A study on the effect of Alanine Aminotransferase measurements (ALT) and genreal lifestyle factors on Liver Damage and its prediction.

Mei Huang (300504502), Naruebet Thanuwohan(300397282), Jack Blair (300507331)

2022-11-02

Executive Summary

Goal: predict liver damage, find most important data for predicting liver damage
Method: Using a K Nearest Neighbor *classification model*, GridSearch for *finding model settings*, and SelectKBest for *feature selection*
Findings: we can predict, to a high degree of accuracy (~97%), using features related to alanine aminotransferase measurements, alcohol consumption, diet, gender, and smoking data, liver damage, and to some extent, liver disease.

Background and Data Description

Sources: National Center for Health Statistics
We have combined data from the following data sets:

- BIOPRO (Biochemistry Profile) - This ALT variable is in this dataset.
- ALQ (Alcohol Use)
- DBQ (Diet Behavior)
- DEMO (Demographic Data)
- SMQ (Smoking - Cigarette Use)
- PAQ (Physical Activity)

We included data from the years 2011 to 2020.

Note: Many features of all the data sets have been excluded due to high amounts of missing data. It is key to note that many variables were removed so that the data sets could merge smoothly, essentially making sure we have the same variables for each yearly data set.

Questions: "Can we, to a degree of accuracy, predict liver damage using data on Alanine Aminotransferase measurements (ALT) as well as several other lifestyle factors such as alcohol use or physical activity?"

Method: We use ALT to label liver damage

	Age >= 21	Age 18-20
Male - IU/L	48	37
Female - IU/L	31	30

Classification from reference 1.

We have created a new binary variable "LiverDamage" with this classification, with the value 1 meaning that a person has liver disease, and the value 0 meaning that a person does not have liver disease.

	1	2	3	4	5	6
SEQN	109266	109292	109313	109317	109319	109323
LBXSATSI	15	21	24	9	21	25
Gender	2	1	1	2	1	1
Age	29	58	63	28	22	22
CountryOB	2	1	2	1	1	1
DietHealth	3	4	3	4	4	5
MilkConumption	2	3	3	1	1	2
TakeoutMeals	0	1	4	2	1	5
ReadyMeals	0	10	0	20	0	0
FrozenMeals	5	3	12	3	10	0
Cigs100	2	2	2	1	2	1
AvgAlc12Month	1	6	2	3	4	6
DrinkEveryday	2	2	2	2	2	2
VigWork	2	2	1	2	2	1
ModWork	2	2	1	2	1	1
TravelBikeWalk	2	2	2	2	2	1
VigRec	1	2	1	2	1	2
ModRec	2	2	1	2	1	1
MinSedentary	480	600	300	420	600	420
YEARS	17_20	17_20	17_20	17_20	17_20	17_20
LiverDamage	0	0	0	0	0	0

Description:

	Description
SEQN	Sequential Number (ID)
LBXSATSI	Alanine Aminotransferase (ALT) (U/L)
Gender	Gender of patient
Age	Age of patient
CountryOB	Country of birth.(1:Born in 50 US states or Washington, DC,2:Others,77:Refused,Don't Know)
DietHealth	How healthy is the diet(1:Excellent,2:Very Good,3:Good,4:Fair,5:Poor,7:Refused, 9:Don't know)
MilkConumption	Past 30 day milk product consumption(0:Never,1:Rarely,2:Sometimes,3:Often,4:Varied,7:Refused,9:Don't Know)
TakeoutMeals	Meals from fast food or pizza place
ReadyMeals	Ready-to-eat foods in past 30 days
FrozenMeals	Frozen meals/pizza in past 30 days
Cigs100	Smoked at least 100 cigarettes in life (1/Yes, 2:No, 7:Refused, 9:Don't know)
AvgAlc12Month	Days have 4/5 drinks - past 12 monthes
DrinkEveryday	Ever have 4/5 or more drinks every day
VigWork	Vigorous work activity(1/Yes, 2:No, 7:Refused, 9:Don't know)
ModWork	Moderate work activity(1/Yes, 2:No, 7:Refused, 9:Don't know)
TravelBikeWalk	Walk or bicycle(1/Yes, 2:No, 7:Refused, 9:Don't know)
VigRec	Vigorous recreational activities((1/Yes, 2:No, 7:Refused, 9:Don't know)
ModRec	Moderate recreational activities((1/Yes, 2:No, 7:Refused, 9:Don't know)

