Vietnam General Confederation of Labor

TON DUC THANG UNIVERSITY

FACULTY OF INFORMATION TECHNOLOGY

FINAL REPORT

MACHINE LEARNING

Instructor: **Mr. LE ANH CUONG**

Student: **DUONG NGOC BAO NHI – 521K0143**

Student: **PHAM LE QUOC DAT – 521K0128**

Class :**21K50301**

Year **: 2023-2024**

HO CHI MINH CITY, 2023

Vietnam General Confederation of Labor

TON DUC THANG UNIVERSITY

FACULTY OF INFORMATION TECHNOLOGY

FINAL REPORT

MACHINE LEARNING

Instructor: **Mr. LE ANH CUONG**

Student: **DUONG NGOC BAO NHI – 521K0143**

Student: **PHAM LE QUOC DAT – 521K0128**

Class :**21K50301**

Year **: 2023-2024**

HO CHI MINH CITY, 2023

ACKNOWLEDGEMENT

I would like to thank the lecturer Le Anh Cuong. During the process of studying and learning about the subject, I have learned a lot of knowledge as well as many interesting things from your lessons. He has helped us expand new and in-depth knowledge about this subject. During the learning process, we are still weak, but from the knowledge learned, I have gained important knowledge about specialized skills as well as knowledge to explore and solve problems.

In the learning process, shortcomings are inevitable. I hope to receive feedback from the teacher to make my report more complete.

In addition, I would like to thank the school for creating conditions for us to have a good learning environment so that I can gain more of this knowledge.

We wish you health, happiness and success in your teaching career.

Ho Chi Minh city, 14th December, 2023

Author

(Sign and write full name)

Duong Ngoc Bao Nhi

Pham Le Quoc Dat

CONFIRMATION AND ASSESSMENT SECTION

Instructor confirmation section

_____
_____
_____
_____
_____
_____

Ho Chi Minh 14 December, 2023

Duong Ngoc Bao Nhi

Pham Le Quoc Dat

Evaluation section for grading instructor

_____
_____
_____
_____
_____
_____

Ho Chi Minh, 14 December 2023

Duong Ngoc Bao Nhi

Pham Le Quoc Dat

THE PROJECT IS COMPLETED

AT TON DUC THANG UNIVERSITY

Our team would like to assure that this is our own research project and is under the scientific guidance of Lê Anh Cường Teacher. The research content and results in this topic are honest and have not been published in any form before. The data in the tables for analysis, comments, and evaluation were collected by the author from different sources and clearly stated in the reference section.

In addition, the report also uses a number of comments, assessments as well as data from other authors and other organizations, all with citations and source notes.

If any fraud is detected, our team will take full responsibility for the content of our IT Project Report 2. Ton Duc Thang University is not involved in copyright violations caused by us during the implementation process (if any).

Ho Chi Minh city, 14th December, 2023

Author

(Sign and write full name)

Duong Ngoc Bao Nhi

Pham Le Quoc Dat

TABLE OF CONTENTS

Task 2: Predict Mortality Rate base on Medical history

Dataset used: COVID-19 Dataset by MEIR NIZRI

Original data from Gobierno de México - Información referente a casos COVID-19 en México

0. Dataset Information

Attributes

These fields are excluded from our analysis due to irrelevancy to the problem:

Field	Description
USMER	Indicates whether the patient is treated in USMER or not.
MEDICAL_UNIT	Type of institution of the National Health System that provided the care.
PREGNANT	Whether the patient is pregnant or not.

