UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

Felipe Mateos Castro de Souza - 11796909
Yago Primerano Arouca - 12608643

# Planejamento de Estudo Exploratório para Classificação de Corpos Celestes

São Paulo
2024

**Sumário**

## 1 Questão de pesquisa geral

A questão de pesquisa geral se concentra em determinar a eficácia dos modelos de aprendizado de máquina na classificação de corpos celestes com base exclusivamente em características espectrais capturadas.

## 2 Definição e coleta dos dados necessários

A partir da questão de pesquisa formulada, optamos por utilizar um conjunto de dados disponibilizado pelo Kaggle, uma plataforma de ciência de dados e inteligência artificial amplamente reconhecida por hospedar competições com prêmios, além de oferecer recursos para compartilhamento e análise de conjuntos de dados. Os dados específicos que escolhemos estão disponíveis em examinem conjuntos de dados. Os dados estão disponíveis em https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17, e foram selecionados por serem originários do Sloan Digital Sky Survey (SDSS), um extenso levantamento astronômico de grande relevância na história da astronomia.

A esfera celeste é uma representação imaginária de uma esfera de raio arbitrário que é concêntrica à Terra, onde todos os objetos celestes, como estrelas, planetas e constelações, são considerados como se estivessem fixos na superfície dessa esfera. Ela serve como uma ferramenta fundamental na astronomia As posições dos objetos celestes são descritas em termos de coordenadas celestiais, como ascensão reta e declinação, que são análogas às coordenadas geográficas usadas na Terra. Essas coordenadas são fundamentais para localizar e rastrear objetos no céu. Esses conceitos são importantes para compreender a forma como os dados foram coletados e facilitar a interpretabilidade de suas variáveis.

O conjunto de dados completo consiste em registros de 100.000 corpos celestes detectados. Para esta análise, optamos por utilizar uma amostra aleatória simples, composta de 10% desse total, o que equivale a 10.000 corpos. É importante destacar que essa amostra será utilizada exclusivamente para a fase de análise exploratória e será descartada em análises subsequentes que incluirão a aplicação de um modelo de aprendizado de máquina.

## 3 Análise exploratória dos dados

A amostra coletada já está organizada em formato de tabela, onde cada linha representa uma observação de um corpo celeste. A tabela está estruturada com as seguintes colunas:

- **alpha** ***(real, quantitativo de razão)***: ângulo de ascensão reta do objeto em relação à esfera celeste.
- **delta** ***(real, quantitativo de razão)***: ângulo de declinação do objeto em relação à esfera celeste.
- **u** ***(real, quantitativo intervalar)***: filtro ultravioleta no sistema fotométrico.
- **g** ***(real, quantitativo intervalar)***: filtro verde no sistema fotométrico.
- **r** ***(real, quantitativo intervalar)***: filtro vermelho no sistema fotométrico.
- **i** ***(real, quantitativo intervalar)***: filtro infravermelho próximo no sistema fotométrico.
- **z** ***(real, quantitativo intervalar)***: filtro infravermelho no sistema fotométrico.
- **redshift** ***(real, quantitativo intervalar)***: valor de desvio para o vermelho (velocidade recessional do objeto) com base no aumento do comprimento de onda.
- **MJD** ***(inteiro, quantitativo intervalar)***: Data Juliana Modificada (em dias) indicando quando o pedaço de dados do SDSS foi obtido.
- **run_ID** ***(inteiro, quanlitativo nominal)***: número da execução para identificar a varredura específica.
- **rerun_ID** ***(inteiro, quanlitativo nominal)***: número da reexecução para especificar como a imagem foi processada.
- **cam_col** ***(inteiro, qualitativo nominal)***: coluna da câmera que especifica a linha de varredura dentro da execução.
- **field_ID** ***(inteiro, quanlitativo nominal)***: número para identificar cada campo.
- **spec_obj_ID** ***(inteiro, quanlitativo nominal)***: identificador único de objeto espectroscópico usado para captura.
- **obj_ID** ***(real, quanlitativo nominal)***: identificador de objeto, ou seja, um identificador único para cada corpo celeste.
- **fiber_ID** ***(inteiro , quanlitativo nominal)*** : especifica a fibra que apontou a luz no plano focal em cada observação.

- **plate** ***(inteiro, quanlitativo nominal)***: identifica o número da placa no sistema SDSS.
- **class** ***(texto, qualitativo nominal)***: especifica a classificação do objeto como galáxia, estrela ou quasar, representados, respectivamente, por 'GALAXY', 'STAR' ou 'QSO'.

