Bullying Risk Predictive Model: Dash Application

Daniel J. Meier

Bellevue University

DSC 410: Predictive Analytics

Frank Neugebauer

June 1, 2024

Bullying Risk Predictive Model: Dash Application

What does the application do?

As prevalent as bullying is in our society, it behooves us to take the necessary steps to work towards reducing the prevalence of bullying behaviors as much as possible. Part of that process is to utilize an application that can provide inferences as to whether someone is at risk for bullying within a school setting based on a series of questions that they answer through this application. These questions pertain to the following topics:

1) Have they experienced cyberbullying?
2) Have they been physically attacked?
3) Number of close friends
4) Frequency of how often their parents demonstrate understanding of their problems
5) Age
6) Gender
7) How often do they feel lonely?
8) How often have they missed school without permission?
9) How often do they consider other students' kind and helpful?

Model Used:

The model that is being utilized in this application is a Random Forest Classifier that was tuned with GridSearchCV and utilized an under sampled target class. With the data that was available, the number of students who said "No" to being bullied at school were significantly higher than those who responded "Yes", which meant that there was an imbalance in the target. This presented a problem in the model learning, as without any adjustment to the target class, the models performed well with the majority class but much poorer on the minority class. I tested out both Logistic Regression and Random Forest Classifier as this was a classification problem (predicting Yes or No answers), so models that work with classification problems needed to be utilized. Random Forest performed better across the board than the best performing iteration of the Logistic Regression models.

Project Methodology:

After performing Data Wrangling by cleaning data and removing some features that were either duplicated by data already present (for example, multiple columns pertaining to missing school were in the original dataset) along with data that had a significant amount of missing values (multiple columns pertaining to weight), I moved on to Feature Engineering to encode all of the features (input and target). Finally, I ran six iterations each of Logistic Regression and Random Forest, concluding that the Random Forest Classifier that was tuned with the target class majority being under sampled performing the best overall. While it did not have the highest overall accuracy, its ability to recall Yes responses was significantly better than any other performing model, which considering the topic, it is best to be able to identify if someone is being bullied or is at risk for being bullied. The following outlines the model iterations that were conducted between each model type, which amounted to 12 iterations total:

1) All features, base parameters
2) All features, tuned model, base parameters
3) Removed two lowest scoring features based on feature importance metrics, base parameters
4) Removed two lowest scoring features based on feature importance metrics, tuned model
5) Under sampled the majority class to bring target classes into balance, base parameters, all features
6) Under sampled the majority class to bring target classes into balance, tuned model, all features

Next Steps:

The next step to make this a fully operational product/application is to acquire more data. My concern is that the best performing model out of the 12 iterations conducted between the two types of models was one where I essentially cut data out. Ideally, I would be able to acquire a significant amount of data that involved students reporting that they have been bullied so the model has more to learn from. Another additional step would be to incorporate additional questions that could help provide additional features for the model to learn from, along with helping to potentially identify other factors that have high correlations/relationships to a student being bullied. I initially dropped questions pertaining to weight because it was missing nearly 2/3 of the data in those respective columns, so it was not conducive to leave in because of how significant it was.

