**Cognition and Machine Learning**

**COGS514**

**Final Project   
(Document Part)**

**2020/2021**

Name: Orçun Sami Tandoğan  
Number: 2535037

**Project Title:** Prediction of diagnosis of diabetes based on selected physiological attributes and comparison of Indian and Bangladesh dataset to open discussion to predict whether a patient has type I or type II diabetes.

**Methods:**

Linear Regression (Ridge)

Logistic Regression (Lasso, Ridge)

Support Vector Machine

Decision Trees

Random Forest  
**Protocol**

Each dataset is assessed by methods which are indicated.

Special preprocessing has been done for each dataset.

For each of them, accuracy and confusion matrixes are produced.

**Results**

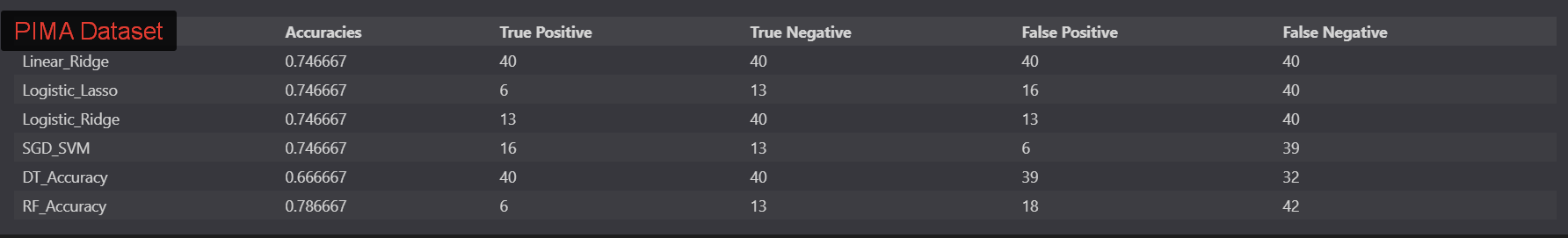


Figure 1.0 Confusssion Matrix of PIMA Indian Dataset

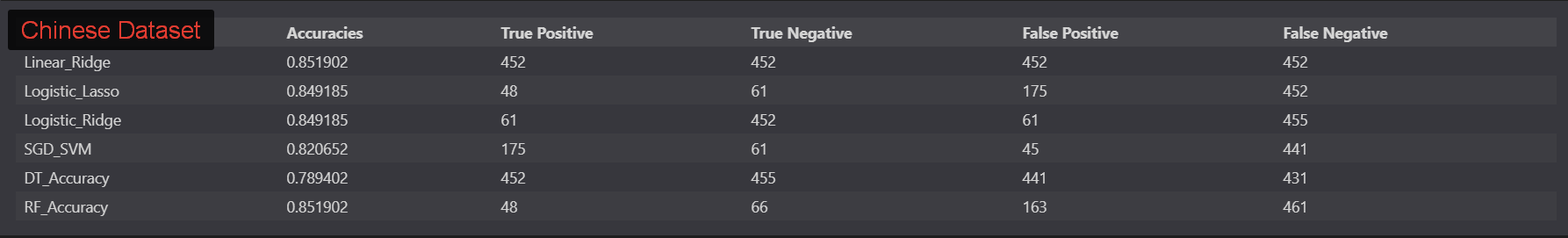


Figure 2.0 Confusssion Matrix of Chinese Dataset

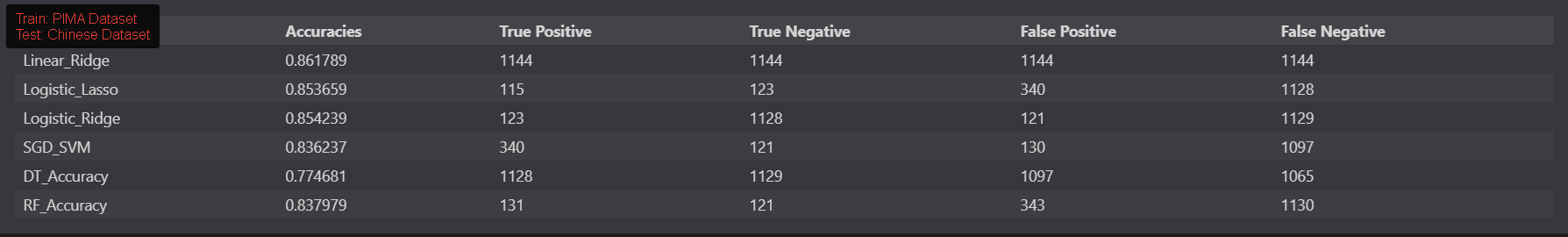


Figure 3.0 Confusssion Matrix of PIMA Trained On Chinese Dataset

**Discussion:**

I strongly recommend you to investigate python code file. This final project has been prepared like it is going to be uploaded to GitHub; so it will be easier to investigate project by ipynb file. This is only sum from my project.

PIMA and Chinese datasets consist Diabetes patients. There were various features in each of them and these features are assessed to find best methods to diagnose diabetes. Project was aiming to predict not only Diabetes, but also specify type of diabetes if it is Type I.

I added column to PIMA datasets for diagnosing Type I diabetes according to literature.

In both datasets, there are common features. After each datasets are assessed withing itself; both datasets are again preprocessed and only selected the common features. These features are used for training methods and predicting Chinese datasets from training PIMA datasets. Interested result is that training on PIMA datasets and using on Chinese datasets gave us more accurate results than assessing datasets withing itself. It must be reason there confusing and not clear part in the features of Chinese dataset like follow-up issue. Chinese datasets consists of various features which actually do not help us to predict so you are also going to find different preprocessing on Chinese dataset. My ipynb file is so long, but there are clear descriptions in there. If I would like to transfer them to here, it will be so messy so I just want to indicate what I have done in sum:

