***SCRIPT***

Julie

***SLIDE 1***

Hello everyone, we are group 9 and we choose to talk about lung cancer which is one of the most common cancers worldwide.

If you didn’t know the five-year survival rate is not above 15%.

However, early-diagnosed lung cancer has a cure rate of more than 70%, so efforts to prevent and treat it are very important.

***SLIDE 2***

Our goal is to predict if patients have lung cancer or not.

So that patients can get early in-depth tests.

***SLIDE 3***

After some readings, it is important for us to help in the diagnosis of lung cancer thanks to a machine learning approach.

We choose to work on Al generator data combined with clinical data.

***SLIDE 4***

- ﻿﻿EDA: Make visualizations

- ﻿﻿Model Selection: we built models such as random forest, gradient boosting, ada boost, and MLP.

- ﻿﻿Model Evaluation: we choose the best model considering accuracy and AUC scores

- ﻿﻿System Building: To finally build our system

***SLIDE 5***

After the preprocessing and for the EDA step, we set 6 hypotheses as follows. The older the person, the smoker, the higher the bmi, and the lower the test result, the higher the probability of cancer. And treatments will increase survival rates. In addition, survival rates vary depending on the cancer size.

***SLIDE 6***

In case of age, we can check simply through histogram. People with cancer are more distributed in older areas. Since then, we can know that age and lung cancer have a relationship.

***SLIDE 8***

Past smoking and current smoking were reclassified as 'smoking'. Since both groups had a high rate of smoking, the relationship between cancer and smoking was difficult to know. Therefore, we need more data about periods or the amount of smoking to verify. .

***SLIDE 9***

Violin plot was drawn with BMI data based on the presence of cancer. Since only plotting is not enough, I used a chi-square test and an independent t-test. As a result, both p-value was lower than .05 and I can say BMI and cancer have a relationship.

***SLIDE 10***

Next, I looked into the relationship between the FEV test and cancer. When considering the entire data, since it’s hard to know the relationship, I set a given condition that the test result is under 70%. If then, as we can know through a box plot, it’s more likely to have cancer.

***SLIDE 11***

Another test to know lung function is DLCO. Also in this test, for the same reason, I set a given condition that the test result is under 80. If then, as we can know through a box plot, it’s more likely to have cancer.

***SLIDE 12***

The next part is to know whether treatments help to survive. After making heatmaps, I calculated some conditional probabilities of each treatment. As you can see, radiation therapy helps to survive, operation helps a little, and chemotherapy does not.

***SLIDE 13***

Finally, I drew the survival curve for small-sized and non-small-sized cancers. Since the two curves overlap and there seems to be no difference, I performed a log-rank test. As a result, it can be confirmed that there is no difference in practice because the p-value was higher than .05.

***SLIDE 14***

We separated our problem into two steps. First classifying whether cancer or not. And also secondly we classified whether non-small or small, because there is significant difference in treatment. By doing this, we could get AUC and F1 scores too.

***SLIDE 15***

At all steps, oversampling method SMOTE is used to balance the target. There was always a large imbalance in target. Without SMOTE the ratio of data and the performances were almost same, while AUC was around 0.5

***SLIDE 16***

As explained before, we will use these as metrics; accuracy, AUC, and F1 score. They are calculated as you see here. Because the higher the more preferable, we will check them.

***SLIDE 17***

And this is the result. Random forest got the best performance. Since it is high enough, we can conclude that we can classify cancer patients at least in our data.

***SLIDE 18***

Using the same method, we could predict their stage. Random forest performed best, but not as high as before. It seems that it is hard to distinguish, because there are few early symptoms.

Random forest says treatment itself and size itself are not important to the death rate. It seems rather than themselves, changes caused by them are more important.

***SLIDE 19***

To build our system, we saw that the random forest model has the best accuracy, AUC and F1-score.

Moreover, the most important features about the prediction are BMI, FEV, DLCO and Age.

There is a correlation between the decrease in lung function and smoking.

***SLIDE 20***

Finally, in hospital we can test FEV and DLCO and in the future it will be easier if we can make kits like for corona cases.

Moreover, we could suggest the best combination of treatments to cure lung cancer when it is possible (early stage).

Thanking