Insomnia among university students

Data science in action bonus project

Lorenzo Baglini, Edoardo Borriello, Giacomo Flores

The purpose of the survey we have carried out is to research what variables may have the greatest impact on problems of insomnia. We thought it was important to go and analyze this correlation because from the investigation we have carried out it has resulted as many of our colleagues suffer from this problem. With the data at our disposal, we have therefore tried to draw some results by comparing various causes that could give rise to the issue.

As first thing we decided to stipulate, with other guys attending the course of Data Science and Management, a file of questions to answer, going to create a database on which to base our research.

We managed to get 26 answers for each question, obtaining a more or less complete dataset that allowed us to move forward in the project.

After the data collection phase, we examined the file and made a selection of the questions, eliminating the ones that were not useful for our research.

After an initial sorting, we ranked the questions according to those we felt were most or least useful.

Finally, we eliminated the less Useful questions and the open questions, creating a file of 28 columns.

First of all, after importing the dataset, we divided it between a training part and a test part. Afterwards we have applied the feature scaling to avoid bias towards variables with higher magnitude in order to use the method of K-Nearest Neighbors, which, with the data in our possession, we thought was the most reliable one.

We have then chosen five combinations of questions comparing the values concerning the accuracy of the model in the training part and in the test part.
We decided to use this type of analysis because, with the data in our possession, we thought it was the most reliable and accurate method to analyze and study them.
Finally we have chosen the combinations of questions that we thought more suitable to our analysis, and we have then selected them taking in consideration the values regarding the accuracy of the model both in the training and in test part.

The combinations are respectively composed by the following questions:

1. Do you live with your family?
2. Do you watch TV series or movies while you eat
3. How much do you care about nutrition and having a healthy lifestyle?
4. How many hours do you spend in hobbies per day?
5. How many hours on average do you spend using your phone in a day?
6. Do you consider yourself kind of bored?

Code description

To begin, we imported the libraries needed to study our dataset through K-NN analysis method.

```
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

Next, we imported the excel spreadsheet containing our cleaned dataset.

```
# Importing the dataset
file_name = "risposte.xlsx"
sheet = "risp"
df = pd.read_excel(io=file_name, sheet_name=sheet, header=0)
```

Then, using astype(), we converted the data type of the data column we needed for our analysis from string to Int, and more specifically to binary.

```
#Binarizing Responses
df["Insomnia"] = (df["Insomnia"] == "Yes").astype(int)
df["Gender"] = (df["Gender"] == "Male").astype(int)
df["In family"] = (df["In family"] == "Yes").astype(int)
df["Bored"] = (df["Bored"] == "Yes").astype(int)
df["Leave phone"] = (df["Leave phone"] == "Yes").astype(int)
df["sp satisfied"] = (df["sp satisfied"] == "Yes").astype(int)
df["movies or series"] = (df["movies or series"] == "Movies").astype(int)
df["streaming or cinema"] = (df["streaming or cinema"] == "Streaming").astype(int)
df["while eat"] = (df["while eat"] == "Yes").astype(int)
df["correct raccomandation"] = (df["correct raccomandation"] == "Yes").astype(int)
df["alone or not"] = (df["alone or not"] == "Alone").astype(int)
df["listening h"] = (df["listening h"] == "Yes").astype(int)
df["unsubscribed"] = (df["unsubscribed"] == "Yes").astype(int)
df["digital platforms"] = (df["digital platforms"] == "Yes").astype(int)
```

Finished the introductory part, using the command .iloc, we have selected the different pairs of data column assigning them to the variable X chosen as predictors for the model, we then associated the couples one after the other to the target variable y which represents the answers to the question "Have you got insomnia problems?".

```
# Living with own family + Watching series while eating
X = df.iloc[:,[5, 21]].values
y = df.iloc[:,1].values
```

After that, we split the data into training and test.
In numerical terms, we determined that 25% of the data should go to form the test part and, consequently, the remaining 75% was allocated to the training part.
Furthermore, we set the random state equal to 0, in order to avoid the data to be mixed improperly.

```
# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)
```

At this point, it was necessary to standardize the values contained in the dataset in order to avoid the possibility of errors caused by bias towards variables with higher magnitude

For example, the datas regarding the age of the survey participants would tend to create bias in the results, since they are values between 21 and 28, unlike all the other responses that contain much lower numbers.

