Análise Estatística no Contexto de Predição de Mudança de Emprego

Thaís C. Sampaio, João Pedro Campos
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará, Fortaleza, Brasil
Emails: {thaisc, joaopedroscampos}@alu.ufc.br

***Resumo*—Este trabalho trata do pré-processamento de dados para a predição de mudança de emprego de cientistas de dados através de modelos estatísticos, investigando seus preditores e a relação entre eles, assim escolhendo os melhores parâmetros para o modelo.**

I. Introdução

Uma empresa pretende contratar cientistas de dados e deseja, através de análises estatísticas, criar um modelo preditivo para identificar se um candidato que participou do treinamento aceitará um novo emprego e aceitará trabalhar para a nova empresa ou não aceitará o novo emprego e não continuará após o treinamento. Para isso, a empresa coletou informações de todos os candidatos, como experiências passadas, nível de educação e dados demográficos. Com posse desses dados, deseja-se verificar, também, a relação e a importância de cada preditor, fornecendo parâmetros que podem ser usados como decisão durante a contratação e reduzir impactos em custo e tempo, além de melhorar a qualidade dos treinamentos dos funcionários.

Este trabalho está organizado da seguinte forma: Na Seção II é explorada uma análise geral dos dados, discorrendo sobre os métodos utilizados durante a análise. Na seção III, descreve-se os resultados obtidos através das simulações computacionais. Na seção IV, encontram-se as referências bibliográficas, assim como o *link* para o banco de dados e para as simulações em *Python 3* desenvolvidas para este estudo.

II. Metodologia

O dataset *HR Analytics: Job Change of Data Scientists* está disponível em [1]. Esse conjunto de dados, inicialmente, é dividido entre dados de treino e dados de teste. Ele está desbalanceado e o não processamento prévio dos dados pode afetar os resultados. Na subseção a seguir, será feita uma investigação inicial para que seja realizado o pré-processamento do *dataset*.

A. Investigação Inicial

O conjunto de dados encontrado na plataforma do Kaggle possui três arquivos, o conjunto de dados de treino, o conjunto de dados de teste e o exemplo para submissão das respostas do desafio proposto na plataforma. Assim utilizaremos o conjunto de dados de treino, pois esse possui todas as colunas enquanto o conjunto de dados de teste não possui a coluna para predição, as classes. O conjunto de dados de treino, inicialmente, possui $N_{total} = 268.212$ observações. Através da biblioteca Pandas em Python, verifica-se $N_{null} = 20733$ observações nulas, permanecendo $N_{valid} = 247479$ observações válidas. Existem $D = 14$ variáveis preditoras. Ainda utilizando Pandas, pode-se verificar o tipo de cada preditor. Existem 2 variáveis numéricas e 12 variáveis categóricas que serão convertidas em preditores numéricos posteriormente. A caracterização e descrição dos preditores do banco de dados podem ser conferidas na Tabela I e na Tabela II, respectivamente. Destaca-se a observação da coluna de índice 13, pois o *target*, especificamente, é o parâmetro que se deseja prever ao fim dessa análise. O *target* possui duas classes, sendo: $0-$ não está procurando por um emprego novo e $1-$ está procurando por um emprego novo.

Tabela I
Caracterização das colunas do banco de dados

Índice	Coluna	Tipo de dados
0	enrolle_id	Categórico nominal
1	city	Categórico nominal
2	city_development_index	Numérico
3	gender	Categórico nominal
4	relevent_experience	Categórico binário
5	enrolled_university	Categórico nominal
6	education_level	Categórico ordinal
7	major_discipline	Categórico nominal
8	experience	Categórico ordinal
9	company_size	Categórico ordinal
10	company_type	Categórico nominal
11	last_new_job	Categórico ordinal
12	training_hours	Numérico
13	target	Numérico binário

Tabela II
Descrição dos dados nas colunas

Índice	Descrição
0	Identificação para cada candidato
1	Identificação das cidades dos candidatos
2	Índice de desenvolvimento das cidades dos candidatos
3	Gênero ou sexo do candidato
4	Experiência relevante do candidato
5	Tipo de curso matriculado, se houver
6	Nível de escolaridade do candidato
7	Área de conhecimento principal da formação do candidato
8	Experiência total do candidato em anos
9	Número de funcionários da empresa atual, se tiver empregado
10	Tipo da empresa atual, se tiver empregado
11	Diferença em anos do último emprego
12	Horas de treinamentos completados
13	Classes sobre a decisão sobre aceitar ou não o novo emprego

