**About the Data:**

This data comes from the young people survey in Slovakia, sourced from Kaggle. The data was collected in 2013 with 1,010 students at FSEV UK aged 15 to 30. After removing all the rows with null values, we had 674 full responses to use in our models. All students were of Slovakian nationality. The survey asked 150 questions that can be split up into the following groups: ‘Music Preferences’, ‘Movie Preferences’, ‘Hobbies & Interests’, ‘Phobias’, ‘Personality traits, Views on life, & Opinions’, ‘Spending Habits’ and ‘Demographics’. To narrow down our scope, we chose to limit our model features to ‘Personality traits, Views on life, & Opinions’ and chose Gender, Education, Phobia of Heights, Phobia of Spiders and Phobia of Snakes as our targets. The features we used all are on a five-point Likert scale from Strongly disagree (1) to Strongly agree (5), we removed the three variables that did not comply with this same format.

**Our Process:**

We used logistic regression, k-nearest neighbors and neural network models to try and best fit our five targets. These three models were chosen based on their varied approaches to predicting data. We chose to try and predict our targets with the Personality traits, views on life, & opinions variables. Because there are 57 different questions in this group, we ran a logistic regression model in R to inform our models about what variables were statistically significant in predicting each target. For gender, we used a binary logistic regression. For Education, we used a multinomial logistic regression analysis because the dependent variable was categorical. For the Phobias, we used a proportional odds logistic regression because the targets were categorical with more than 2 ordered levels. We simplified the levels from the 5 Likert scale values down to 3: ‘Agree’, ‘Neutral’ & ‘Disagree’. Limiting the number of states of our targets gave us higher number of instances of each value, which helped the accuracy of our model.

For our logistic regression model, we used our informed set of features to predict the target and then used gridsearch to tune the parameters of the model. For our k-nearest neighbors (KNN) model, we used our informed features and then looped through different k values to see which had the highest accuracy. The neural networks also used the informed features, we then split the data into training and testing data and then scaled it. We found that our dataset was too small for deep learning, which yielded significant overfitting. For each of our different models, we used two layers. The first layer varied in the number of units, but all used a ReLU activation function. The second layer was a softmax layer to predict our desired target. We then fit our neural network model with varying number of epochs per model. We evaluated this model with accuracy and loss but used accuracy to compare the results to the other two models.

**Gender:**

Respondents were asked their gender and could either answer Male or Female. We found that the list of features below was statistically significant in predicting the gender of the respondent. Out of the three models, KNN was by far the most accurate in predicting gender, correctly predicting gender 84.6% of the time. Logistic regression was the second-best model with an accuracy of 76.2%, lastly the neural network model predicted gender correctly 74.0% of the time.

**Education:**

Respondents were asked their highest education achieved and could answer one of: ‘Currently a Primary school pupil’, ‘Primary school’, ‘Secondary school’, ