**Nome Completo dos Alunos:** Bruno Seiti Ueda Palma, João Pedro Libonati, Leonardo Ribeiro Camboim, Thiago Almeida Freitas

**Trabalho Final**

**Definição do Problema de Pesquisa:**

“Há algum padrão de mortes entre os vingadores?”: Nós interpretaremos isso como sendo o risco de morte destes vingadores, com base em outras variáveis.

Para isso vamos analisar uma variável relacionada com “mortes”. Observando o dataset, temos a série de colunas “DeathX” (sendo X um número de 1 a 5), sendo uma variável composta que podemos usar para o risco de morte. Nós simplificaremos esta variável composta para “DeathCount”, assim termos como predizer esta variável simples, na qual representa o risco de morte. Se caso a variável for maior que 0, isso indica que o vingador morreu, e igual a 0 não morreu.

**DataPrep:**

Usamos o Alteryx, porque, além de que seu uso nos levará a aprender uma ferramenta nova, o grupo terá mais sinergia com esta ferramenta do que outras tecnologias para a atividade proposta.

Com o Alteryx, fizemos uma análise dos dados e entendemos que as colunas a serem usadas seriam como a seguir: Appearences, Year, Year since joining, Honorary\_v2\*, Gender\_v2\*, Current\_v2\*, Month\*, DeathCount\*, ReturnCount\*.

\* mostra as colunas que criamos com outras colunas. Algumas usando mais de uma coluna e outras com dados extraídos das colunas como da coluna “Full/Reserve Avengers Intro”, que possuia o mês e o ano em string, extraímos o mês e o convertemos para inteiro.

**Modelos:**

Depois de preparar os dados, resolvemos fazer uma regressão logística para chegar em um modelo que pudesse predizer se um vingador morreria ou não (coluna criada DeathCount sendo maior que 0), dependendo das outras variáveis.

Então, usamos as ferramentas Logistic Regression e Stepwise, para criar o modelo de regressão logística. Porém, observamos que o R-Quadrado era de 0.1136. Portanto, isto indica que nosso modelo iria falhar consideravelmente quando fosse aplicado.

Com o workflow do modelo criado, adicionamos um algorítmo usando a ferramenta Score e Create Samples para testarmos várias iterações deste workflow. Com 20 iterações, o modelo acertou 0.714 das vezes.

**Conclusão:**

Não foi encontrado um padrão considerável para predizer as mortes dos heróis com uma boa certeza. Podemos considerar este modelo criado casualmente, e não para algo sério.

**Melhorias Futuras:**

O motivo de não termos encontrado nenhum padrão significante pode se dar por conta da base analisada não ser grande o suficiente. Melhorias podem ser:

- Aumentar a população por coleta de dados ou por oversampling;

- Repensar o workflow utilizando algum método desconhecido para o grupo;

- Fazer uma interpretação do dataset aprofundadamente, como por exemplo usar outras colunas que entendemos que não teria uso significativo para a criação do modelo ou entendendo a coluna “Notes” e criar outras colunas com base nesta.

Mas além disso, vimos que talvez pudessem ser usadas as colunas que formam uma variável composta, como dito na “Definição do Problema de Pesquisa”. Usando essas colunas, talvez fosse possível usar Clusters ou a regressão logística mesmo, porém pensando em padrões de retorno e mortes.