```
# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

We then used KNeighborsClassifier changing the default value assigned to the n_neighbors parameter (5) in relation to the training part, in order to eliminate as much as possible wrong values and predictions. The best accuracy was obtained by setting this value equal to 1.

```
# Fitting K-NN to the Training set
from sklearn.neighbors import KNeighborsClassifier
IM = KNeighborsClassifier(n_neighbors = 1, metric = 'minkowski', p = 2)
IM.fit(X_train, y_train)
```

We then used x to predict the test results, which was then useful in calculating the accuracy of our analysis

```
# Predicting the Test set results
y_pred = IM.predict(X_test)
```

Through the Confusion Matrix we were able to assess whether the predictions were indeed accurate.

The results showed that, out of seven total observations, four were labeled as positive both by the predicted class and by the true class and, for the other three, their nature of negativity was confirmed in both classes of verification.

As mentioned before, value 1 referred to the number of neighbors was the best one from the point of view of accuracy, since with other values such as 3, 5 or 7 there was always some discordant result between predicted and true class.

```
# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
```

Next, we wrote the code for the creation of the graphs related to the training part. First of all we have set the coordinates to follow for the creation of the chart and then we have chosen red and green as colors to distinguish the two parts. So, through a for loop, we created the graphs assigning to each one of them a name and to the axis x and y the relative questions from the data columns to compare.

```
# Visualising the Training set results
from matplotlib.colors import ListedColormap
X_set, y_set = X_train, y_train
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:, 0].max() + 1, step = 0.01),
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:, 1].max() + 1, step = 0.01))
plt.contourf(X1, X2, IM.predict(np.array([X1.ravel(), X2.ravel()]).T).reshape(X1.shape),
             alpha = 0.75, cmap = ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColormap(('red', 'green'))(i), label = j)
plt.title('Insomnia (Training set)')
plt.xlabel('Living with own family')
plt.ylabel('Watching series while eating')
plt.legend()
plt.show()
```

After that we used .score to calculate the accuracy of our training part by comparing the X_train and the y_train, this allowed us to select only the pairs of data columns with a level of accuracy between 70% and 90%

```
#Calculate training accuracy
accuracy_train = IM.score(X_train, y_train)
print(accuracy_train)
```

Afterwards, as previously done for the training part, we wrote the code for the creation of graphs, changing the colors to yellow and green to distinguish the two sets of data.

```
# Visualising the Test set results
from matplotlib.colors import ListedColormap
X_set, y_set = X_test, y_test
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:, 0].max() + 1, step = 0.01),
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:, 1].max() + 1, step = 0.01))
plt.contourf(X1, X2, IM.predict(np.array([X1.ravel(), X2.ravel()]).T).reshape(X1.shape),
             alpha = 0.75, cmap = ListedColormap(('orange', 'blue')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColormap(('orange', 'blue'))(i), label = j)
plt.title('Insomnia (Test set)')
plt.xlabel('Living with own family')
plt.ylabel('Watching series while eating')
plt.legend()
plt.show()
```

As before for the training set, we calculated the test accuracy by comparing the x_test and the y_test. We decided to examine, as far as possible, the data with the test accuracy that is closest to the training set.

Unfortunately, given the scarcity of data collected this was not always possible, but in many cases, as we will see later with the graphs, the comparison led to good results in terms of completeness and accuracy of the analysis.

```
#Calculate test accuracy
accuracy_test = IM.score(X_test, y_test)
print(accuracy_test)
```

Experimental design

We decided to conduct our research on the data of insomnia in the boys of our age because we believe it can be a problem really common among us. Starting from the initial dataset, in fact, we began to think which question, among the many, could be the one on which to build our analysis and, although more than one could have served as the basis for our research, in the end, as mentioned above, our choice fell on insomnia.

As mentioned in the introduction, the basis of everything was the creation of a dataset with responses from us students in the course. Although not very rich in responses, the dataset in question turned out to be sufficient to conduct a fairly comprehensive research on the above topic, through the use of the K-NN we used to study the data.

Of course, from the starting dataset we had to eliminate many of the questions that had been collected as they could not contribute in any way to a meaningful analysis on the problem of insomnia. To choose our data, therefore, we decided to start from the problem and think about what, out of the data at our disposal, could have triggered it, such as family living, diet or time of phone use. To choose the data to be analyzed we also had to stick, as explained earlier in the description of the code, to the accuracy of the training and testing so in some cases we were forced to ignore data that could characterize the problem of insomnia but that when compared in the code, did not give the desired result.