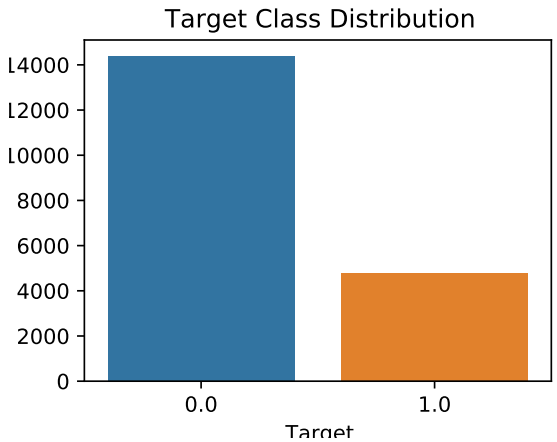

Figura 1. Visualização do *target* por frequência

A análise do *target* será discutida posteriormente e a sua distribuição inicial pode ser observada na Figura 1.

B. Tratamento dos Dados

Como dito anteriormente, para realizar uma análise estatística confiável, é necessário pré-processar os dados, visto que os dados do mundo real geralmente estão "bagunçados". Transformar dados brutos em um conjunto de dados "limpos"evita problemas como ruídos, registros ausentes, incorretos ou duplicados, melhorando a qualidade dos dados.

Um importante passo para o pré-processamento é a limpeza dos dados [2]. Outro passo é a transformação dos dados, que converte variáveis em formatos ou unidades que são mais úteis para os métodos estatísticos escolhidos pelo analista de dados. Essas etapas podem ser repetidas até que sejam organizadas para seu propósito, tomando cuidado para não modificar o conjunto de dados e impactar a análise.

Dessa forma, iniciou-se o pré-processamento com o tratamento dos elementos faltantes nas colunas de dados categóricos nominais de índice 7, 9 e 10, atribuindo uma nova variável que engloba todos os espaços sem informação e depois foi feita a transformação dos dados categóricos para dados numéricos. A coluna de índice 7 foi escolhida para acréscimo de uma classe extra de elementos desconhecidos, pois a área de conhecimento principal da formação do candidato não possuía uma opção para quem não tinha nível de escolaridade de graduação. E as colunas de índices 9 e 10, relacionadas a última empresa ou empresa atual do candidato, foi escolhida para acréscimo de uma classe extra de elementos desconhecidos, pois não tinha a opção para quem não tinha experiência de trabalho. A seguir, a solução para os dados numéricos faltantes foi feita removendo as linhas com esses dados. Ressalta-se que, para o pré-processamento dos dados, também seria possível substituir os valores numéricos faltantes pelo valor médio (ou mais frequente) de cada coluna, entretanto, essa opção não é o melhor método para lidar com valores nulos, pois não leva em consideração a covariância entre os valores [3].

C. Análise dos Dados

Primeiramente, foi realizada a análise mono-variada incondicional dos dados foi feita obtendo o histograma dos preditores e calculando a média μ_d, desvio padrão σ_d e obliquidade γ_d para cada preditor D, utilizando as N observações, obtendo D histogramas, médias, desvios padrões e obliquidades. Calculando a média a partir da equação 1, encontra-se o valor com maior probabilidade de ser encontrado entre os dados.

$$\mu_d = \frac{1}{M}\sum_{i=0}^{M} x_i \tag{1}$$

Onde M é o número de linhas equivalente ao número de candidatos e x_i é a n-ésima amostra do conjunto de amostras $X = \{x_1, x_2, ..., x_i\}$.