ID	Field	Description
1	**SEX**	1. Female 2. Male
2	**PATIENT_TYPE**	Type of care the patient received in the unit: 1. Returned home 2. Hospitalized
3	**DATE_DIED**	The date of death or '9999-99-99' otherwise.
4	**INTUBED**	Whether the patient was connected to the ventilator.
5	**PNEUMONIA**	Whether the patient already have air sacs inflammation or not.
6	**AGE**	Age of the patient.
7	**DIABETES**	Whether the patient has diabetes or not.
8	**COPD**	Whether the patient has Chronic Obstructive Pulmonary Disease or not.
9	**ASTHMA**	Whether the patient has asthma or not.
10	**INMSUPR**	Whether the patient is immunosuppressed or not.
11	**HIPERTENSION**	Whether the patient has hypertension or not.
12	**OTHER_DISEASE**	Whether the patient has other disease or not.
13	**CARDIOVASCULAR**	Whether the patient has heart or blood vessels related disease.
14	**OBESITY**	Whether the patient is obese or not.
15	**RENAL_CHRONIC**	Whether the patient has chronic renal disease or not.
16	**TOBACCO**	Whether the patient is a tobacco user.
17	**CLASIFFICATION_FINAL**	Covid test results. Values 1-3 mean that the patient was diagnosed with COVID in different degrees. 4-5 means the test is inconclusive. 4 = INVÁLIDO POR LABORATORIO (INVALID BY LABORATORY) 5 = NO REALIZADO POR LABORATORIO (NOT PERFORMED BY LABORATORY) 6 = CASO SOSPECHOSO (SUSPECTED CASE) (Probably negative) 7 means the patient is negative with COVID.
18	**ICU**	Whether the patient had been admitted to an Intensive Care Unit.

0.1 : Libraries

0.1 : Data Filtering and Transformation

Filtering:

- 98, 99 = Missing values

- For CLASIFFICATION_FINAL, we remove entries == 4 or 5 because they may not be fit for analysis

```
df = pd.read_csv('Covid Data.csv')

df_filtered = df[~(
    (df['INTUBED'] == 99)  | (df['PNEUMONIA'] == 99)  | (df['DIABETES'] == 98)  |
    (df['COPD'] == 98)  | (df['ASTHMA'] == 98)  | (df['INMSUPR'] == 98)  |
    (df['HIPERTENSION'] == 98)  | (df['OTHER_DISEASE'] == 98)  |
    (df['CARDIOVASCULAR'] == 98)  | (df['OBESITY'] == 98)  |
    (df['RENAL_CHRONIC'] == 98)  | (df['TOBACCO'] == 98)  |
    (df['CLASIFFICATION_FINAL'] == 4)  | (df['CLASIFFICATION_FINAL'] == 5) |
    ((df['ICU'] == 99))
)]
```

Transformation

- All columns except AGE is converted to a boolean value ●

Column names are renamed for clarity

```
df_cleaned = pd.DataFrame()
df_cleaned['IS_MALE'] = df_filtered['SEX'].apply(lambda x: True if x == 2 else False)
df_cleaned['IS_SENT_HOME'] = df_filtered['PATIENT_TYPE'].apply(lambda x: False if x == 2 else True)
df_cleaned['IS_DEAD'] = df_filtered['DATE_DIED'].apply(lambda x: False if x == '9999-99- 99' else True)
df_cleaned['IS_INTUBED'] = df_filtered['INTUBED'].apply(lambda x: False if (x == 2 or x == 97) else True)
df_cleaned['HAS_PNEUMONIA'] = df_filtered['PNEUMONIA'].apply(lambda x: False if x == 2 else True)

df_cleaned['AGE'] = df_filtered['AGE']

df_cleaned['HAS_DIABETES'] = df_filtered['DIABETES'].apply(lambda x: False if x == 2 else True)
df_cleaned['HAS_COPD'] = df_filtered['COPD'].apply(lambda x: False if x == 2 else True) df_cleaned['HAS_ASTHMA'] =
df_filtered['ASTHMA'].apply(lambda x: False if x == 2 else
True)
df_cleaned['HAS_INMSUPR'] = df_filtered['INMSUPR'].apply(lambda x: False if x == 2 else True)
df_cleaned['HAS_HYPERTENSION'] = df_filtered['HIPERTENSION'].apply(lambda x: False if x == 2 else True)
df_cleaned['HAS_OTHER_DISEASE'] = df_filtered['OTHER_DISEASE'].apply(lambda x: False if x
== 2 else True)
df_cleaned['HAS_CARDIOVASCULAR'] = df_filtered['CARDIOVASCULAR'].apply(lambda x: False if x == 2 else
True)
df_cleaned['HAS_OBESITY'] = df_filtered['OBESITY'].apply(lambda x: False if x == 2 else True)
df_cleaned['HAS_RENAL_CHRONIC'] = df_filtered['RENAL_CHRONIC'].apply(lambda x: False if x
== 2 else True)
df_cleaned['USE_TOBACCO'] = df_filtered['TOBACCO'].apply(lambda x: False if x == 2 else True)
df_cleaned['IS_POSITIVE'] = df_filtered['CLASIFFICATION_FINAL'].apply(lambda x: True if (x in [1, 2, 3]) else
False)
df_cleaned['IS_ICU'] = df_filtered['ICU'].apply(lambda x: False if (x == 2 or x == 97) else True
```