Results

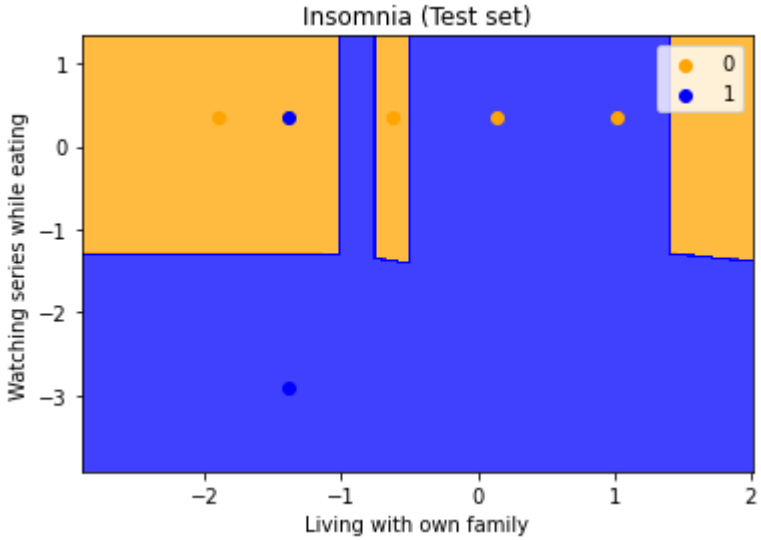
To obtain the best results that could explain, and highlight, a correlation with the occurrence of insomnia in the students, we tested more than twenty pairs of answers to questions related to the issue.
Below are graphs relating first to the training portion then to the testing portion of each of the five pairs of variables that we identified as the best predictors for the chosen target variable.

In order to establish which were the best predictors, we decided to consider those that presented a level of accuracy in the training set between 0.7 and 0.9 but, not having a large amount of responses for each question, we made some small exceptions. The reason behind this choice is explained by the fact that, having few values, the accuracy index followed very wide ranges, that is, about 0.12 between one value and the next.

After identifying the predictors to be analyzed, we selected the five best models based on the distance presented between the values on the accuracy indices obtained in the training set and in the test set.

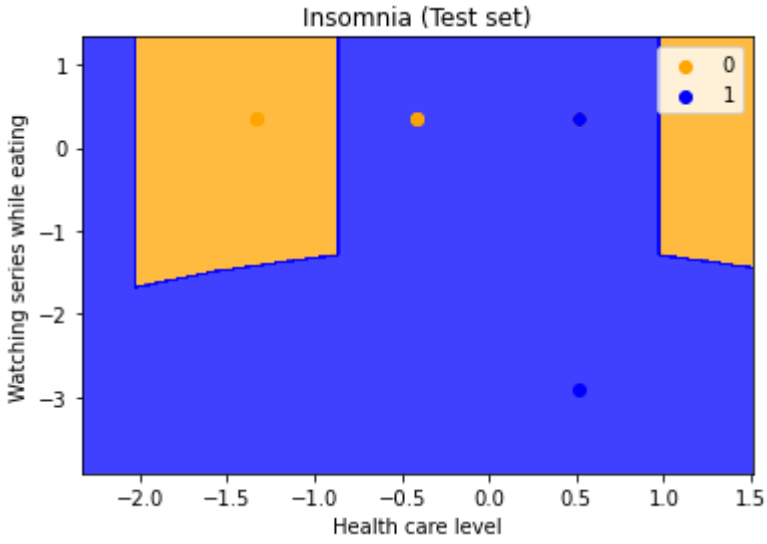
Among the five models chosen as the best, the one obtained by taking as predictors the variable that insists on living with their own family and the variable that investigates whether or not they watch TV series and movies even while eating, is the most fitting.

It is possible for us to arrive at this statement given that, the model in question, presents a good value of accuracy in the training set (0.79), and a value not too distant for the test set (0.57), considering that the values of the latter increased by about 0.07 between one and the other.