	Description
MinSedentary	Minutes sedentary activity
YEARS	Dataset years used
Liverdamage	Our target Variables(1:Yes, 0:No)

```
## 'data.frame':    5604 obs. of  21 variables:
##  $ SEQN          : int  122462 122742 121203 112021 91482 83951 119212 115470 86168 77539 ...
##  $ LBXSATSI      : int  17 29 11 17 18 16 14 16 15 15 ...
##  $ Gender        : int  2 1 2 2 2 1 1 1 1 2 ...
##  $ Age           : int  30 57 80 52 54 24 35 50 80 38 ...
##  $ CountryOB     : int  2 1 1 1 2 1 1 2 1 2 ...
##  $ DietHealth    : int  2 4 2 4 5 3 1 3 2 2 ...
##  $ MilkConumption: int  1 1 2 2 3 2 1 0 3 0 ...
##  $ TakeoutMeals  : int  1 19 0 2 7 8 2 0 2 0 ...
##  $ ReadyMeals    : int  5 0 0 0 3 2 2 0 1 0 ...
##  $ FrozenMeals   : int  0 0 0 0 0 7 1 2 0 4 ...
##  $ Cigs100       : int  1 2 2 2 1 1 1 2 1 1 ...
##  $ AvgAlc12Month : int  1 1 1 2 2 4 1 6 1 2 ...
##  $ DrinkEveryday : int  2 2 2 2 2 2 1 2 2 2 ...
##  $ VigWork       : int  2 1 2 2 1 1 2 1 2 1 ...
##  $ ModWork       : int  2 1 2 1 2 1 2 1 1 1 ...
##  $ TravelBikeWalk: int  2 2 1 2 2 1 2 1 2 1 ...
##  $ VigRec        : int  1 2 2 2 2 1 1 1 2 2 ...
##  $ ModRec        : int  2 1 2 1 2 1 2 1 1 1 ...
##  $ MinSedentary  : int  360 420 300 300 480 240 360 240 420 600 ...
##  $ YEARS         : chr  "17_20" "17_20" "17_20" "17_20" ...
##  $ LiverDamage   : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##
##    0    1
## 5064  540
```

The combined data set we generated has 8006 rows with 21 variables, we have 13 categorical variables, and 8 numeric variables. There is no missing data, since we have deleted all the missing data in the process of combining data so as to allow us to merge the data properly. We have split the data with a 70/30 training/test split which has resulted in our training data set having a total of 5604 rows, with 540 of those patients diagnosed as having liver damage.

Ethics, Privacy and Security

Ethics

The National Health and Nutrition Examination Survey (NHANES) has ensured that Personally Identifiable Information (PII) has been removed from the data sets, with the main concern regarding privacy being the possibility of de-anonymization. The possibility of de-anonymization is heightened with the wide breadth of data present on individuals, including behavioral and demographic (National Center for Health Statistics, 2021). Therefore, one of the key ethical considerations was to ensure that any linkage between data sets or analysis could not inadvertently be used for the de-anonymization of an individual.

Additionally, as we are dealing with health and medical data and information, we have been careful with our analysis and have ensured that the analysis is conducted with a focus on equity, e.g., not averaging across genders, as the medical and health data pertaining to men is not necessarily the same as for women, etc.

Privacy

We have evaluated the security of our project's tech stack, data, and information according to the CIA triad framework: confidentiality, integrity, and availability.

Confidentiality

Access to our code base was controlled using GitHub. The remote GitHub repository was kept private, with access controlled by the repository owner to ensure only approved people, who were team members, were to have access.

All the team members had GitHub accounts which were secured using 2FA authentication and a minimum 8 alphanumeric character length password with one capital, number, and special character. This was to ensure that access to team members' accounts, and thereby the team repository, remained secure.

Additionally, all team members were accessing the local and remote repositories using approved devices to ensure clean devices with up-to-date security policies and the latest firmware updates. These devices were used strictly for work purposes (e.g., this analysis). Each team member device was secured using BitLocker encryption and password-controlled account access.

Team communication was conducted via email and Discord. Access to the team channel is invite-only and approval is controlled by a team administrator. The Discord of each team member was set up with 2FA.

Integrity

We have imported our data directly from the NHANES website, the data is then saved as .rds files, to take a "snapshot" of the data at that period of time. This is important for the reproducibility of our analysis and results, if the NHANES data hosted on the CDC website were to be changed in any way or become compromised.

In addition, our use of git version control also provides logs to ensure non-repudiation (i.e., the inability to deny). The tracking of every version of our scripts ensures that any changes to the data are kept and maintained.

Availability

Our project work was conducted using git version control to ensure traceable and restorable changes to our code base. Additionally, all the local git repositories are backed up in real-time with Microsoft's OneDrive cloud service technology. Our remote repositories are stored in the cloud with GitHub. This ensures that if there was hardware corruption, that the files would still be recoverable using our Cloud service.

Exploratory Data Analysis

	minimum	first quartile.25%	median	third quartile.75%	maximum	IRQ	sd	skewness	kurtosis
SEQN	62169	75997	87485	114317	124822	38320	20313	0	2
LBXSATSI	4	15	20	28	319	13	18	5	42
Gender	1	1	1	2	2	1	0	0	1
Age	20	31	44	60	80	29	17	0	2
CountryOB	1	1	1	1	77	0	1	59	4072
DietHealth	1	2	3	4	9	2	1	0	3
MilkConumption	0	1	2	3	4	2	1	-1	2
TakeoutMeals	0	0	1	3	21	3	3	3	14
ReadyMeals	0	0	0	2	90	2	6	6	63
FrozenMeals	0	0	0	2	90	2	6	6	55
Cigs100	1	1	2	2	9	1	1	1	22
AvgAlc12Month	1	1	2	3	15	2	2	2	10

	minimum	first quartile.25%	median	third quartile.75%	maximum	IRQ	sd	skewness	kurtosis
DrinkEveryday	1	2	2	2	9	0	0	1	59
VigWork	1	1	2	2	2	1	0	-1	2
ModWork	1	1	2	2	9	1	1	1	16
TravelBikeWalk	1	1	2	2	9	1	0	0	14
VigRec	1	1	2	2	2	1	0	-1	2
ModRec	1	1	2	2	9	1	1	1	16
MinSedentary	1	180	300	480	600	300	158	0	2
LiverDamage	0	0	0	0	1	0	0	3	8