1. Statistical Analysis

1.1 Correlation (Pearson)

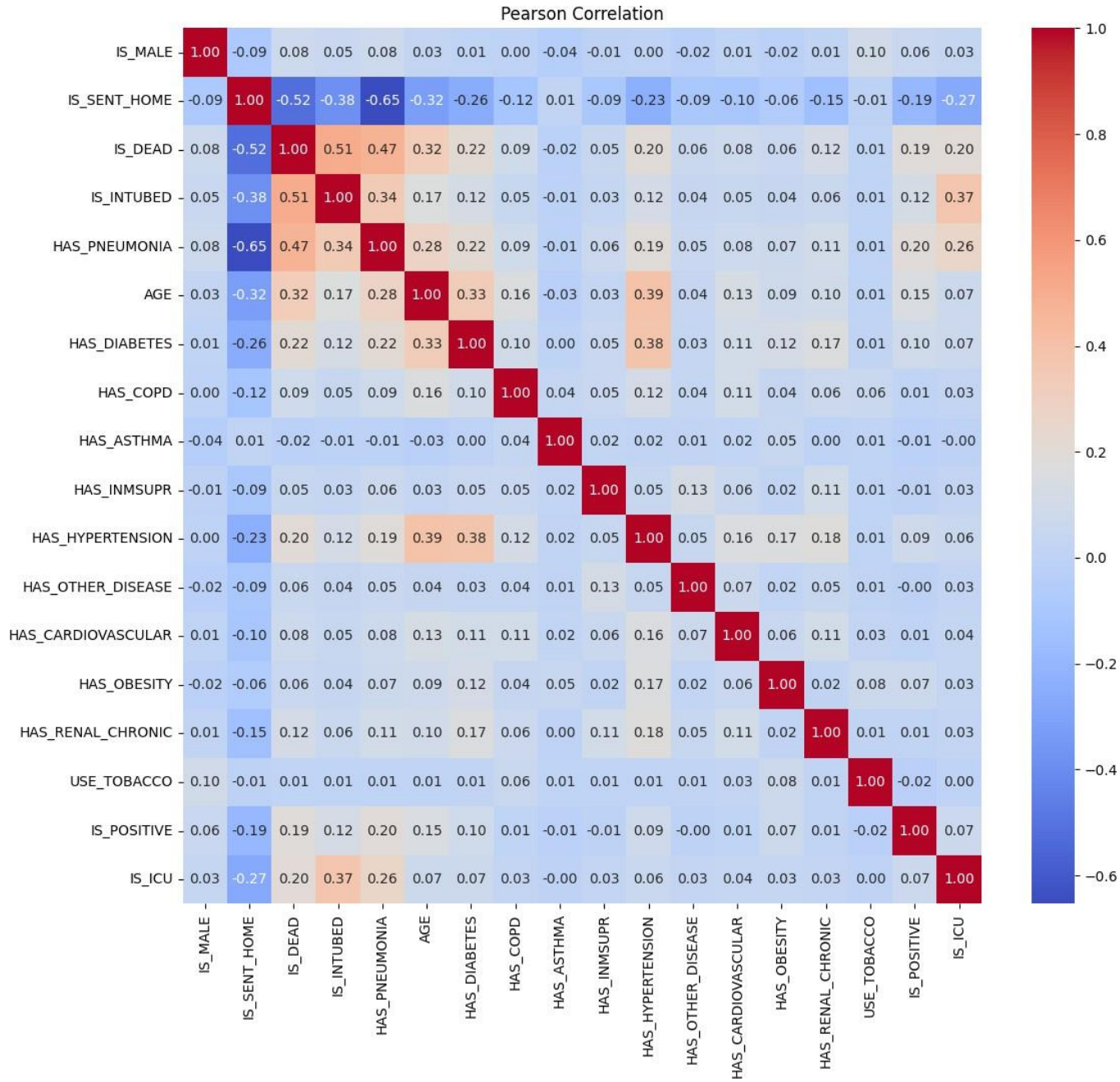

Figure 1: Pearson Correlation between the variables

These are the observations we made from figure 1:

1. IS_SENT_HOME moderately **negatively** correlates with Disease-type attributes and IS_DEAD (-0.52). Highest is with HAS_PNEUMONIA (-0.65).

