ANALYSIS ON Level OF CANCER BASED ON VARIOUS FACTORS

Danvanth S

2nd year Integrated M. Sc

(Decision and Computing Sciences)

Coimbatore Institute of Technology

Coimbatore, India

[1933006mdcs@cit.edu.in](mailto:1933006mdcs@cit.edu.in)

Vishnu Sudharsan A

2nd year Integrated M. Sc

(Decision and Computing Sciences)

Coimbatore Institute of Technology

Coimbatore, India

[1933033mdcs@cit.edu.in](mailto:1933033mdcs@cit.edu.in)

Abstract

Throughout our lives, healthy cells in our bodies divide and replace themselves in a controlled fashion. Cancer starts when this process is out of control. A tumor is a mass composed of a cluster of such abnormal cells. Mortality rate due to cancer is 158.3 per 100,000 men and women per year. WHO says that Tobacco use is the most important risk factor for cancer and is responsible for approximately 22% of cancer deaths. This paper deals with the analysis on level of cancer due to these factors. We use Multiple Linear Regression(MLR) model to analyse and predict the level of cancer. MLR is a regression model that estimates the relationship between dependent variable and two or more independent variables .

Review of the literature

Regression analysis model has been used in medical statistics to see the relationship among multiple variables.

Dr. Amita Goel et al.(2017) states that ,Using regression models in the right way and with right data set, Data Analysis and Prediction can produce the most accurate result.

Gibbs Y. Kanyongo et al.(2006) used regression analysis to establish the relationship between home environment and reading achievement.

D A Lawlor et al.(2000) used this method to find relationship between the effectiveness of exercise as an intervention in the management of depression.

# Giorgia Barbara Piccoli et al. used regression analysis correlating dialysis schedules and pregnancy outcomes.

# B.Yildiz et al.(2017) did a analysis of regression and machine learning models on commercial building electricity load forecasting. The regression models which were the main focus of the study performed fairly well in comparison to other more advanced machine learning models.

INTRODUCTION

Cancer has a major impact on society across the world. Cancer Statistics describe what happens in large groups of people and provide a picture in time of the burden of cancer on society.

The most common cancers are breast cancer, lung and bronchus cancer, prostate cancer, colon and rectum cancer. Prostate, lung cancer account for an estimated 43% of all cancer diagnosed in men in 2020. For women three most common cancer are breast , lung and colorectal and they will account for an estimated 50% of all new cancer diagnoses in women 2020. Lung cancer is the second most common cancer in both men and women. [Multiple linear regression](https://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-multiple-linear-regression/) is the most common form of linear regression analysis.  As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.  The independent variables can be continuous or categorical. Multiple linear regression (MLR) is used to determine a mathematical relationship among a number of random variables. In this analysis we are using MLR to analyse the relation among the level of cancer and other factors causing the cancer.

Objective of the study

* To study the data of patients affected by lung cancer.
* To analyse the factors affecting the level of cancer.
* To fit the regression equation for the level and other factors.