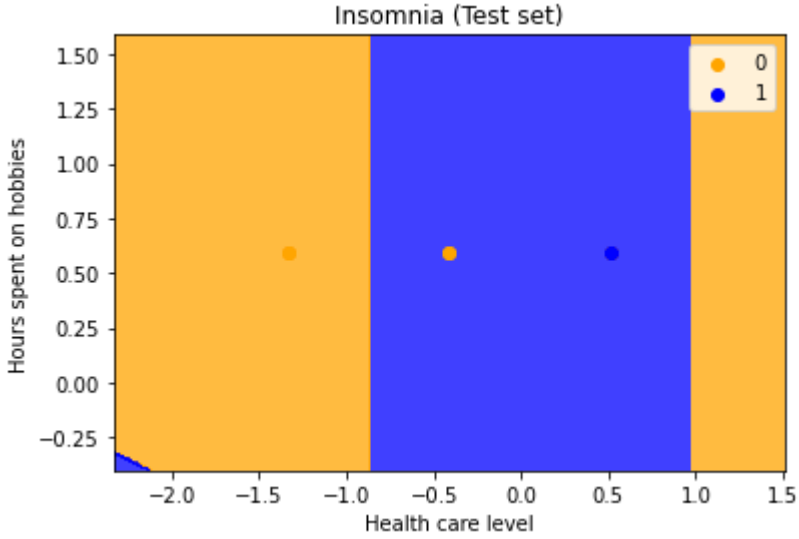


In this case, instead, we compared the level of care and attention to one's own health and whether or not to watch TV series and movies even while eating, obtaining an even higher result in the training set (0.89), although the test set remains unchanged (0.57) compared to the one seen previously. We note, therefore, a wider range between the two data that make them the less accurate model than the first seen.

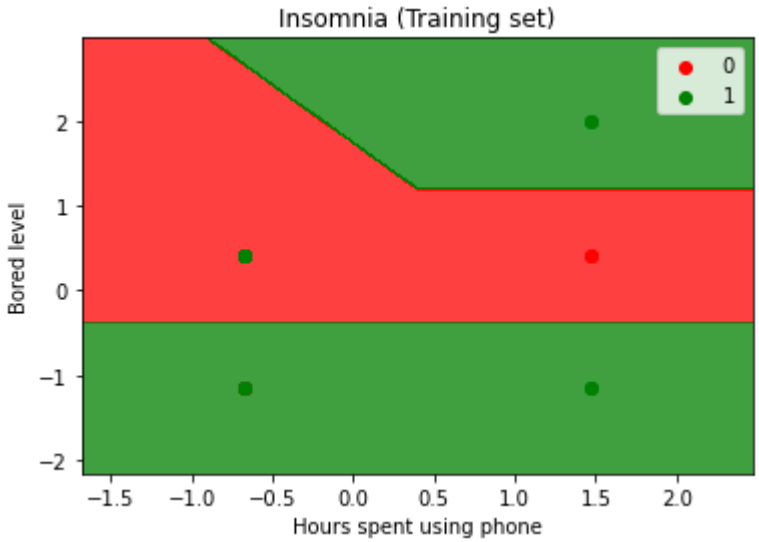
Despite this, we observed that there was again a high correlation between insomnia and the chosen variables.



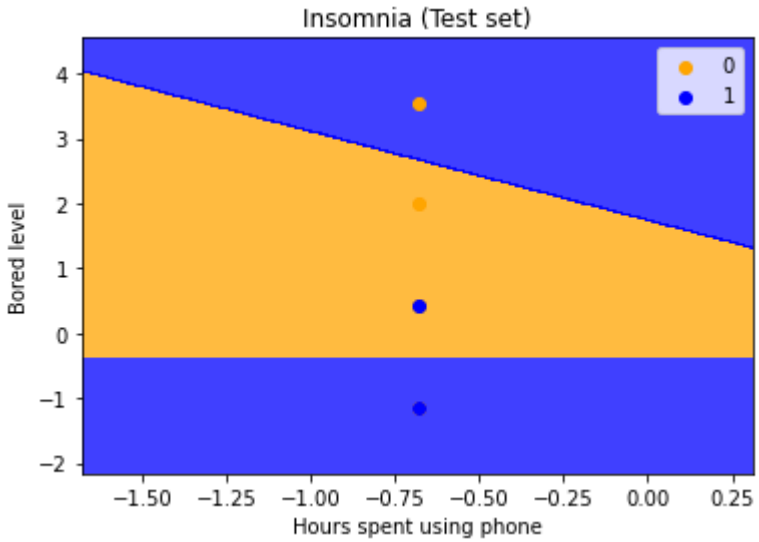
In this instance we always find the level of attention to health which, even when compared to another variable such as hours spent on hobbies, returns the same result in the training set and in the test set.