   This indicates that people who are sent home (not hospitalized) are healthier than hospitalized patients. And it seems that people admitted are more likely to be dead.

2. AGE moderately **positively** correlates with IS_DEAD (0.32), HAS_PNEUMONIA (0.28), HAS_DIABETES (0.33), HAS_HYPERTENSION (0.39) and moderately **negatively** correlates with IS_SENT_HOME (-0.32).

This indicates that young people are in general more healthy than old people, and is less prone to diseases such as Pneumonia, Diabetes or Hypertension. Older people are also more likely to be admitted, instead of being sent home.

3. IS_POSITIVE suprisingly doesn't correlate as much as we expected, to both the mortality rate and diseases. The correlations that stands out are IS_DEAD (0.19), IS_INTUBED (0.12), HAS_PNEUMONIA (0.2) and AGE (0.15).

   This may indicate that COVID positivity is more isolated than being enabled by other diseases than we thought. Notably still, it is more correlated with Respitory-conditions like Pneumonia and people who are positive are more likely to be put on a ventilator, as we would think. Older people are more likely to be positive, probably because of their declining immune system.

4. HAS_HYPERTENSION (High-blood pressure) is moderately **positively** correlated with HAS_DIABETES (0.38)

   This makes common sense.

5. People who are admitted to the ICU (Intensive Care Unit) are more likely to be put on ventilator (correlation 0.37)

6. IS_DEAD is **positively** correlated with IS_INTUBED (0.51), HAS_PNEUMONIA (0.47) and AGE (0.32)

   This indicates that people who are older, had Pneumonia and was put on ventilator are more likely to die. These three conditions usually come together.

7. The gender of the patient, and whether the person is a smoker, do not affect much to the mortality rate, shown by the low correlation (0.08 and 0.01 respectively). Do note that males are more likely to smoke.

8. HAS_OTHER_DISEASE and HAS_INMSUPR suprisingly has low correlation with IS_DEAD (0.06 and 0.05 respectively) and other attributes.

From here, these are our conclusions:

- When it comes to predicting mortality rate (IS_DEAD), AGE , IS_INTUBED and HAS_PNEUMONIA has the most impact due to their high correlation.
- Being COVID-Positive doesn't necessarily lead to higher rate of death, rather, it is as bad as being Obese or having other diseases.
- Most other diseases and Smoking doesn't affect the mortality rate as much as people might expect. ●

However, correlation does not imply causation.