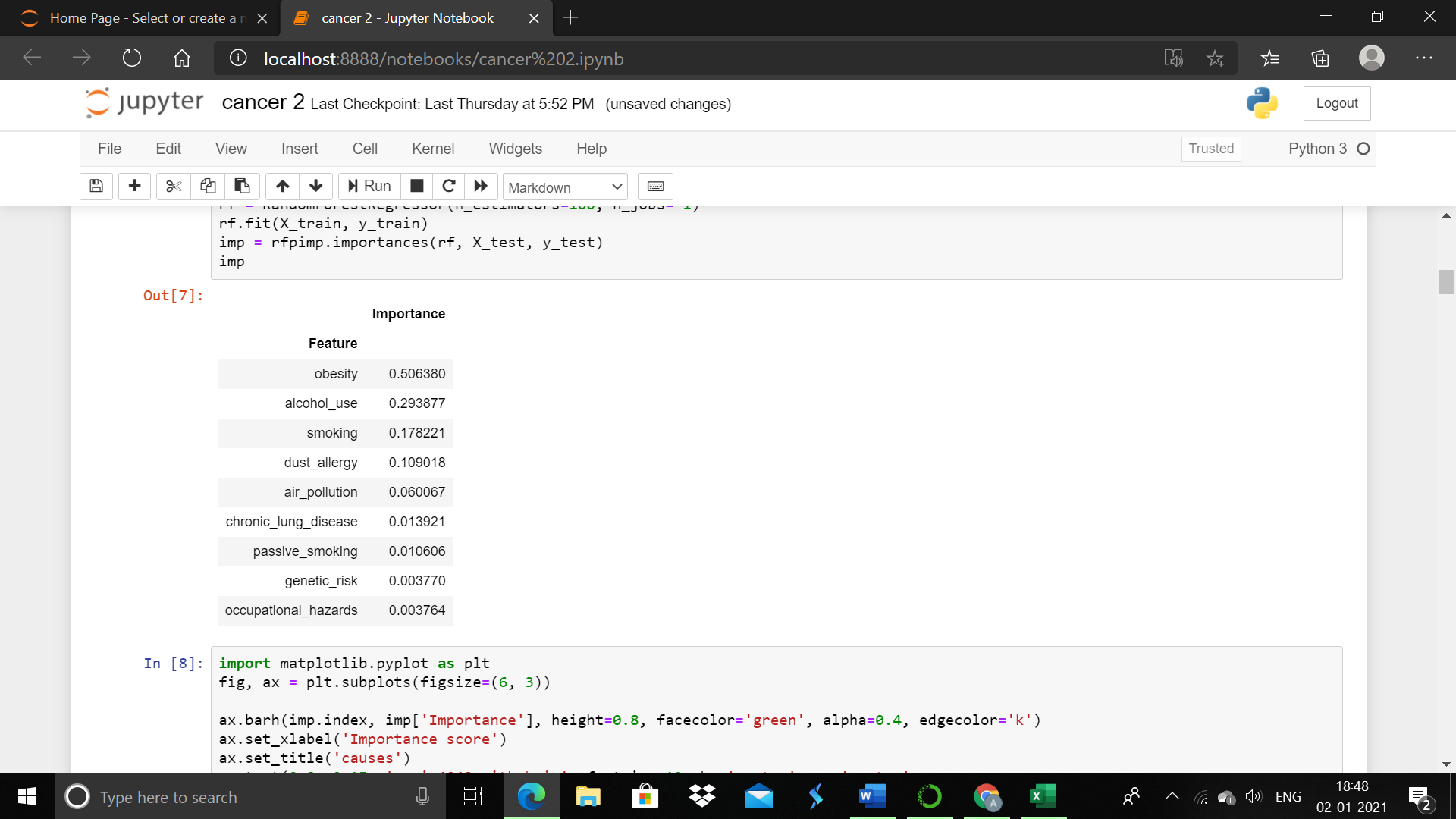
Data and methodology

The data of 1000 patients suffering from cancer were collected. The data contains patient id, age, factors like air pollution, alcohol use, obesity, occupational hazards , genetic risk smoking and dust allergy. It also contains the symptoms and level of the cancer. The level of cancer is divided into low, medium and high. We encoded this categorical data into numerical data by integer encoding. Low was given the value 1, medium was given 2 and high was given 3. The data was reconstructed by removing the unwanted columns for the analysis.

## Importance score:-

Now the data contains 9 factors causing cancer. The importance of these factors are checked using rfpimp library in python. We obtain the following information:-