There were 0 values in Glucose, BP, Skin thickness ,Insulin & BMI which needs to be eliminated. Plotting made so easy to preprocess PIMA dataset; and the methods are applied which results in Figure 1.0.

However, Chinese dataset was not so easy as PIMA dataset. There are not clear parts like the article indicates diabetes has been shown after FPG level is higher than 7.0; but other diagnostic features for diabetes were not specified. Like censors are used, and other diagnosing feature which is `Diagnosing during follow up` are only `follow-up` information; and these follow-up`s were not suits with Final FPG levels during final visit. In this reason, I assume that first FPG levels have been assessed during follow up; so I created new column for diagnosing diabetes. There is also nice quantile parts in which the dataset is preprocessed as clear as it could be. Then methods are applied within the Chinese dataset.

After all, as it is previously mentioned, PIMA is trained for Chinese dataset with two differences:

First, only common features are used. Second, PIMA datasets` outcome are considered as only Type I diabetes; so after we test PIMA dataset on Chinese dataset, we can might have been assume that difference between of accuracy of PIMA datasets training on Chinese datasets, and Chinese dataset training within itself should have gave us type I accuracy. However, interesting thing is happened and accuracy of last work has been better than Chinese datasets within itself which can be seen in Figure 3.0. More detailed answers can be found in Jupiter file at attachments.

**References:**

1. Naz, H., & Ahuja, S. (2020, April 14). Deep learning approach for DIABETES prediction USING pima Indian dataset. Retrieved May 20, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7270283/
2. Type 2 diabetes. (2021, January 20). Retrieved May 20, 2021, from <https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193>
3. Dryad data -- Association of body mass index and age with incident diabetes in Chinese adults: A POPULATION-BASED cohort study. (n.d.). Retrieved May 21, 2021, from https://datadryad.org/stash/dataset/doi:10.5061/dryad.ft8750v