICA (LICENCIATURA)<br>903 = LETRAS-PORTUGUÊS (BACHARELADO)<br>904 = LETRAS-PORTUGUÊS (LICENCIATURA)<br>905 = LETRAS-PORTUGUÊS E INGLÊS (LICENCIATURA)<br>906 = LETRAS-PORTUGUÊS E ESPANHOL (LICENCIATURA)<br>1401 = FÍSICA (BACHARELADO)<br>1402 = FÍSICA (LICENCIATURA)<br>1501 = QUÍMICA (BACHARELADO)<br>1502 = QUÍMICA (LICENCIATURA)<br>1601 = CIÊNCIAS BIOLÓGICAS (BACHARELADO)<br>1602 = CIÊNCIAS BIOLÓGICAS (LICENCIATURA)<br>2001 = PEDAGOGIA (LICENCIATURA)<br>2401 = HISTÓRIA (BACHARELADO)<br>2402 = HISTÓRIA (LICENCIATURA)<br>2501 = ARTES VISUAIS (LICENCIATURA)<br>3001 = GEOGRAFIA (BACHARELADO)<br>3002 = GEOGRAFIA (LICENCIATURA)<br>3201 = FILOSOFIA (BACHARELADO)<br>3202 = FILOSOFIA (LICENCIATURA)<br>3502 = EDUCAÇÃO FÍSICA (LICENCIATURA)<br>4004 = CIÊNCIA DA COMPUTAÇÃO (BACHARELADO)<br>4005 = CIÊNCIA DA COMPUTAÇÃO (LICENCIATURA) |
| co_ies | Numérica | 8 | Código da IES (e-Mec) | - |
| co_catad | Numérica | 8 | Código da categoria administrativa da IES | 93 = Pessoa Jurídica de Direito Público - Federal<br>116 = Pessoa Jurídica de Direito Público - Municipal<br>118 = Pessoa Jurídica de Direito Privado - Com fins lucrativos - Sociedade Civil<br>121 = Pessoa Jurídica de Direito Privado - Sem fins lucrativos - Fundação<br>10001 = Pessoa Jurídica de Direito Público - Estadual<br>10002 = Pessoa Jurídica de Direito Público - Federal<br>10003 = Pessoa Jurídica de Direito Público - Municipal<br>10004 = Pessoa Jurídica de Direito Privado - Com fins lucrativos - Associação de Utilidade Pública<br>10005 = Privada com fins lucrativos<br>10006 = Pessoa Jurídica de Direito Privado - Com fins lucrativos - Sociedade Mercantil ou Comercial<br>10007 = Pessoa Jurídica de Direito Privado - Sem fins lucrativos - Associação de Utilidade Pública<br>10008 = Privada sem fins lucrativos<br>10009 = Pessoa Jurídica de Direito Privado - Sem fins lucrativos - Sociedade |
| co_orgac | Numérica | 8 | Código da organização acadêmica da IES | 10019 = Centro Federal de Educação Tecnológica<br>10020 = Centro Universitário<br>10022 = Faculdade<br>10026 = Instituto Federal de Educação, Ciência e Tecnologia<br>10028 = Universidade |
| co_munic_curso | Numérica | 8 | Código do município de funcionamento do curso | Ver tabela de MUNICÍPIOS |