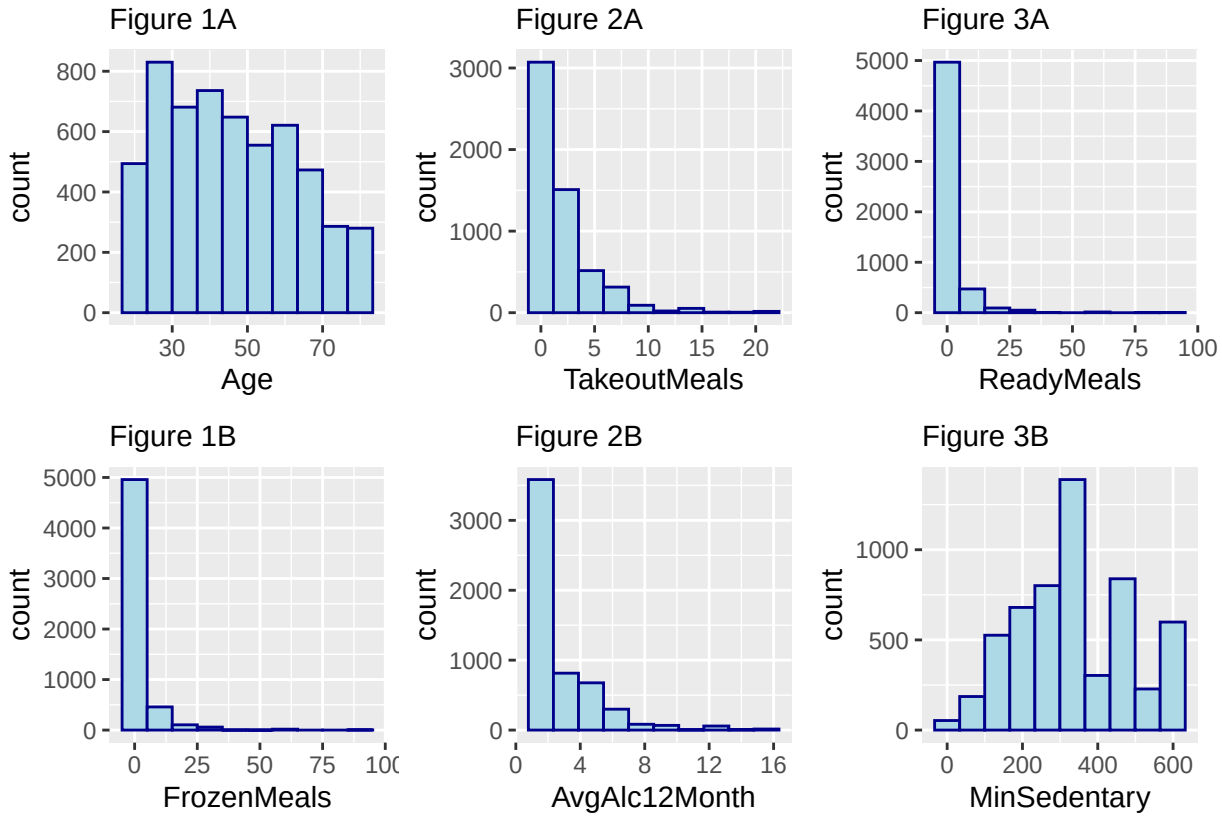
This data set only contains Age above 20, so it is only suitable to predict Age between 20-80.

	LiverDamage	ModRec
LBXSATSI	0.688	0
ModWork	0.000	1

There are only two correlations as can be seen from the table above. LiverDamage is based off of LBXSATSI(ALT) and age, so this correlation is to be expected. ModRec and ModWork is slightly surprising but expected in a way, as they are similar but different variables.

Our data set is quite small and we still have no idea which feature might be useful, so the following visualization plots will be for all the features we have

Visualization of numeric variables distribution



Figures 1A(Age) and 3B(MInSedentary) are spread across all the data range, while all the other numeric variables are tailed to the right.