Através do desvio padrão, encontra-se o grau de dispersão do *dataset* (ou o quanto os dados estão distantes da média). Ele é dado por:

$$\sigma = \sqrt{\frac{1}{M}\sum_{i=0}^{M} (x_i - \mu)^2} \tag{2}$$

A obliquidade (*skewness*) é dada pela Equação 3, e é utilizada para medir a assimetria das caudas da distribuição. Quando o valor da obliquidade é próxima a zero ou igual a zero a distribuição é simétrica, quando o valor da obliquidade é maior que zero a distribuição apresenta dados centralizados à esquerda e quando o valor da obliquidade é menor que zero a distribuição apresenta dados centralizados à direita.

$$\text{skewness} = \frac{\sum (x_i - \bar{x})^3}{(n-1)v^{3/2}} \text{ onde } v = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{3}$$

Em seguida, realizou-se a análise condicional monovariada, que leva em consideração as variáveis de cada classe. Assim, obteu-se $D \times L$ histogramas, médias ($\mu_{d|l}$), desvios padrões ($\sigma_{d|l}$) e obliquidades ($\gamma_{d|l}$). Para o caso do *target* destacado anteriormente, por exemplo, foi observada a distribuição para cada preditor considerando as duas classes 0 ou 1, presentes na Figura 3 e na Figura 4, respectivamente.

Seguidamente, foi feita a análise bi-variada incondicional dos dados, obtendo-se os gráficos de dispersão para cada par de variável e a matriz de correlação cruzada entre todos os preditores, observada na Figura 5. A correlação bidimensional, considerando duas variáveis x e y, é dada por:

$$\rho(x,y) = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}, \tag{4}$$

Onde

$$cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - x_{mean})(y_i - y_{mean})}{N-1}, \tag{5}$$

O gráfico de dispersão pode ser conferido em [4] e não foi anexado neste trabalho por questões de visualização.

Por último, foi realizada a análise incondicional multivariada, que é feita a partir do calculo dos Componentes Principais e recebe o nome de Análise dos Componentes

Principais e pode ser chamado pela sigla PCA do inglês *Principal Components Analysis*. Porém, antes de iniciar essa parte da análise, é necessário certificar que os dados foram pré-processados e não possuem elementos faltantes, caso contrário, não é possível executar o PCA. Certificado que o conjunto de dados foi pré-processado, é necessário fazer uma transformação nos dados, tornando a média dos parâmetros nula e sua variância unitária, o que pode ser chamado de padronização dos dados, que é o nome dado ao ato de deixá-los na mesma ordem de grandeza. A padronização dos dados é feita a partir da seguinte equação:

$$n_i = \frac{x_i - \mu_X}{\sigma_X} \tag{6}$$

Na equação 6 n_i é a n-ésima amostra padronizada, do conjunto de amostras padronizadas $N = \{n_1, n_2, ..., n_i\}$, a partir da subtração n-ésima amostra x_i, do conjunto de amostras $X = \{x_1, x_2, ..., x_i\}$, pela média μ_X e da divisão do resultado pelo desvio padrão σ_X. O PCA é uma técnica que permite diminuir a dimensionalidade do conjunto de dados e consequentemente diminuir a complexidade do modelo contribuindo para um menor trabalho computacional. As informações redundantes são eliminadas ao utilizar os dados correlacionados combinados a partir de seus autovetores em conjuntos de dados linearmente independentes, não correlacionados, que podem ser chamados de componentes principais. Assim o resultado é um novo conjunto de variáveis que tem variância máxima. Essa recombinação dos dados pode ter uma consequência negativa ao aprendizado e deve ser analisado, pois a redução da dimensionalidade também traz diminuição das informações contidas no novo conjunto e essa diminuição de informação pode ser calculada através de um gráfico da variância acumulada em função do número de componentes da nova dimensão. Portanto essa relação do número de componentes do novo conjunto de dados com a quantidade de informação e com a complexidade do modelo são objetos de estudo para encontrar uma combinação ideal para construir o modelo de aprendizado.

