Nosso projeto se trata de explorar um DataSet que nos traz informações como: idade, sexo, número de dependentes, IMC, fumante e as despesas médicas dessa pessoa. A ideia é tentar prever a probabilidade de uma pessoa ser fumante dado a despesa médica dela. Após os dados serem organizados e carregando-os para que sejam manipulados de melhor forma, foi separado uma parte para testes e o DataSet foi lido. Na parte de análise exploratória, foi feita a correlação de todos os dados da tabela, na busca de encontrar uma alta correlação entre fumantes e as despesas médicas, e após analisar o resultado do plot percebeu-se que a correlação entre fumantes e despesas é alto o que possibilitou continuar o projeto com uma certa segurança quanto a estes dados, depois plotou-se a porcentagem de fumantes e não fumantes do total de entrevistados. Então plotamos um histograma das despesas, e a partir dele foi possível perceber que as despesas poderiam ser divididas em 4 grandes grupos, os que gastam menos ou 15000 dólares, entre 15000 e 30000, entre 30000 e 50000 e mais do que 50000. Dai surgiu o questionamento se seria possível prever se uma pessoa é fumante ou não dado as despesas médicas da mesma. Tendo em vista que agora os dados de a pessoa ser fumante ou não e as despesas da mesma, após a divisão das categorias, são qualitativas usamos a regressão logística. Fazendo um plot do gráfico de quantidade de fumantes ou não por grupo, foi percebido que os grupos mais afetados pelo fato de ser fumante ou não, são os grupos que gastam entre 15000 e 30000 e 30000 e 50000 dólares, logo estes são os dados que impactam no resultado final. Percebeu-se também que a faixa etária em que as pessoas mais fumam é entre 18 e 20 anos, ao fazer um plot entre homens e mulheres fumantes, percebeu-se que o resultado de ambos foi bem similar, o que caracteriza que não há mais homens fumantes ou mulheres, foi feito o mesmo plot novamente porém ao invés de usar sexo foi usado número de filhos, o resultado obtido foi que conforme o número de filhos aumenta o número de fumantes diminui como uma escada. Inicialmente foi pensado em fazer o projeto utilizando Naive Bayes, porém após uma reflexão mais profunda sobre o projeto, percebemos que queríamos entender qual o impacto de uma variável no dado a ser estudado que é uma pessoa ser fumante ou não.

A conclusão encontrada pelo grupo foi fazer uma Regressão, todavia pelo fato de termos muitas variáveis para levar em consideração, teria que ser feita uma regressão múltipla. Como o resultado a ser recebido seria binário, sim ou não no caso sim ou não, o método adotado foi a Regressão Logística. Foi feita um ajuste nos dados para que não haja uma tendência para o lado dos não fumantes, tendo em vista que a porcentagem de não fumantes é bem maior que a de fumantes, esse ajuste apenas fez com que os resultados dos dados ficassem mais justos,