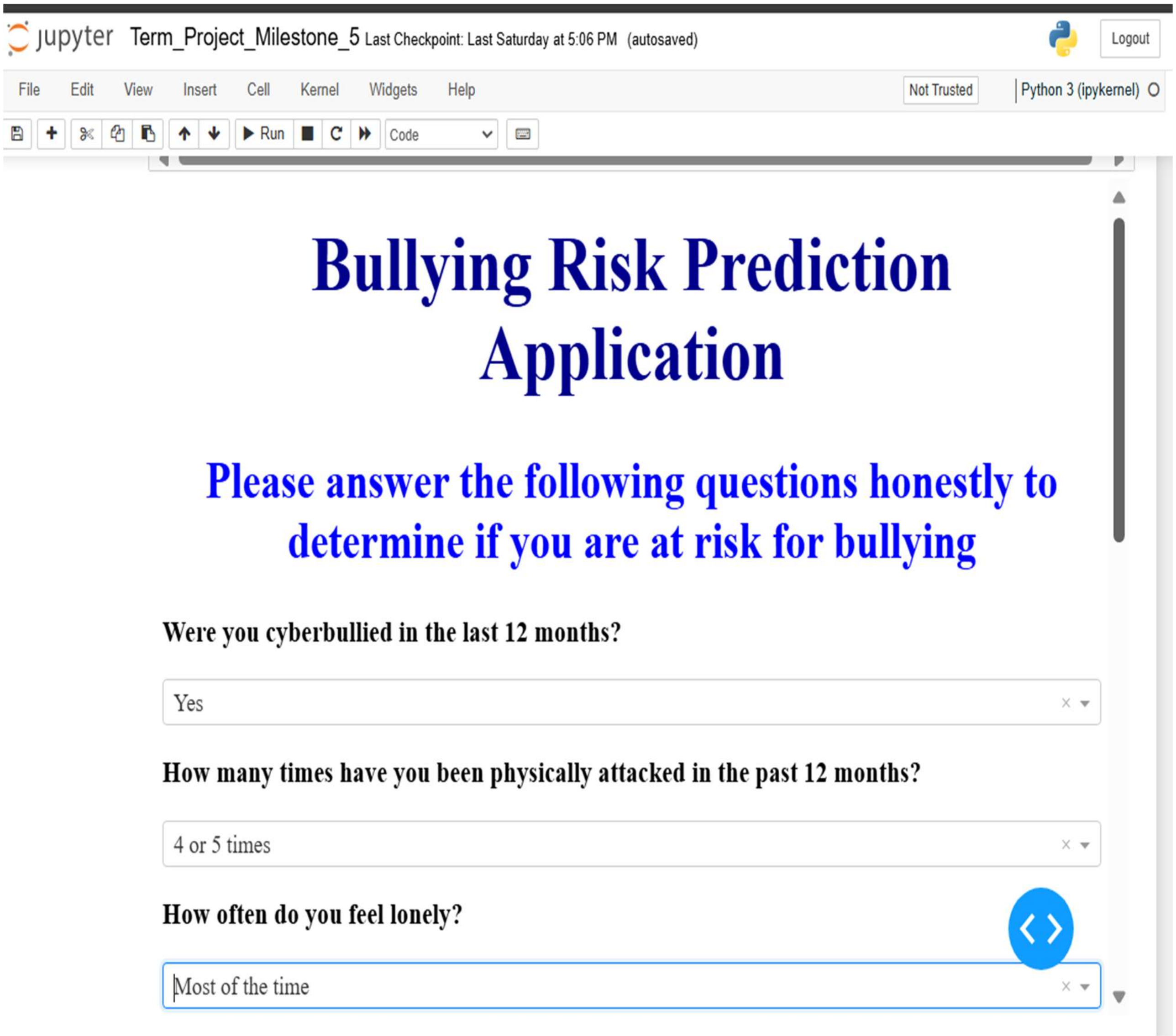
What I learned/still have questions about:

There are a couple things that I feel are imperative to point out as the most significant things I learned. First, the iterative process of trying to churn out the best performing model possible is a very intense process, but very much worth it in the end. It demonstrated the determination and critical thinking skills required to keep doing anything and everything to tweak model parameters, test/train ratio, oversample/under sample, or adjust the model type entirely to get the best performing model you can.

The second thing I learned was about incorporating all of this into a functional application, which I think is neat to see come to life after starting with all the initial thoughts about how to proceed with this topic to begin with. Being able to take an idea and see it play out into a working application really demonstrates the amount of work it takes to get something like this up and running. While there is a lot to be desired at how well the model behind the scenes is performing and making the dash application look more visually appealing, learning about the essence of what makes this run has been a great experience.

One thing I feel I really have questions about is diving deeper into the making of the application itself and being able to make it accessible to other people. Going through this process really has gotten me thinking about all the other applications that I see out there and one day being able to author my own application would be cool, so diving even deeper into more ins and outs of the dash application process would be something I would love to learn more about. It has certainly been a lengthy process to get to where things stand now, but I am very grateful for everything this has taught me, and I plan to take this and continue to grow from it.

Screenshots of the Running Application:



Bullying Risk Prediction Application

Please answer the following questions honestly to determine if you are at risk for bullying

Were you cyberbullied in the last 12 months?