Visualization of categorical variables distribution

Figure 1A

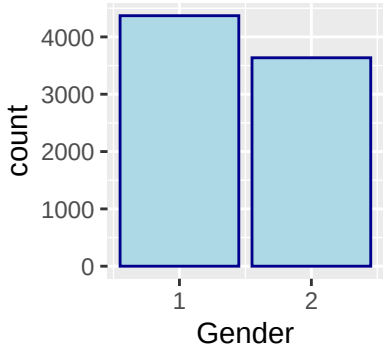


Figure 2A

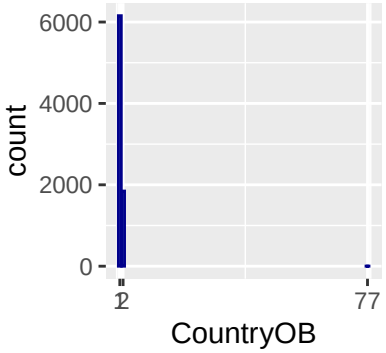


Figure 3A

Figure 1B

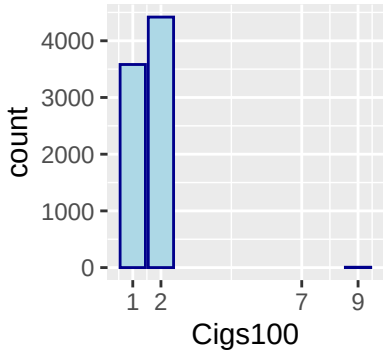


Figure 2B

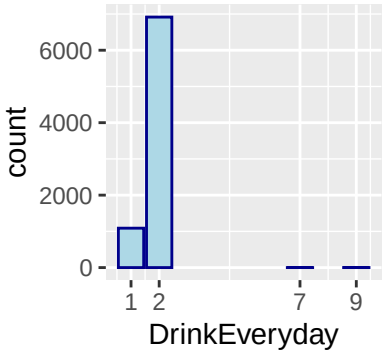


Figure 3B

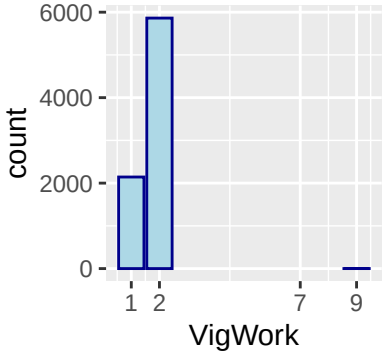


Figure 1C

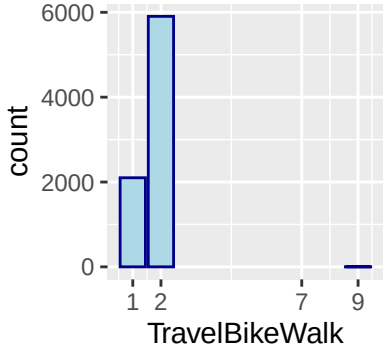


Figure 2C

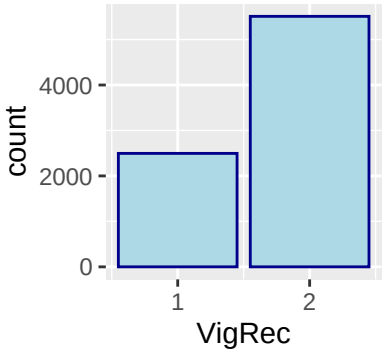


Figure 3C

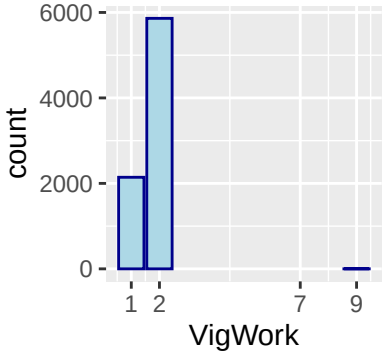


Figure 1D

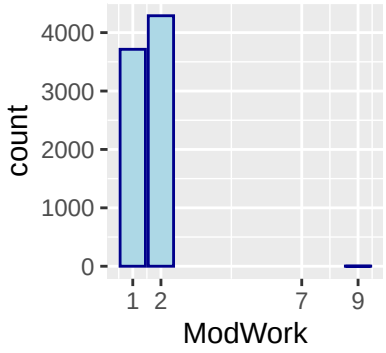


Figure 2D

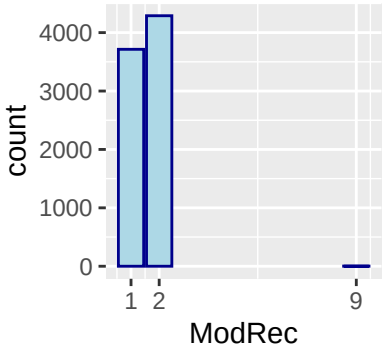
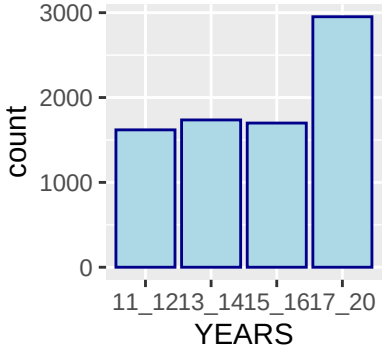
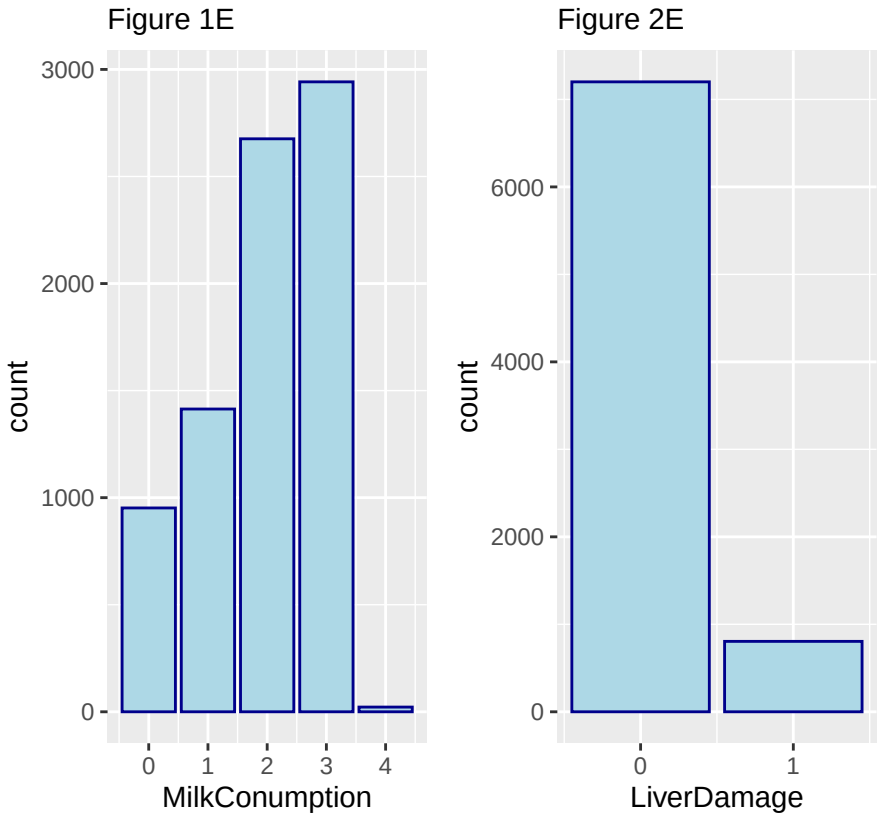