| | | | | |
|---|---|---|---|---|
| co_uf_curso | Numérica | 8 | Código da UF de funcionamento do curso | 11 = Rondônia (RO)<br>12 = Acre (AC)<br>13 = Amazonas (AM)<br>14 = Roraima (RR)<br>15 = Pará (PA)<br>16 = Amapa (AP)<br>17 = Tocantins (TO)<br>21 = Maranhão (MA)<br>22 = Piauí (PI)<br>23 = Ceará (CE)<br>24 = Rio grande do norte (RN)<br>25 = Paraíba (PB)<br>26 = Pernambuco (PE)<br>27 = Alagoas (AL)<br>28 = Sergipe (SE)<br>29 = Bahia (BA)<br>31 = Minas gerais (MG)<br>32 = Espírito santo (ES)<br>33 = Rio de janeiro (RJ)<br>35 = São paulo (SP)<br>41 = Paraná (PR)<br>42 = Santa catarina (SC)<br>43 = Rio grande do sul (RS)<br>50 = Mato grosso do sul (MS)<br>51 = Mato grosso (MT)<br>52 = Goiás (GO)<br>53 = Distrito federal (DF) |
| co_regiao_curso | Numérica | 8 | Código da região de funcionamento do curso | 1 = Norte<br>2 = Nordeste<br>3 = Sudeste<br>4 = Sul<br>5 = Centro-Oeste |
| co_curso | Numérica | 4 | Código do curso no Enade | |
| ano_in_grad | Numérica | 8 | Ano de início da graduação | - |
| tp_pres | Numérica | 4 | Tipo de presença | 111 = Aluno não selecionado<br>222 = Ausente<br>334 = Aluno fora do cadastro e implantado sem liminar (participação indevida)<br>555 = Presente<br>556 = Presente com resultado desconsiderado devido a problemas administrativos<br>999 = Aluno fora do cadastro e implantado com liminar |
| nt_ger | Numérica | 8 | Nota bruta da prova - Média ponderada da formação geral (25%) e componente específico (75%) (0 a 100) | - |
| tp_inscricao | Numérica | 8 | Indicador de concluinte / ingressante | 0 = Concluinte<br>1 = Ingressante |
| id_enem | Numérica | 8 | Identificação do ano de Enem selecionado | - |
| enem_nt_cn | Numérica | 3 | Nota da prova de Ciências da Natureza* | - |
| enem_nt_ch | Numérica | 3 | Nota da prova de Ciências Humanas* | - |
| enem_nt_lc | Numérica | 3 | Nota da prova de Linguagens e Códigos* | - |
| enem_nt_mt | Numérica | 3 | Nota da prova de Matemática* | - |

## 8.2 ANEXO B – DICIONÁRIO DE VARIÁVEIS CONCEITO ENADE

| Coluna | Descrição |
|---|---|
| **Ano** | Ano de realização do exame |
| **Cód Area** | Código da Área de enquadramento do curso no Enade |
| **Área** | Descrição da Área de enquadramento do curso no Enade |
| **Cód. IES** | Código da IES (e-Mec) |
| **Nome da IES** | Descrição da IES (e-Mec) |
| **Cod. Município** | Código do município de funcionamento do curso |
| **Nome do Município** | Descrição do município de funcionamento do curso |
| **UF do Curso** | Código da UF de funcionamento do curso |
| **Inscritos** | Quantidade de alunos inscritos no exame |
| **Participantes** | Quantidade de alunos participantes no exame |
| **Nota Bruta - FG** | Nota Bruta - Formação geral: Média ponderada da parte objetiva (60%) e discursiva (40%) na formação geral (0 a 100) |
| **Nota Padronizada - FG** | Nota Padronizada - Formação geral: Nota convertida de 0 a 5, conceito Enade, baseado na Nota bruta na formação geral |
| **Nota Bruta - CE** | Nota Bruta - Componente específico: Média ponderada da parte objetiva (85%) e discursiva (15%) no componente específico (0 a 100) |
| **Nota Padronizada - CE** | Nota Padronizada - Componente específico: Nota convertida de 0 a 5, conceito Enade, baseado na Nota bruta na formação geral |
| **Nota Bruta - Geral** | Nota Bruta da prova - Geral: Média ponderada da formação geral (25%) e componente específico (75%) (0 a 100) |
| **Nota Padronizada - Geral** | Nota Padronizada prova - Geral: Formação geral: Nota convertida de 0 a 5, conceito Enade, baseado na Nota Bruta da Nota Geral |
| **Conceito Enade (Contínuo)** | Conceito Enade (Contínuo): Nota contínua já convertida de 0 a 5, baseado na Nota Bruta da Nota Geral |
| **Conceito Enade (Faixa)** | Conceito Enade (Faixa): Faixa enade convertida em números inteiros baseado na tabela 1 deste trabalho |
| **Observação** | Observação |