Table Importance level



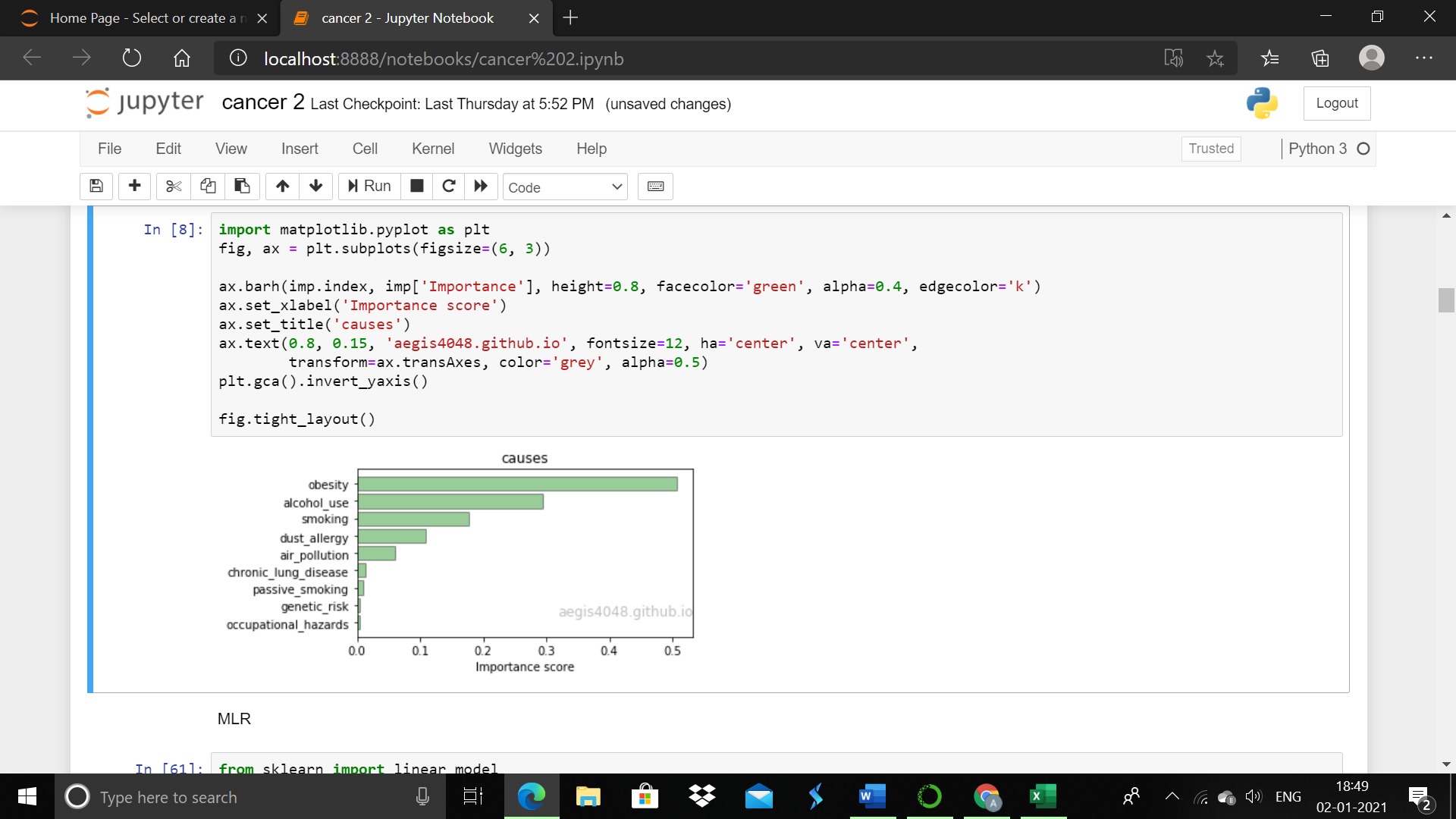


Figure Graphical representation of importance score

Based on the importance score obesity, smoking, alcohol use and dust allergy were selected.

## Regression model

## The regression model is fitted by taking level as dependent variable and obesity, smoking, alcohol use and dust allergy as independent variable.

Level =

The following coefficients were obtained,

Intercept 0.434952

alcohol\_use 0.056922

dust\_allergy 0.043546

obesity 0.190350

smoking 0.069722

## Hypothesis testing (f-statistic)

We are talking null hypothesis as :- All coefficients are equal to zero and alternate hypothesis as :- Atleast one coefficient is not zero.