Figure 3D





We can see from Figure 1A, we have slightly more males than females. From figure 3A we can see that most people have "Good" health (group 3). From figure 3D we can see that most of the data is evenly spread over the years, except the |17_20" group, which is actually around 4 years' worth of data, instead of the standard two-year periods we see.

Then we have a look at if there is any different distribution for different categorical variables.

1 = Male, 2 = Female
We can see that there is about the same percentage of each gender that have/do not have liver damage.

Gender	LiverDamage	n	percentage
1	0	3939	90.14
1	1	431	9.86
2	0	3262	89.71
2	1	374	10.29

1 = United states, 2 = Other

CountryOB	LiverDamage	n	percentage
1	0	5593	90.85
1	1	563	9.15
2	0	1606	86.95
2	1	241	13.05

CountryOB	LiverDamage	n	percentage
77	0	2	66.67
77	1	1	33.33

How healthy is the diet(1:Excellent,2:Very Good,3:Good,4:Fair,5:Poor,7:Refused, 9:Don't know)
With this, we can see that as diet health gets worse, the percentage of people who have liver damage increases. i.e., Excellent = 6.7%, Very good = 7.48%, Good = 9.67%, Fair = 12.78%, Poor = 14.14%

DietHealth	LiverDamage	n	percentage
1	0	515	93.30
1	1	37	6.70
2	0	1521	92.52
2	1	123	7.48
3	0	3016	90.33
3	1	323	9.67
4	0	1734	87.22
4	1	254	12.78
5	0	413	85.86
5	1	68	14.14
9	0	2	100.00

Past 30 day milk product consumption(0:Never,1:Rarely,2:Sometimes,3:Often,4:Varied,7:Refused,9:Don't Know) We can see from the table that those who often drink milk or drink no milk at all are less likely to have liver damage.

MilkConumption	LiverDamage	n	percentage
0	0	871	91.49
0	1	81	8.51
1	0	1252	88.54
1	1	162	11.46
2	0	2405	89.87
2	1	271	10.13
3	0	2655	90.24
3	1	287	9.76
4	0	18	81.82
4	1	4	18.18

Smoked at least 100 cigarettes in life (1/Yes, 2:No, 7:Refused, 9:Don't know) By not so much, the group that has smoked at least 100 cigarettes have a higher percentage with liver damage (10.25% vs 9.91%)

Cigs100	LiverDamage	n	percentage
1	0	3215	89.75
1	1	367	10.25
2	0	3980	90.09
2	1	438	9.91
9	0	6	100.00

Ever have 4/5 or more drinks every day The group that had 4/5 drinks per day has 12.5% with liver damage whereas the group that did not have 4/5 drinks per day had 9.67% with liver damage.

DrinkEveryday	LiverDamage	n	percentage
1	0	952	87.50
1	1	136	12.50
2	0	6246	90.33
2	1	669	9.67
7	0	1	100.00
9	0	2	100.00

Vigorous work activity(1/Yes, 2:No, 7:Refused, 9:Don't know) The group that did vigorous work had 11.25% with liver damage, whereas the group that did not do vigorous work had 9.62% with liver damage.

VigWork	LiverDamage	n	percentage
1	0	1902	88.75
1	1	241	11.25
2	0	5298	90.38
2	1	564	9.62
9	0	1	100.00

Walk or bicycle(1/Yes, 2:No, 7:Refused, 9:Don't know) The group that did walk/bike had 10.38% with liver damage, whereas the group that did not walk/bike had 9.94% with liver damage.

TravelBikeWalk	LiverDamage	n	percentage
1	0	1882	89.62
1	1	218	10.38
2	0	5318	90.06
2	1	587	9.94
9	0	1	100.00

Vigorous recreational activities((1/Yes, 2:No, 7:Refused, 9:Don't know) The group that did vigorous recreational activities had 9.01% with liver damage, whereas the group that did not do recreational activities work had 10.53% with liver damage.

VigRec	LiverDamage	n	percentage
1	0	2271	90.99
1	1	225	9.01
2	0	4930	89.47
2	1	580	10.53

There seems to be a general increase of people with liver damage over the years until the year period "17_20", which sees a decrease.

