### Introduction

Brief into about the project and datasets…

### Data Preparation Findings:

* **Correlation values before and after scaling**: our categorical features have different scales. For example: “Air Pollution” (1 min - 8 max) while “Genetic Risk” (1 min - 7 max) while “Fatigue” (1 min - 9 max) and so on. When we calculated the correlation values, we tried to unify the scales of the categories but surprisingly the correlation values stayed the same.
* **Class labels Correlation**: we found each class label has different correlations with the dataset features, for example the less value for “Dust Allergy” and “Obesity” the more likely the Cancer is still in Low level while the more values of these features the more likely the Cancer reached High level. Also some of the features have high correlation ranking in the original dataset but ranked differently when correlation considered separately per class, for example “Alcohol use” and “Balanced Diet” among the top6 in the original dataset but not in the 3 combined datasets of top2 features, on the other hand “Smoking” is not among the top6 in the original dataset but it is in the 3 combined datasets of top2 features.