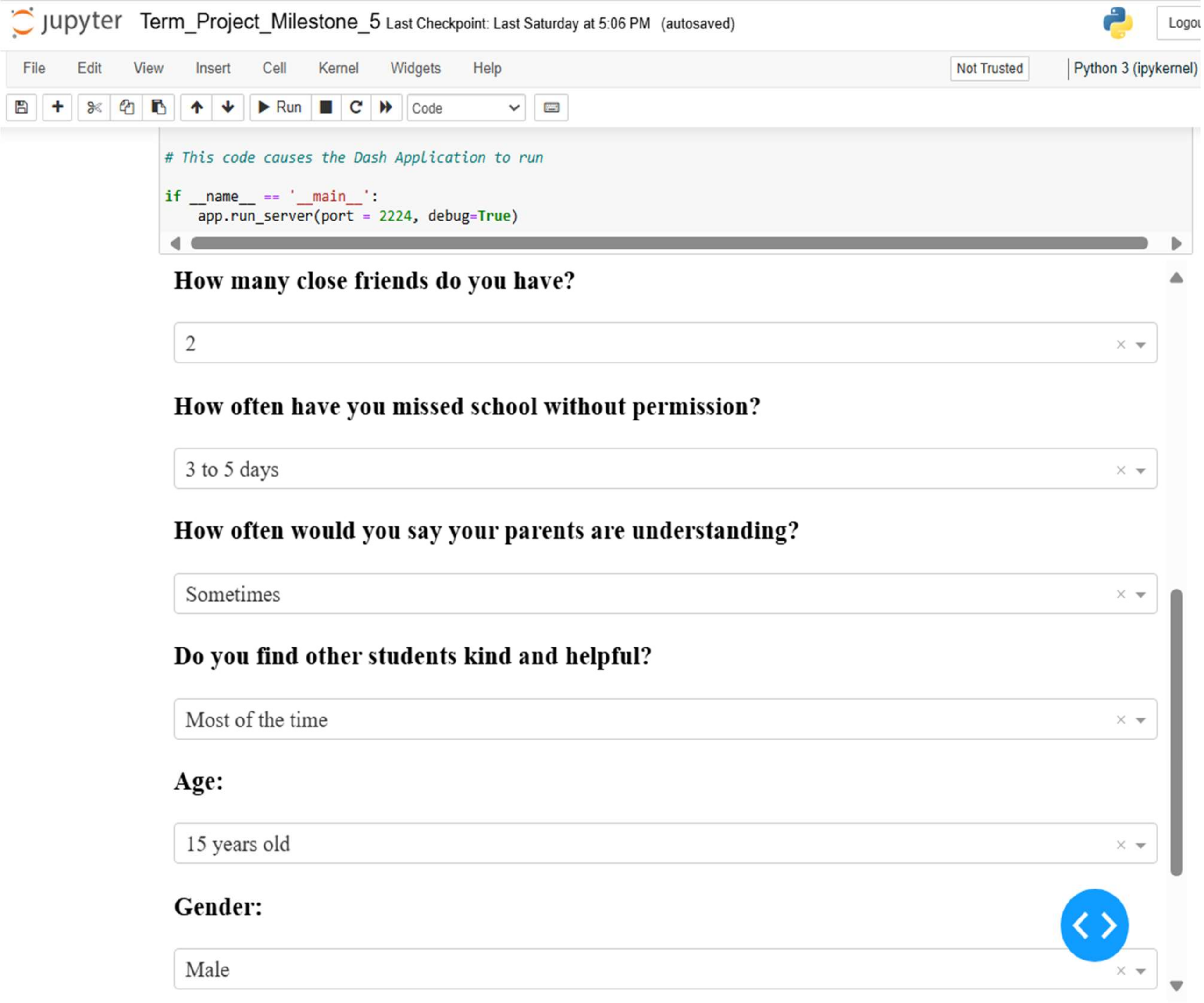
Yes

How many times have you been physically attacked in the past 12 months?

4 or 5 times

How often do you feel lonely?

Most of the time



Term_Project_Milestone_5 Last Checkpoint: Last Saturday at 5:06 PM (autosaved)

File Edit View Insert Cell Kernel Widgets Help

```
# This code causes the Dash Application to run

if __name__ == '__main__':
    app.run_server(port = 2224, debug=True)
```

How many close friends do you have?

2

How often have you missed school without permission?

3 to 5 days

How often would you say your parents are understanding?

Sometimes

Do you find other students kind and helpful?