YEARS	LiverDamage	n	percentage
11_12	0	1469	90.74
11_12	1	150	9.26
13_14	0	1545	89.00
13_14	1	191	11.00
15_16	0	1510	88.88

YEARS	LiverDamage	n	percentage
15_16	1	189	11.12
17_20	0	2677	90.68
17_20	1	275	9.32

Analysis Results

I decided to create a K Nearest Neighbors classification model using a grid search method to find the optimal model. This allows me to find out the significance of variables within the data set as well as providing a method of prediction which are the two objectives of this study.

Model Creation

Step 1: splitting the data.

We have split the data 70/30, meaning we have 70% training data and 30% test data.

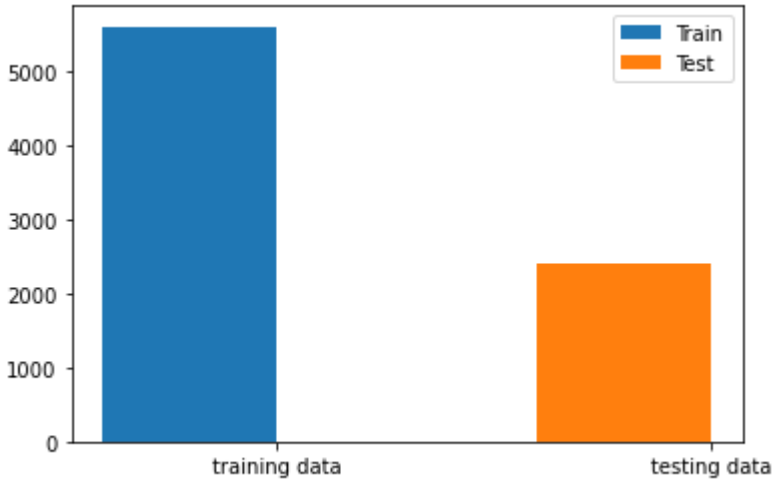


figure 3a

Step 2: data transformation/manipulation.

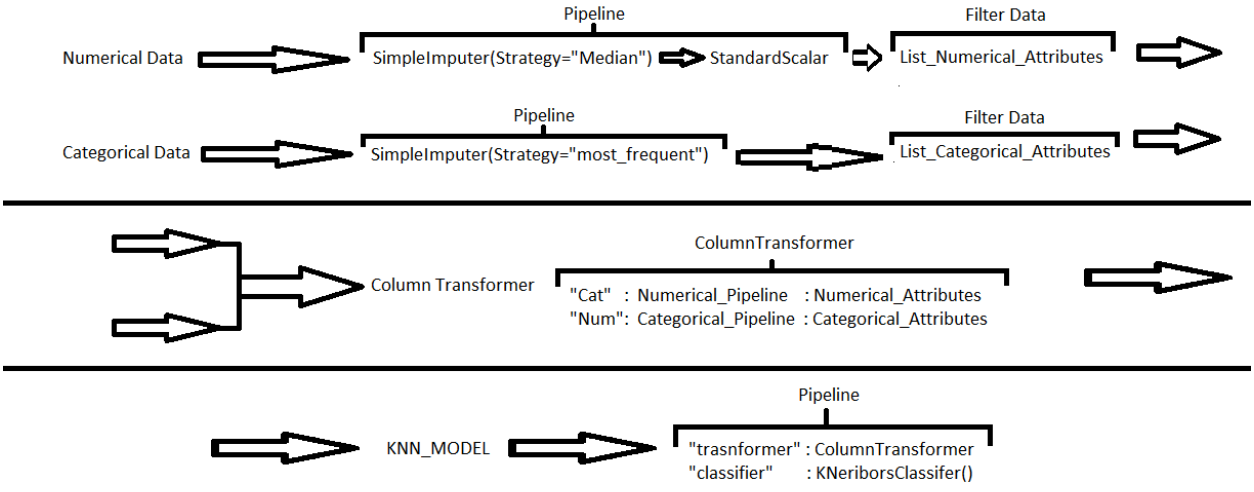
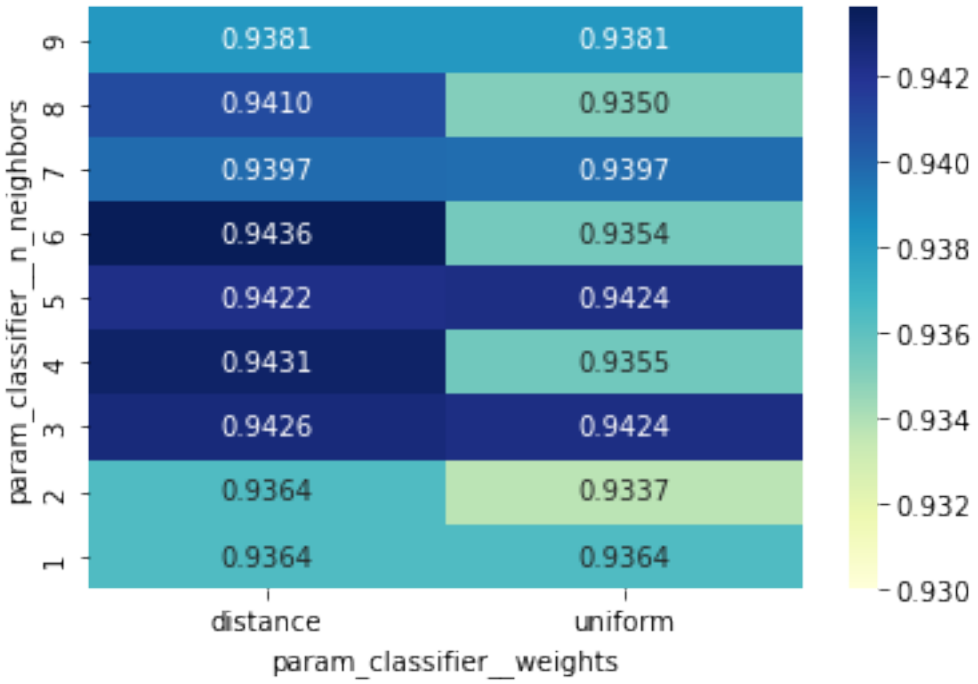


figure 1a

Step 3: grid search.

