

1. How accurate is the model really, and what does an R^2 of 0.62 mean in real life?

An R^2 of 0.62 means the model explains about 62% of the differences in people's stress scores. For a behavioral dataset—where people are all different and stress has many causes—that is actually pretty strong. It won't predict someone's stress perfectly, but it gives a good estimate within about one point on the stress scale.

2. Does higher screen time actually cause stress, or is it just correlated?

The data shows a strong correlation, which means people with high screen time tend to have higher stress. But correlation doesn't automatically mean one causes the other. Many other factors could play a role, and this project can't prove cause and effect. It simply shows a consistent pattern worth paying attention to.

3. Why were exercise and social media use weak predictors?

People often assume those matter the most, but the data didn't support that. Exercise helps reduce stress, but the dataset only measured how many days per week someone exercises—not the intensity or duration. For social media, the dataset only asked which platform they use, not how much or how it makes them feel. So the weak prediction likely comes from limited detail, not because those behaviors don't matter.

4. How did you make sure the data was reliable if it was self-reported?

Self-reported data is always imperfect, but it's common in wellness research. To handle this, I checked the data for outliers, unrealistic values, and missing entries. The dataset was clean, consistent, and large enough to capture general patterns even if a few people misjudged their numbers.

5. Would the model work differently with older adults or different populations?

Probably, yes. Most people in this dataset were younger adults, so the patterns reflect that group. Older adults, teenagers, or stressed professionals might have different habits. The model is a good starting point, but it would need more diverse data to be used broadly.

6. Why did the Random Forest model perform better than linear regression?

Random Forest captures **nonlinear relationships**, meaning it can detect more complicated patterns between lifestyle habits and stress levels.

Linear regression assumes the relationships are straight-line and simple.

Human behavior is rarely simple, so Random Forest did better.

7. What personal information would someone have to give for this to work in real life?

Very little—just the same information the survey collected: screen time, sleep quality, basic demographics, and exercise frequency. It doesn't need medical data or anything deeply personal. It could work with anonymous inputs, which increases privacy.

8. Could this type of prediction be used by employers in harmful ways?

It could be if used irresponsibly, which is why ethical guidelines are important. This type of prediction should only be used voluntarily to support wellness, not to judge performance or discipline employees. The focus must stay on helping people identify stress, not monitoring them.

9. If you added more data, like income or job type, do you think accuracy would improve?

Most likely, yes. Stress has many causes—finances, workload, living situation—none of which were in the dataset. Adding more meaningful features usually improves accuracy because the model sees more of the full picture.

10. How often would someone need to update their data for the model to stay accurate?

Weekly updates would work well. Stress, sleep, and screen time all change quickly, so frequent updates help the model reflect a person's current lifestyle. If this were an app, updating automatically every day would be ideal.