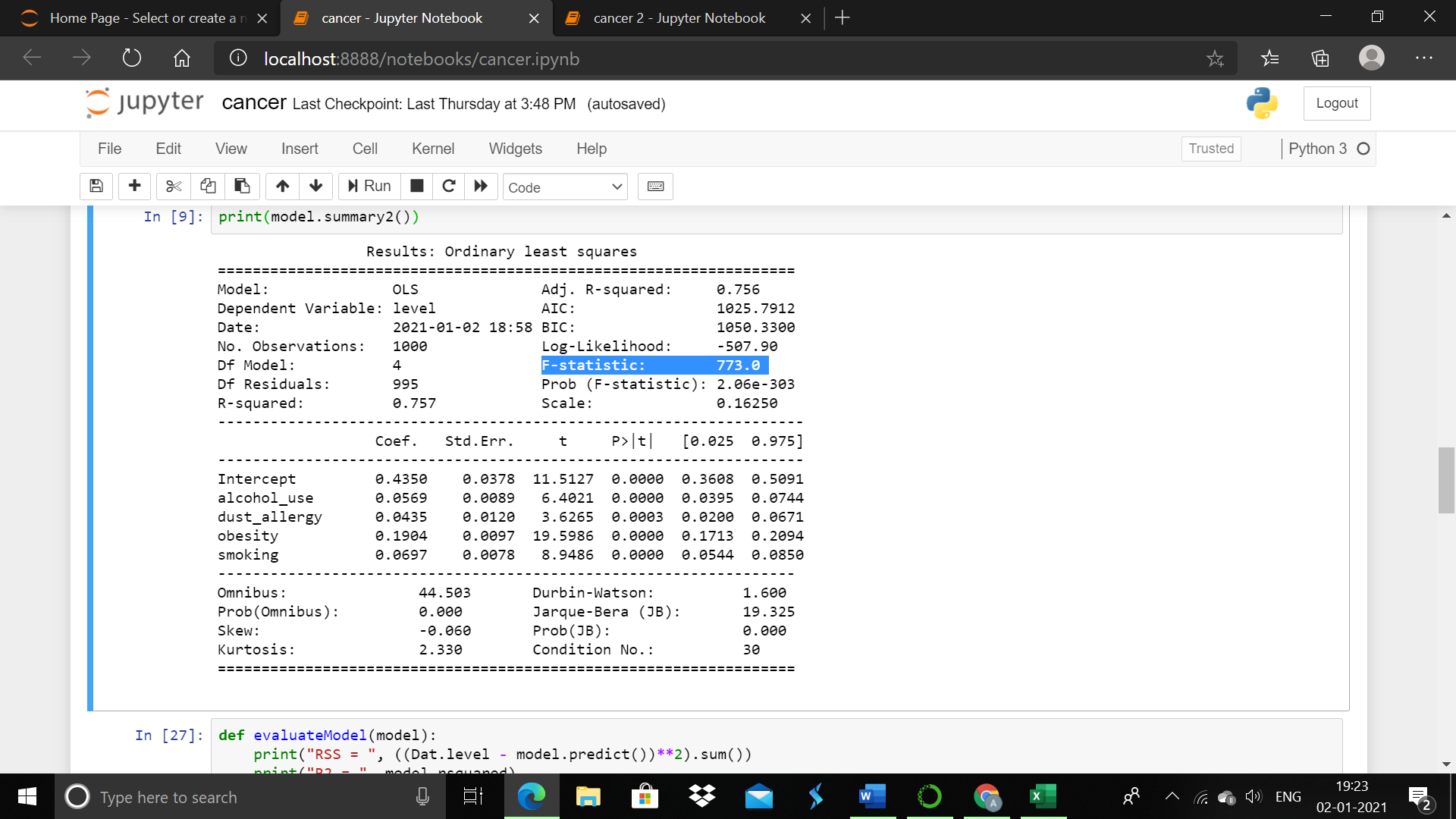


Figure Description of the data

If the value of F-statistic is equal to or very close to 1, then the results are in favor of the Null Hypothesis and we fail to reject it. But as we can see that the F-statistic is many folds larger than 1, thus providing strong evidence against the Null Hypothesis (that all coefficients are zero). Hence, we reject the Null Hypothesis and are confident that at least one predictor is useful in predicting the output.

## FEATURE SELECTION

In this model forward feature selection is used. The RSS can be calculated by

The RSS was found least to following equation.

Level = 0.4349520.069722\*.

## PLOT

Due to the 3D nature of the plot, multiple plots were generated from different angles. Two factors (obesity and alcohol) were used to predict the response variable *level*. With the help of the additional feature *alcohol use* , the linear model experience significant gain in accuracy, now capturing 73% variability of data.

