**Project Report**

**Teammate: Yanwen Lin (yanwenl), Ruijia Chen(ruijiac)**

1. **Introduction**
   1. **Problem Description**

Worker safety is of utmost important amongst companies in the safety space. The ability to reduce fatalities, injuries and even near-miss incidents can save lives and cut costs. Effective prediction of probability of accidents happening by employing time, temperature and some other factors is essential to take action before accidents occurs.

* 1. **Overview of Approaches Taken**

The main programming language we used is Python with the powerful package tool NumPy and Pandas. Besides, we used BeautifulSoup to parse the original XML file and sklearn package to build the machine learning model.

**1.3 Work Flow**

The work flow of our project is as followed. Every step will be detailed introduced in the next section.

****

1. **Detailed features of the project design**

**2.1 Data preprocessing**

Before predicting, we need preprocess data so that we can take advantage of useful information and get rid of time wasting of redundant material.

2.1.1 XML passing   
Through running python program, we convert each group in XML group into a row in csv file. And grouping little csv files with same property into a large csv file by employing same keywords.

After that we can get two different files, sensor reading and incident occur. We find every row in accident file corresponding to specific row in sensor reading file and label 1 in the new column. Then the rest row label “-1” at this column, treating this column as target.

2.1.2 Feature extract

Intuitively, we can only extract the features of time, site (A specific site corresponds to a specific gas type, like site 1 corresponds to CO, so we consider the site and gas type as a single variable), peak value, average value and temperature. But in fact, time is a variable that contains various information. It has a lot of different periodicity. We convert the time to weekday or not, which day in a week sin time, cos time etc. to get all its cyclical effects on whether or not accidents happen.

We plot the distribution of peak value and average value, we find these two parameters are not has regression effect on whether the accident occur. So, we treat these two parameters as one hot coding rather than numeral feature.

Then we normalize the temperature

2.2 Model construction

2.2.1 Training model

We train the model using logistic regression. Logistic regression is a very powerful model which could not only give us the predicted label but also the probability.

2.2.2 Model selection

To do the model selection, we use the cross-validation to achieve the highest performance of the model tested with a variety range of hyper-parameters.

2.2.3 Predicting

We get a relatively good predicting result with the data we get and the logistic model. Here for the test data, we get 78.5% accuracy.