1.2 Counting

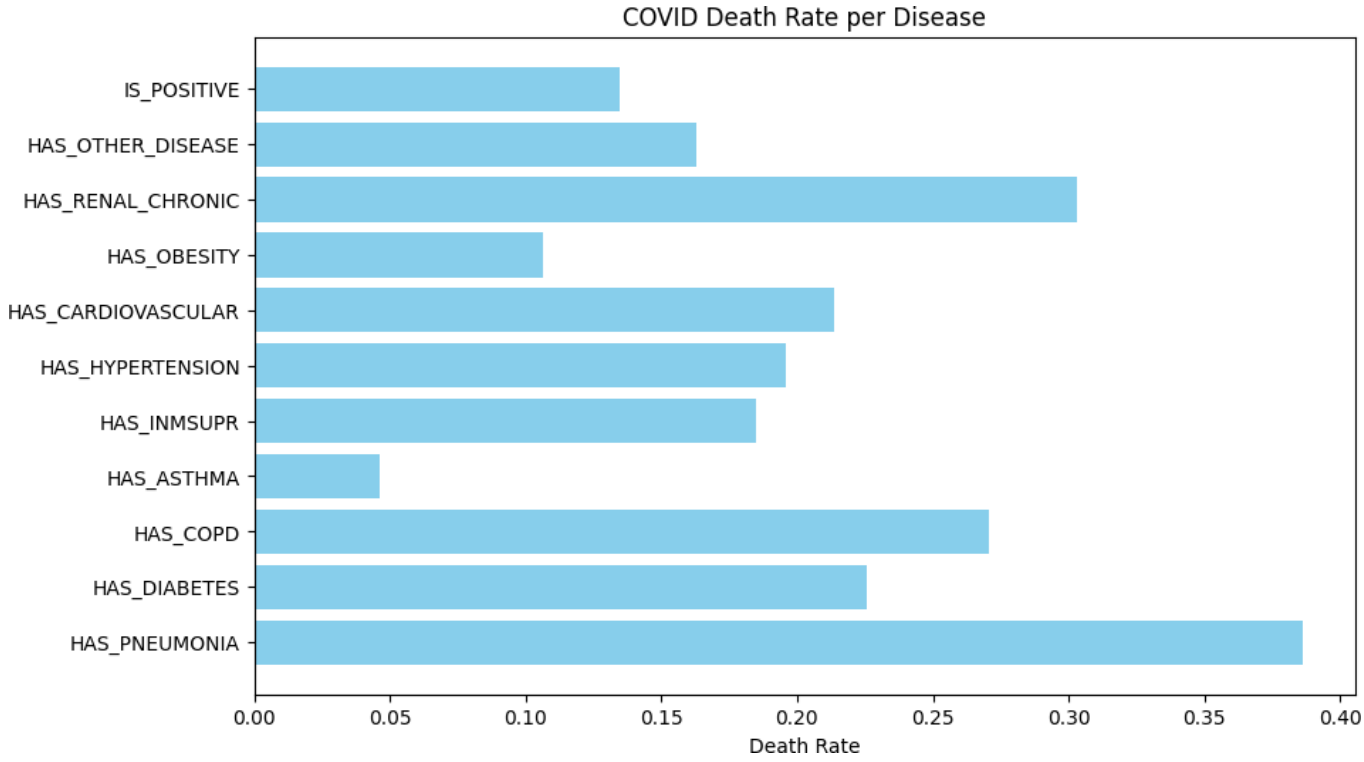

Figure 2: COVID Death rate by Disease

From figure 2, we can see that:

1. People who HAS_RENAL_CHRONIC and/or HAS_PNEUMONIA are more likely to die
2. HAS_ASTHMA and HAS_OBESITY has the lowest death rate.
3. IS_POSITIVE to COVID is the third lowest.

However, this may be due to the sampling, as indicated by the correlation shown before.

2. Basic Machine Learning (Logistic Regression, Naive Bayesian, Decision Tree, Ensemble)

1. **NB_gaussian**: This model has a high number of True Positives (253935) and True Negatives (15289), indicating that it is performing well in correctly predicting both classes. However, it also has a relatively high number of False Positives (22921), which means it is incorrectly predicting the positive class quite often.

2. **NB_complement**: This model also has a high number of True Positives (248674) and True Negatives (18543). However, it has a higher number of False Positives (28182) compared to NB_gaussian, which indicates it may be over-predicting the positive class.

3. **decision_tree**: This model has the highest number of True Positives (271267) among all models, and a relatively low number of False Positives (5589). However, it has a high number of False Negatives (11650), indicating it may be under-predicting the positive class.

4. **random_forest**: This model has a high number of True Positives (270999) and a relatively low number of False Positives (5857). However, it also has a high number of False Negatives (10922), similar to the decision_tree model.

5. **gradient_boosting**: This model has the highest number of True Positives (272492) and True Negatives (11007) among all models, indicating it is performing well in correctly predicting both classes. It also has the lowest number of False Positives (4364) and False Negatives (10493), indicating it has a good balance in predicting both classes.

6. **LR_liblinear** and **LR_newton-cg**: These two models have similar performance, with a high number of True Positives (272462 and 272441 respectively) and True Negatives (10691 and 10717 respectively). They also have a similar number of False Positives (4394 and 4415 respectively) and False Negatives (10809 and 10783 respectively).

In conclusion, the gradient_boosting model seems to perform the best among all models, as it has the highest number of True Positives and True Negatives, and the lowest number of False Positives and False Negatives. However, the choice of the best model also depends on the specific cost associated with False Positives and False Negatives, which can vary depending on the application. For example, in a medical diagnosis application, minimizing False Negatives (i.e., maximizing recall) might be more important than minimizing False Positives (i.e., maximizing precision). Therefore, you should choose the model that best fits your specific needs and constraints.

3. FFNN and RNN

Overfitting Prevention

Result analysis and Methods for Improvement in Accuracy