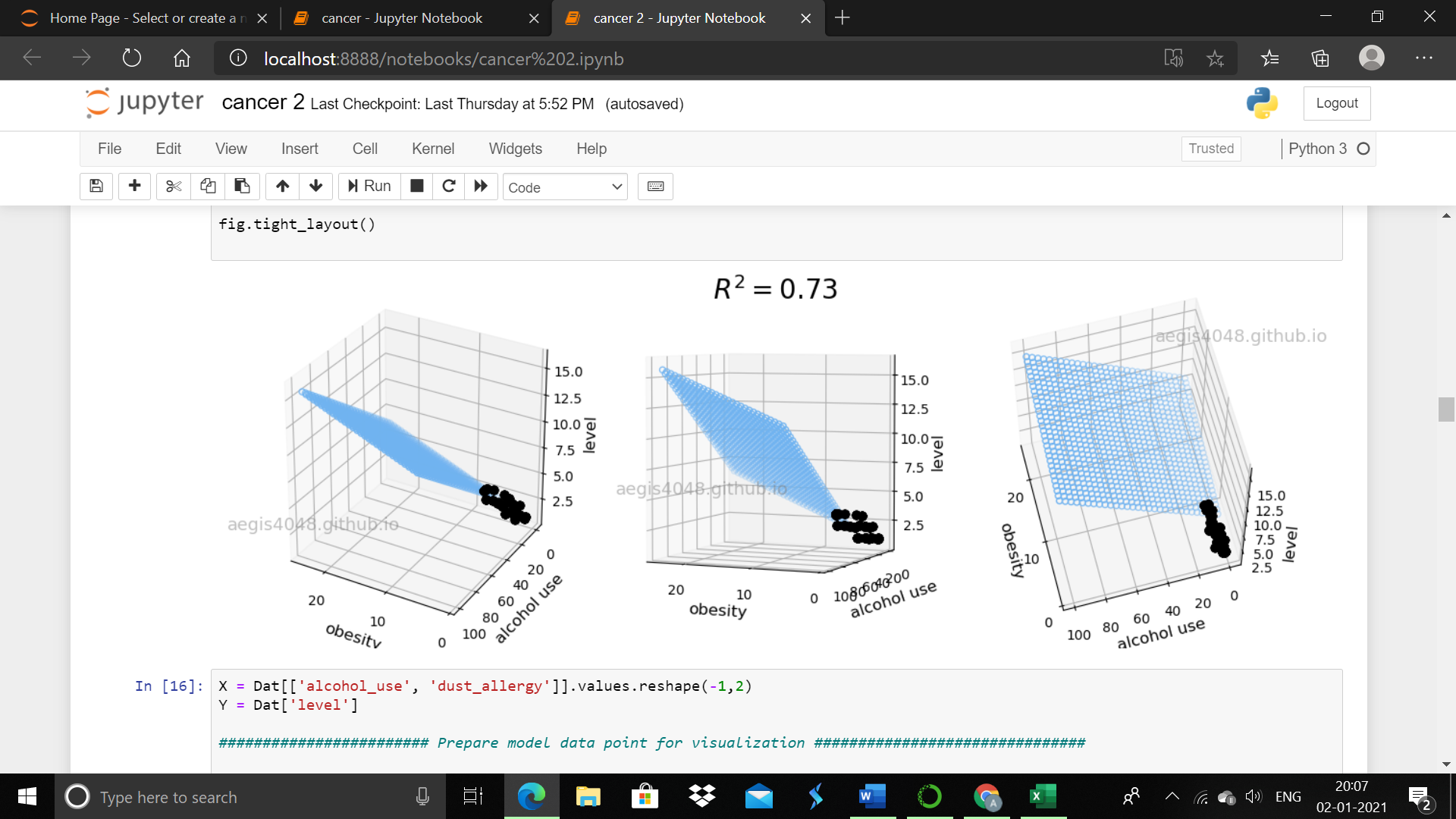


Figure 3D graph for strong features

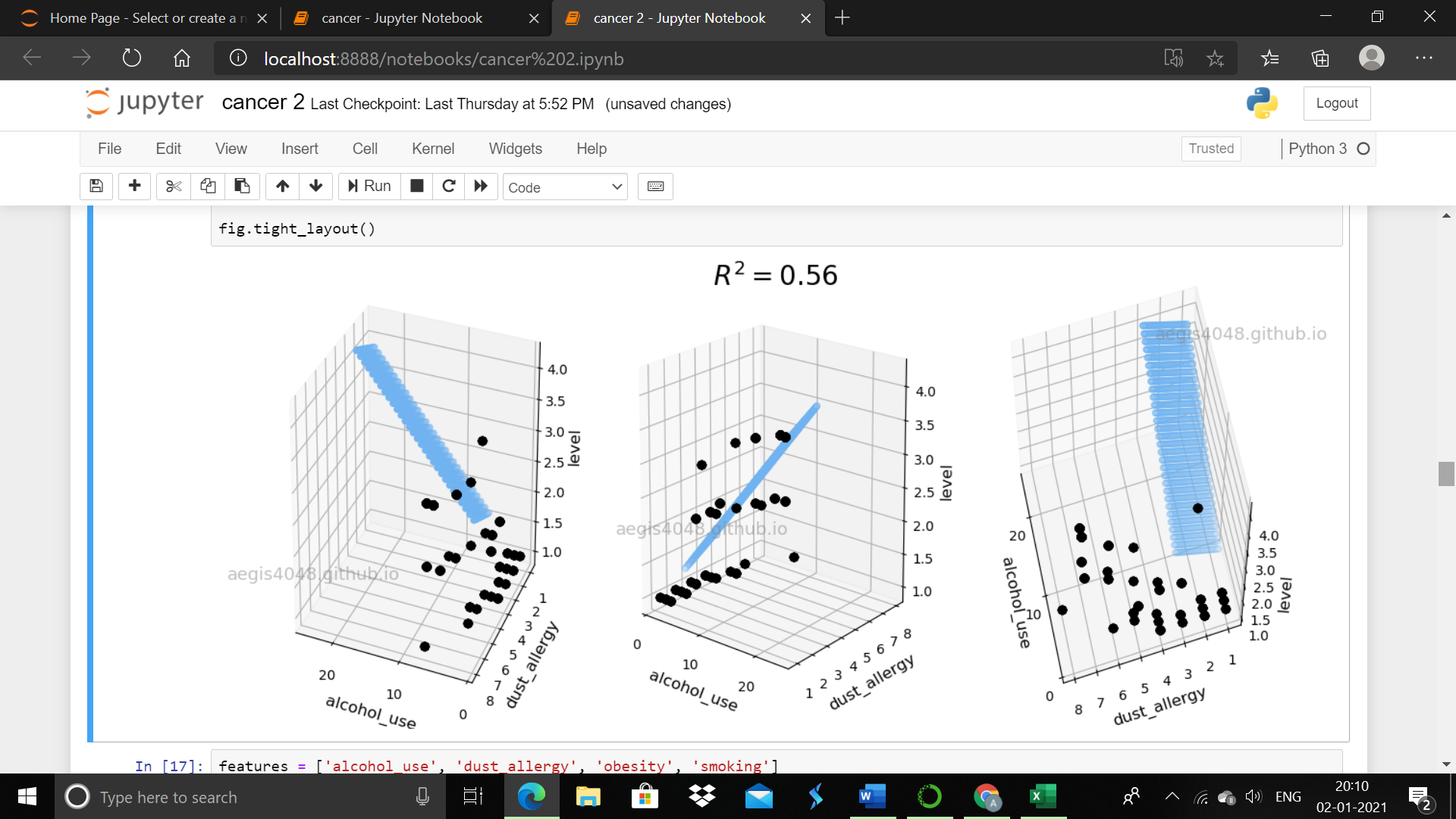
This plot represents 3D linear regression model with strong features.  


Figure 3D graph with weak features

The above plot represents 3D linear regression model with weak features.

CORRELATION

The spearmen correlation for all factors.