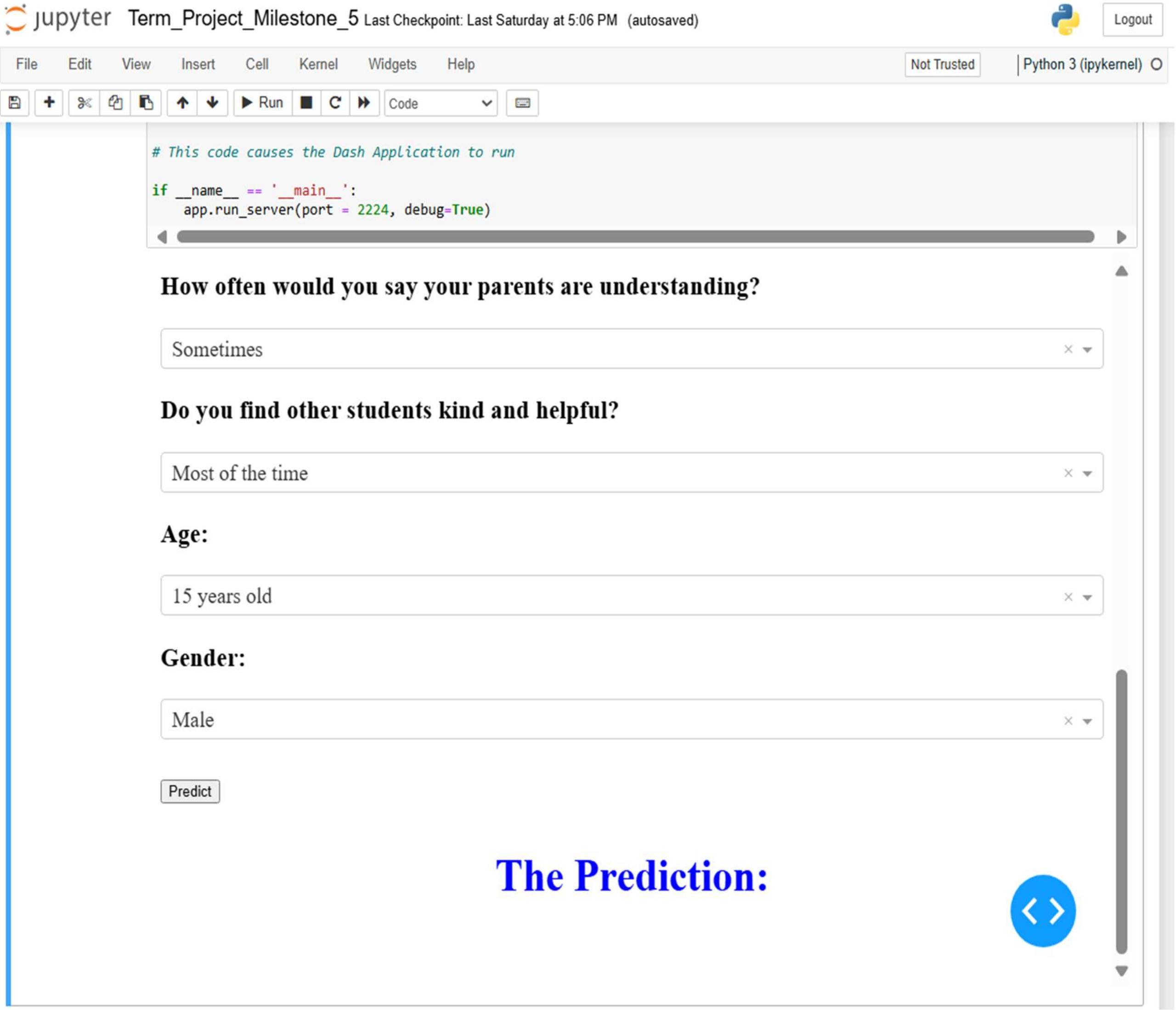
Most of the time

Age:

15 years old

Gender:

Male



jupyter Term_Project_Milestone_5 Last Checkpoint: Last Saturday at 5:06 PM (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)
Run
Code

```
# This code causes the Dash Application to run

if __name__ == '__main__':
    app.run_server(port = 2224, debug=True)
```

How often would you say your parents are understanding?

Sometimes

Do you find other students kind and helpful?