Using sklearn GridSearchCV, I did a grid search of 1:10 neighbors and the weights “uniform” and “distance”. The results can be seen below:



As we can see from the heat map above, our best model is formed with 6 Neighbors and a classifier weight “distance”, which achieves a model accuracy of 94.36%.

Feature Selection

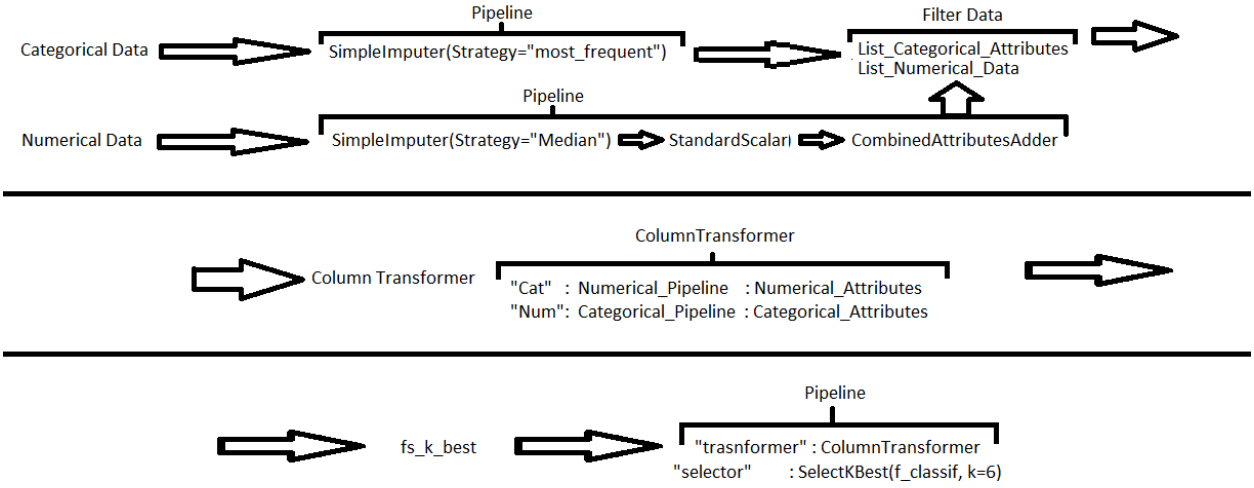


figure 2a

Using the SelectKBest method we are able to find out which features are statistically significant, the results:

As we can see from figure 4a, the statistically significant features based on a significance level of 0.05, are: LBXSATSI, AvgAlc12Month, Gender, DietHealth, TakeoutMeals and Cigs100. We will construct a new model using only these features and see if we cannot improve our model.

Feature Selected Model

We will be using the same KNN Model / pipeline structure as shown in figure 1a but we will remove all non-significant features from the data set.

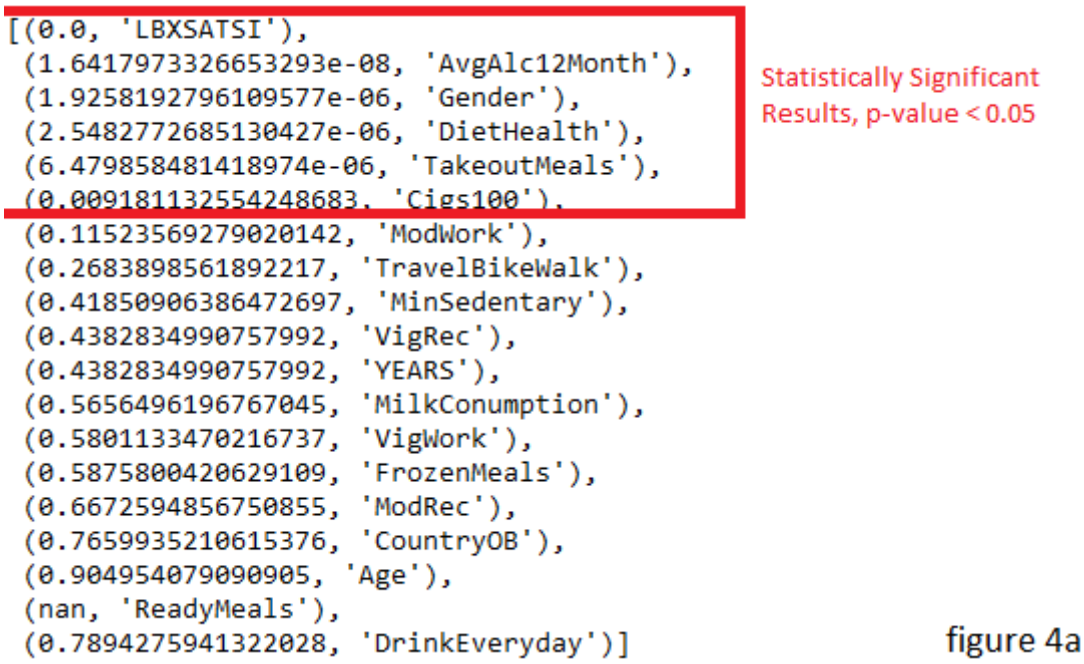


Figure 1: Feature selection results.