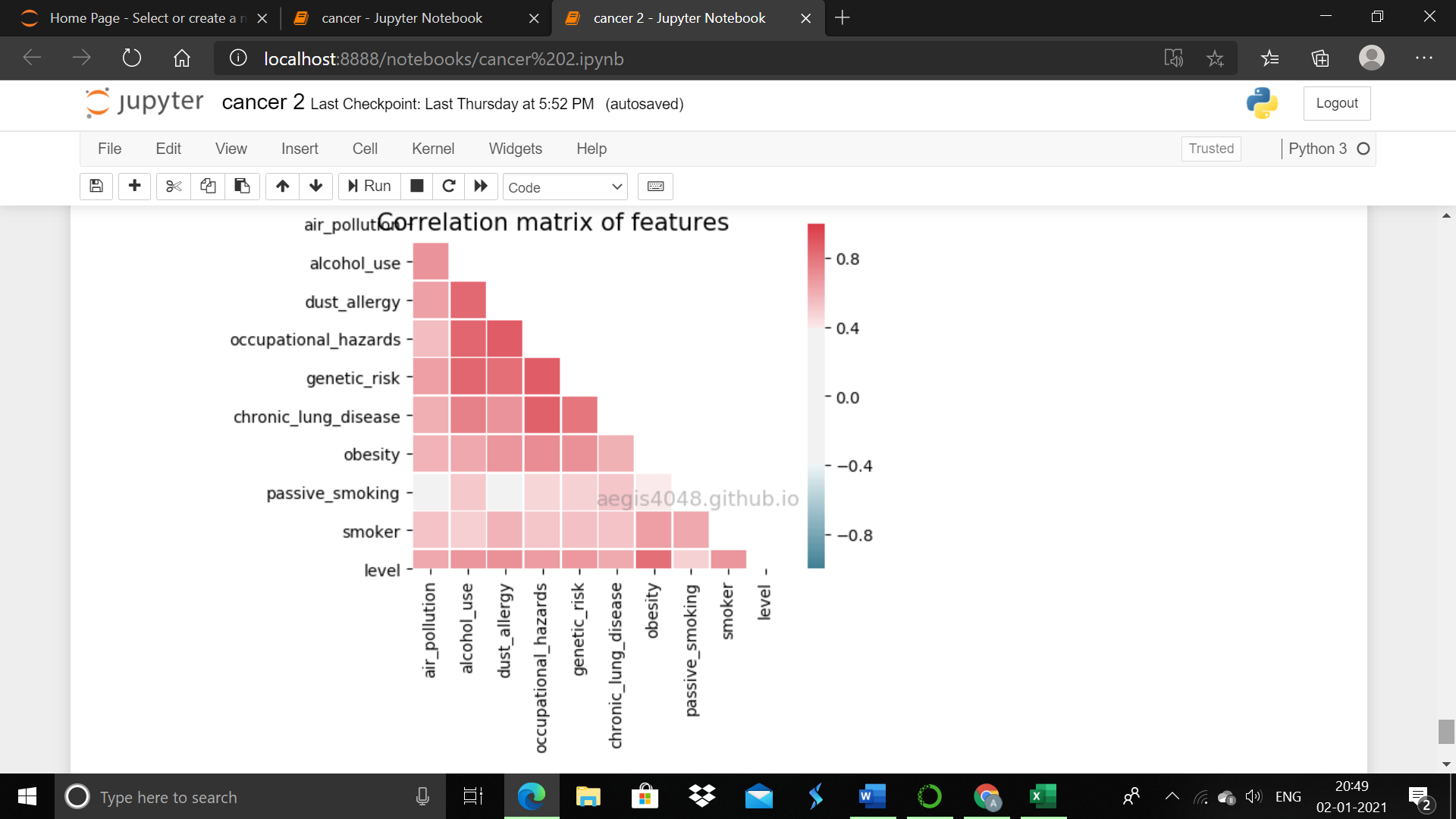


Figure Rank correlation matrix

RESULTS AND DISCUSSION

From the regression equation we take the value of R2 . This model explains 76% of the variation in the level of cancer based on these factors. The regression equation shows that the level of cancer increases 0.05 % when alcohol consumption increases 1%. It increases 0.04% when dust allergy increases 1%, 0.19 % when obesity increases 1% and by 0.06% when smoking increases by 1%.

The plots are done with strong features and weak features.

LIMITATIONS

 Even if the model is good at predicting the response variable with given features (high R-squared), linear model is not robust enough to fully understand the effect of individual features on the response variable. In such circumstance, we can't trust the values of regression coefficients. There is an instability in regression coefficients .This is because the *dust allergy* , *alcohol use*, and *obesity* shows strong linear correlation with one another, as shown in the spearman's correlation matrix.

Refrences

[1] www.cancerresearchuk.org/about-cancer/what-is-cancer

[2] https://en.wikipedia.org/wiki/Cancer

[3] https://www.cancer.org/

[4] https://www.cancer.net/cancer-types

[5] https://en.wikipedia.org/wiki/Analysis\_of\_variance

[6] www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/

[7] https://www.statisticssolutions.com/anova-analysis-of-variance/

[8] https://www.youtube.com/watch?v=ITf4vHhyGpc

[9] https://www.statisticssolutions.com/what-is-linear-regression/

[10] https://en.wikipedia.org/wiki/Linear\_regression

[11] https://onlinecourses.science.psu.edu/stat501/node/250

[12] https://www.youtube.com/watch?v=zPG4NjIkCjc