Most of the time

Age:

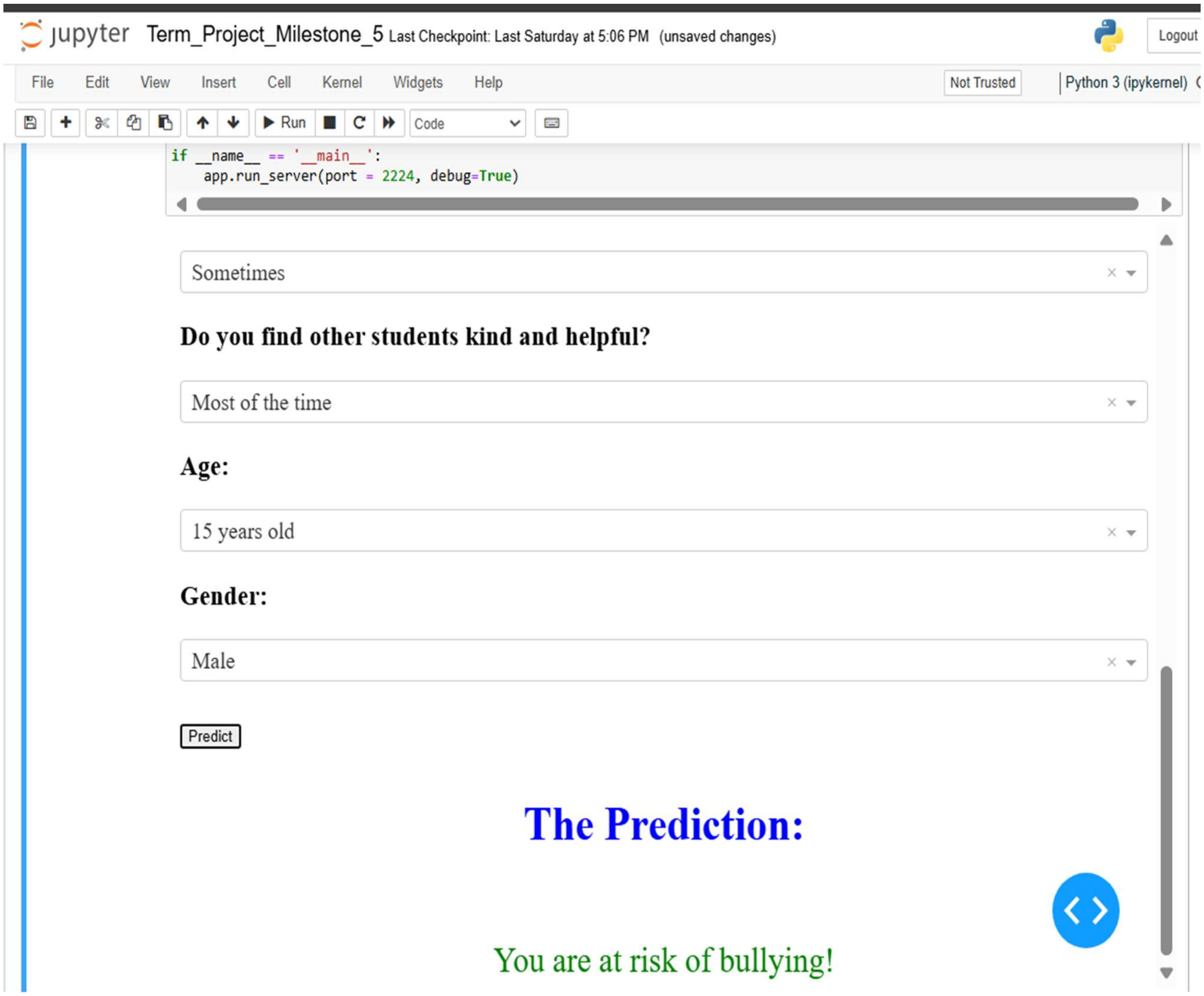
15 years old

Gender:

Male

Predict

The Prediction:



```
if __name__ == '__main__':
    app.run_server(port = 2224, debug=True)
```

Sometimes

Do you find other students kind and helpful?

Most of the time

Age:

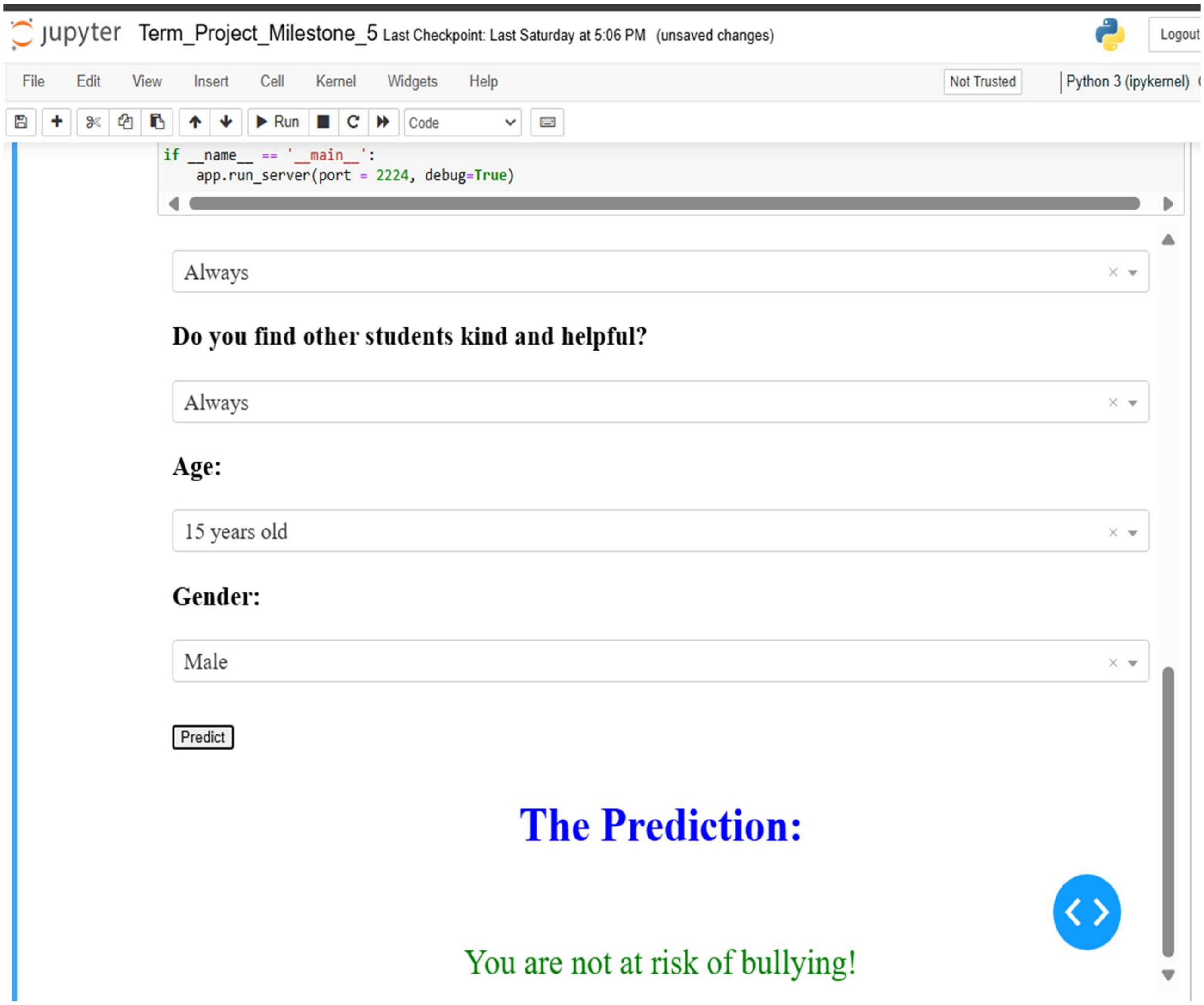
15 years old

Gender:

Male

The Prediction:

You are at risk of bullying!



```
if __name__ == '__main__':
    app.run_server(port = 2224, debug=True)
```

Always

Do you find other students kind and helpful?

Always

Age:

15 years old

Gender:

Male

The Prediction:

You are not at risk of bullying!

References:

Arya, N. (2022, July 25). KDnuggets. *Does the Random Forest Algorithm need Normalization?* https://www.kdnuggets.com/2022/07/random-forest-algorithm-need-normalization.html

Brownlee, J. (2020, August 27). Machine Learning Mastery. *3 Ways to Encode Categorical Variables for Deep Learning.* https://machinelearningmastery.com/how-to-prepare-categorical-data-for-deep-learning-in-python/

Brownlee, J. (2020, August 31). Machine Learning Mastery. *Metrics to Evaluate Machine Learning Algorithms in Python.* https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/

Brownlee, J. (2021, January 27). Machine Learning Mastery. *Undersampling Algorithms for Imbalanced Classification.* https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/

(2024). Plotly. *Dash Layout.* https://dash.plotly.com/layout

Koehrsen, W. (2018, January 9). Towards Data Science. *Hyperparameter Tuning the Random Forest Forest in Python*. https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74