As características do conjunto de dados incluem várias colunas que armazenam atributos identificadores. O uso de IDs em tarefas de previsão em machine learning apresenta desafios significativos, como o risco de vazamento de dados e overfitting, especialmente se os IDs estiverem correlacionados com a variável alvo. Além disso, incluir os IDs como características pode aumentar a dimensionalidade do conjunto de dados, dificultar a interpretação do modelo e prejudicar a generalização para novos dados. Portanto, é crucial abordar essas questões com cuidado durante o processo de modelagem. O uso de identificadores também foge do escopo de se focar exclusivamente em características espectrais. Assim, optamos por descartar as seguintes colunas durante a fase de execução:

- **run_ID**
- **rerun_ID**
- **cam_col**
- **field_ID**
- **spec_obj_ID**
- **obj_ID**
- **fiber_ID**
- **plate**

Além disso, a variável MJD pode ser interpretada como um identificador, uma vez que ela identifica o período temporal em que os dados foram obtidos, em vez de representar uma característica espectral do corpo celeste detectado. Portanto, também optamos por desconsiderá-la em nossa análise.

Como as características a serem analisadas consistem exclusivamente em dados numéricos, conduziremos uma análise para identificar a presença de outliers nesses dados. Se outliers forem identificados, analisaremos a natureza dessas discrepâncias para decidir se devemos ou não descartá-los da análise.

Será realizada uma análise para investigar se técnicas de normalização, com o objetivo de uniformizar a escala de todas as características numéricas, serão benéficas para esses dados. Além disso, examinaremos se há desbalanceamento de classes no conjunto de dados. Se identificado desbalanceamento, avaliaremos se técnicas de balanceamento, como oversampling ou undersampling, podem melhorar a capacidade de predição.

Será verificado se existem dados nulos no conjunto de dados. Se forem identificados dados nulos, será realizada uma análise para determinar se sua presença é provavelmente devido a um fenômeno sistemático ou ao acaso. Se os dados nulos forem atribuíveis ao acaso, as observações contendo esses dados serão removidas, especialmente considerando a disponibilidade de muitos dados. Por outro lado, se os dados nulos indicarem uma causa sistemática, uma análise adicional será conduzida para determinar a melhor abordagem para tratá-los.

Por fim, a análise exploratória futura dos dados, incluindo distribuições, gráficos, etc., será conduzida exclusivamente na amostra que contém 10% dos dados. Essa abordagem visa preservar a integridade dos dados restantes, representando 90% do conjunto total, para análises futuras que dependam de uma quantidade maior de dados, como a aplicação de algoritmos de aprendizado de máquina.

## 4 Questões de pesquisa refinadas

Foram formuladas algumas questões específicas a serem investigadas no conjunto de dados:

- Existem determinadas regiões definidas pelos ângulos alpha e delta que possuem agrupamentos de uma classe específica?
- Será que cada tipo de corpo celeste é caracterizado somente por uma única faixa de cor predominante ou uma combinação delas?
- Existe alguma correlação entre as regiões definidas pelos ângulos alpha e delta de um corpo celeste e a sua emissão luminosa, especificamente em relação à faixa de cor observada?
- É possível que determinadas regiões do espaço, definidos pelos seus ângulos alpha e delta, possuem agrupamentos de corpos que estão conjuntamente se afastando ou se aproximando da Terra, baseado no seu valor de redshift?

## 5 Cronograma

Na primeira e segunda semana, realizaremos uma análise minuciosa dos atributos do conjunto de dados, buscando compreender sua estrutura e conteúdo. Para evitar vieses e garantir representatividade, selecionaremos aleatoriamente uma amostra de 10% dos dados para uma análise preliminar. A partir dessa amostra, exploraremos sua distribuição e qualidade, identificando variáveis relevantes e compreendendo as características espectrais capturadas. Essas análises iniciais nos fornecerão insights valiosos para refinar nossas questões de pesquisa, garantindo que sejam relevantes e direcionadas aos objetivos do projeto.

Na sequência, na segunda semana, concentraremos na exploração e pré-processamento dos dados. Aqui, durante essa exploração iremos averiguar se existem transformações e correções a serem feitas no conjunto de dados e as realizaremos caso necessário, baseando-se nos tipos das variáveis. Essa etapa será crucial para garantir a integridade e a qualidade dos dados utilizados em fases posteriores do projeto.

Na terceira semana, avançaremos para a modelagem e avaliação. Utilizaremos os dados ainda não vistos do conjunto "stellar-classification-dataset-sdss17", dividindo-os em conjuntos de treinamento e teste. Em seguida, treinaremos modelos de aprendizado de máquina utilizando os dados de treinamento e os avaliaremos usando métricas apropriadas, como acurácia, precisão, recall e F1-score. Essa etapa nos permitirá compreender o desempenho dos modelos na classificação de corpos celestes com base em suas características espectrais.

Por fim, na quarta semana, realizaremos uma análise detalhada dos resultados obtidos pelos modelos de aprendizado de máquina. Identificaremos padrões ou insights relevantes sobre a classificação de corpos celestes e compararemos os resultados com as expectativas iniciais. Documentaremos os principais achados e conclusões do projeto e prepararemos o relatório final. De forma resumida, temos:

- Semana 1: Análise exploratória.
- Semana 2: Pré-processamento.
- Semana 3: Criação do modelos de aprendizado de máquina e avaliação.
- Semana 4: Comparação dos resultados e escrita do relatório.