Old data set shape - unsplit data: (8006, 20), X_train data: (5604, 19)
New data set shape: - unsplit data: (8006, 7), X_train data: (5604, 6)

We have removed: [‘ModWork’, ‘TravelBikeWalk’, ‘MinSedentary’, ‘VigRec’, ‘YEARS’, ‘MilkConumption’, ‘VigWork’, ‘FrozenMeals’, ‘ModRec’, ‘CountryOB’, ‘Age’, ‘ReadyMeals’, ‘DrinkEveryday’]

Using the same grid search pipeline as shown in figure 2, we yield an overall 97.73% model accuracy, with our best parameters being 4 neighbors and classifier weight “distance”, as can be seen on the resulting heat map above.

Model accuracy: Test set

For our baseline model (all features included) we have a test set accuracy of 94.09%

For our feature selected model (limited features included) we have a test set accuracy of 97.59%

Notes and possible biases

It should be noted that in the data set there is a total of 4370 males, and 3636 females. This may result in slightly less resilient results when predicting liver damage with female data.

Conclusions

Our original question is:
“Can we, to a degree of accuracy, predict liver damage using data on Alanine Aminotransferase measurements (ALT) as well as several other lifestyle factors such as alcohol use or physical activity?”. Through the use of a KNN classification model, using GridSearchCV to find the optimal model settings and SelectKBest to select the best features, we have found the answers to our question. We can, to a high degree of accuracy, predict liver damage using ALT/lifestyle data. The best features included in the final model were LBXSATSI, AvgAlc12Month, Gender, DietHealth, TakeoutMeals and Cigs100. Alcohol consumption, diet, gender, and smoking data can all be used along with ALT measurements to predict liver damage, and to an extent, liver disease.

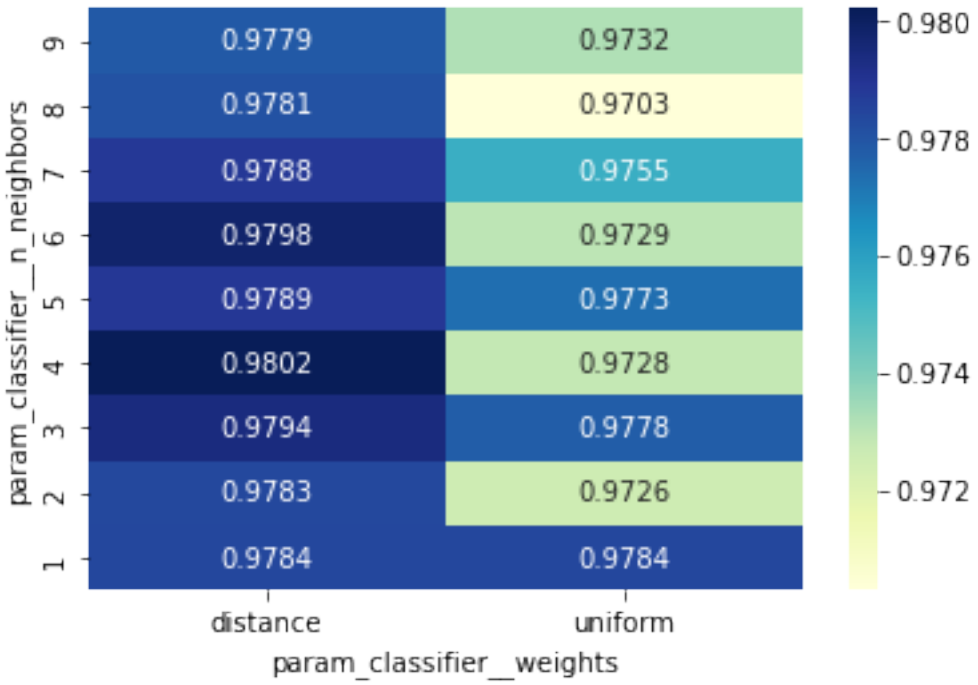


Figure 2: A heatmap of the KNN GridSearchCV Results.

With more testing and refining, this model could be recommended to laboratories testing ALT. With permission, whenever a patient takes an ALT test, the laboratory could use this model, along with the relevant data mentioned above and the ALT test results to predict any major problems with liver damage, and to an extent, liver disease. This could help prevent liver damage/disease and could improve the identifiability of liver damage and or liver disease.

References

reference 1:
Cave, M., Appana, S., Patel, M., Falkner, K. C., McClain, C. J., & Brock, G. (2010). Polychlorinated Biphenyls, Lead, and Mercury Are Associated with Liver Disease in American Adults: NHANES 2003–2004. Environmental Health Perspectives, 118(12), 1735–1742. https://doi.org/10.1289/ehp.1002720

Individual Contributions

- Jack: Combing data set, doing part of EDA Report, data analysis - Refactored all parts of project,
- Mei: Search and Label data set and doing EDA Report
- Naru: Code compiling and Ethics, Privacy, Security part.