III. RESULTADOS

A. Análise Monovariada

Esta análise revela pouca informação sobre o conjunto de dados, sendo possível apenas observar as estatísticas e a distribuição de cada preditor, mostrado na Tabelas III e na Figura 2. A partir dessas observações, pode-se observar que o gráfico que possui maior simetria é o do Nível de Educação, pois também possui o *skewness* mais próximo de zero, enquanto os outros preditores apresentam dados centralizados mais à esquerda ou à direita. A partir da análise monovariada é possível identificar possíveis padrões que justificam a causa dos dados faltantes, dado que podemos observar quais são as classes presentes em cada preditor e identificar se um preditor possui todas as opções disponíveis para os candidatos, caso negativo as opções de classes incompletas pode estar relacionado aos elementos faltantes. A partir dessa análise conseguimos justificar o acréscimo de uma nova classe às

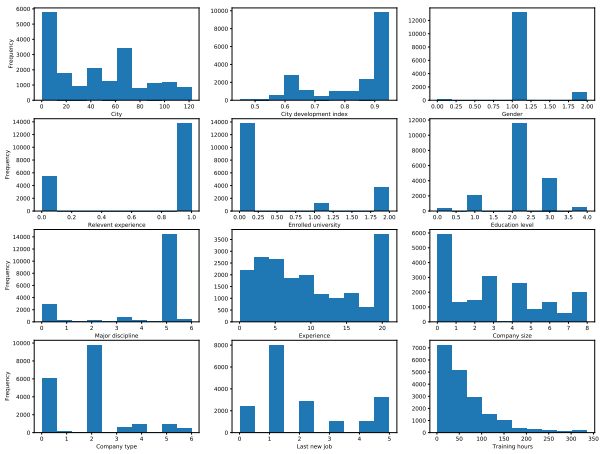

Figura 2. Histograma do conjunto de dados.

colunas de dados categóricos nominais de ındice 7, 9 e 10 como explicado anteriormente.

Tabela III
DESCRIÇÃO DAS L CLASSES

Coluna	Média	Desvio Padrão	Obliquidade
city	44.12	35.46	0.4003
city_development_index	0.8288	0.1234	-0.9954
gender	1.071	0.3040	1.825
relevent_experience	0.7199	0.4491	-0.9795
enrolled_university	0.4641	0.8056	1.265
education_level	2.137	0.6906	-0.08323
major_discipline	3.820	0.7297	-3.405
experience	10.10	6.777	0.4090
company_size	3.253	2.191	0.4124
company_type	1.571	1.157	1.756
last_new_job	2.000	1.676	0.7985
training_hours	65.37	60.06	1.819

A análise condicionada à classe, por sua vez, mostrou resultados mais interessantes, visto que já é possível observar algumas tendências entre os preditores. A partir da análise dos gráficos das figuras 3 e 4 vimos diferenças nas distribuições dos preditores orientados a classe relacionados a cidade e ao índice de desenvolvimento da cidade que pode estar relacionado a cidades que mais tem candidatos que aceitam a proposta do novo emprego. E vimos diferenças nas distribuições dos preditores orientados a classe relacionados ao tamanho da empresa que pode estar relacionado com a capacidade de retenção de um trabalhador, dado que o histograma da classe de candidatos que aceitam a proposta de mudar de emprego está centralizado mais à esquerda onde representa empresas menores.

B. Análise Bivariada

Como foi mencionado anteriormente, alguns preditores possuem correlações significativas, porém nenhum valor grande o suficiente que justificasse a eliminação de algum preditor. Ou seja, não encontramos preditores com correlação linear, pois na matriz de correlação não encontramos valores próximos a 1, nem próximos a -1 e no mapa de calor, Figura 5, não vimos

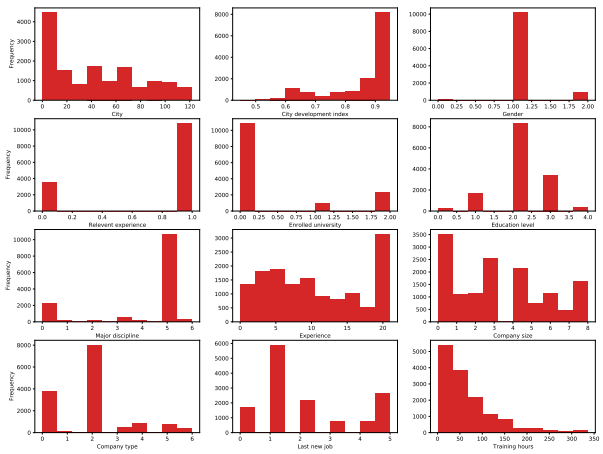

Figura 3. Histograma para a classe 0 (o candidato não está procurando por emprego), considerando todos os preditores.