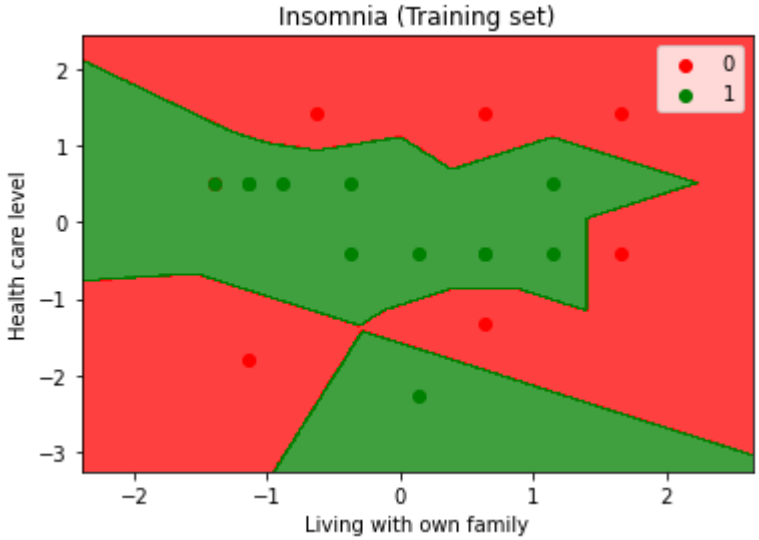
Regarding the model created using hours spent using the cell phone and the level regarding how bored one feels as predictors, it is possible to say that, given the assumptions about data sparsity, 0.63 can be considered an acceptable training set accuracy value.



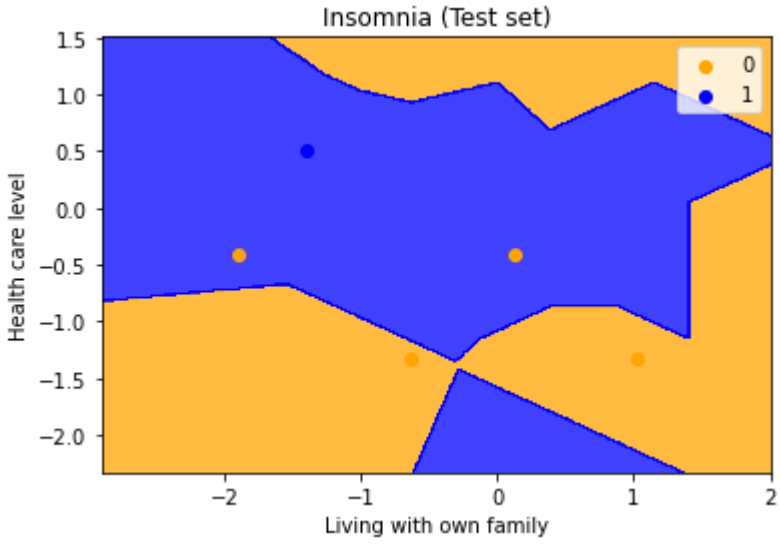
Hypothesis strengthened by the fact that there is, then, a small difference with the accuracy value of the test set (0.57).



The last model examined shows a difference in the training set. Using as predictors the variables related to living with one's family and concerning the level of care and attention towards one's own health, we found, in the training set, an accuracy of 0.94.



In the test set, the accuracy index drops to 0.57, but still remains among the best possible.



Conclusions

Before going into the merits of the question, it should be noted that the questions asked at the beginning of the analysis specifically concerned the daily "routine" of the participants in the project, in order to respect the hypotheses established at the start of the research. We decided, therefore, that our goal would be to look for possible causes of insomnia among the events and habits of everyday life, without going to explore the medical conditions of the participants, which are always the main link with problems of insomnia, especially at this age.

Among the subjects of the survey, in fact, there are a multitude of personalities with different habits and practices. Starting from those who live independently or those with their families, to those who prefer to spend a lot of time at the university or those who stay only the bare minimum. Everyone has, therefore, different rhythms.

It is important, however, to remember that no questions were asked about health conditions, which always remain the main link with problems of insomnia, especially at these ages.

The analyses conducted lead us to think that there may be a strong correlation between living with one's family, watching TV series or movies even while eating and insomnia. This may be related primarily to the stress sometimes caused by living with parents and siblings and to the continuous use of screens, which having a negative effect on the eyes, often lead to difficulty in sleeping.

A predictor that we have found to be a good indicator is that which analyzes the level of care and attention to one's own health. It, coupled with both watching TV series or movies even while eating, and with the hours spent on one's hobbies, returns values that allow us to deduce a possible cause-effect relationship.

Moreover, it is reasonable to state that, in light of the model developed, feeling bored and using a cell phone for a long time could cause problems of insomnia in the student age.

Finally, it is more difficult to demonstrate with evidence the link between the variables and the problem of insomnia, but it remains a hypothesis not to be discarded a priori.

Bibliography

Data collection via Google Form