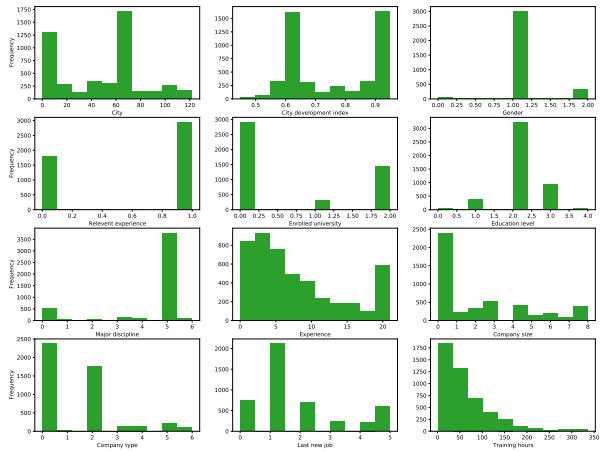

Figura 4. Histograma para a classe 1 (o candidato está procurando por emprego), considerando todos os preditores.

as respectivas cores vermelho vinho e azul anil. Também não observamos figuras com distribuição centralizada em uma diagonal no gráfico de um preditor em função do outro, caso fossem identificadas esses tipos de figuras caracterizariam a correlação linear entre os preditores. Além disso, a quantidade das classes dos preditores é pequena, ou seja, existem muito mais amostras que classes preditores pelo fato da maioria ser de origem categórica.

. Ainda assim, observa-se que os preditores com maior influência em outros preditores são: *last_new_job*, *city_development_index* e *experience*.

Foi observado outro resultado adicionando uma classe em *major_discipline*, ao considerar que as pessoas que não preencheram esse campo possuía pelo menos o menor nível de educação disponível, notou-se uma grande mudança em relação a correlação, observada na Figura 6.

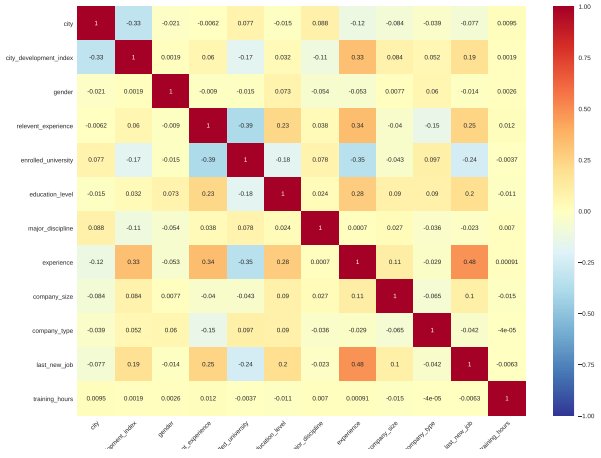

Figura 5. Mapa de calor considerando pares de preditores.

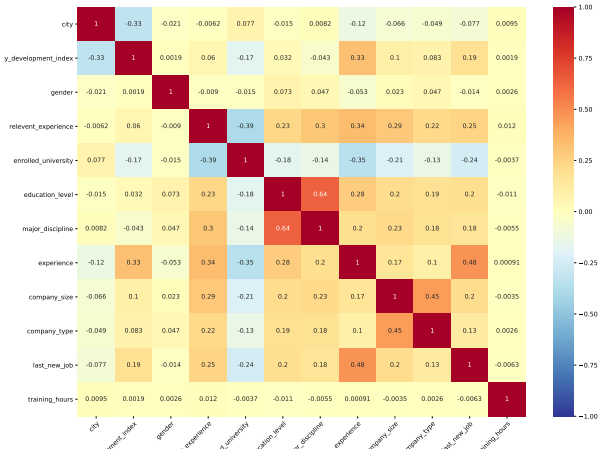

Figura 6. Mapa de calor considerando pares de preditores e uma coluna a mais em *major_discipline*..

C. Análise Multivariada

Os resultados obtidos no PCA foram através do gráfico da Figura 8 que representa a variância cumulativa em função de diferentes números de componentes e por esse gráfico podemos analisar que duas componentes forma um conjunto com muita perda de informação, essas componentes representam aproximadamente 50% da informação dos dados contidos nas 11 colunas escolhidas como preditores, ou seja, todas menos a coluna de gênero. Além disso podemos ver a performance de duas componentes principais no novo conjunto na Figura 8 que representa a distribuição nas novas dimensões e podemos observar que as classes estão misturadas porém a classe de quem não quer o novo emprego está concentrada mais para esquerda e outro parte concentrada na parte superior. Em contrapartida, a classe de quem quer o novo emprego está concentrada na direita inferior. Portanto seria necessário escolher um maior número de componentes principais, sendo o número de 6 componentes, equivalente a aproximadamente 83% de informação, uma boa estimativa para começar as análises de como o modelo de aprendizado se comporta.

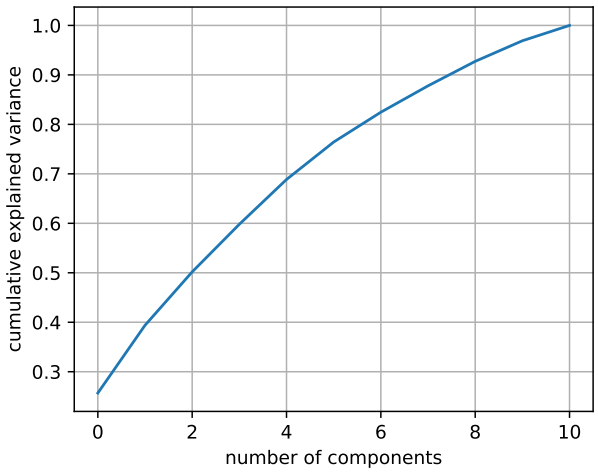

Figura 7. Gráfico das significâncias cumulativas.

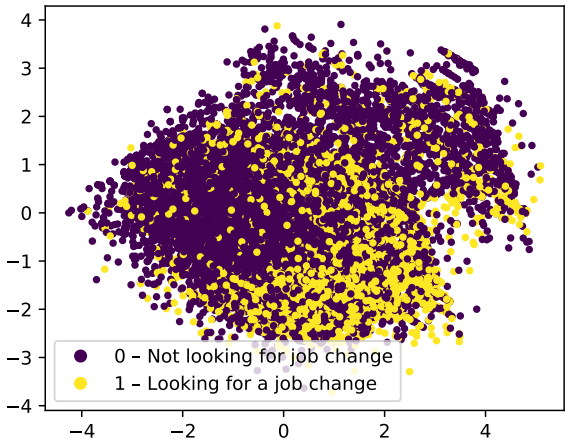

Figura 8. Gráfico do PCA com duas componentes.

D. Conclusões

Ao final do trabalho, observou-se que, visando aplicações no aprendizado de máquina, é interessante retirar a coluna de índice 3 (gênero) pois, além de não ser uma coluna de extrema importância no banco de dados, a quantidade de dados faltantes, cerca de 23% dos dados, e o desbalanceamento dos dados entre os gêneros, tendo aproximadamente 70% de candidatos do sexo masculino e 7% de candidatos do sexo feminino, pode entregar resultados errôneos, considerando também que preditores de gênero e raça envolvem questões éticas.

Ressalta-se, também, que, por ser possível utilizar diferentes métodos de pré-processamento e números de componentes do PCA, o resultado poderá diferir dependendo da técnica utilizada, e o seu desempenho só poderá ser o melhor caso fosse realizada um treinamento diferente para cada possibilidade. Entretanto, como este não é o objetivo do estudo, selecionou-se as técnicas mais simples.

Portanto, a partir do pré-processamento e da análise dos dados, conclui-se que, através de um conjunto de dados, é possível observar as suas características e estatísticas desse conjunto, que podem utilizadas construção de um modelo preditivo de forma eficiente e robusta.

Referências

[1] Möbius. (2020, dec) Hr analytics: Job change of data scientists. [Online]. Available: https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists

[2] M. Critical Data, *Secondary analysis of electronic health records*. Springer Nature, 2016.

[3] T. D. Science. (2020, jul) 7 ways to handle missing values in machine learning. [Online]. Available: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e

[4] T. C. S. João Pedro Campos. (2021, jun) Hr analytics: Job change of data scientists. [Online]. Available: https://colab.research.google.com/drive/1E0gyoCnVJRMTfvKDow-1Rrh77ol3dX7M?usp=sharing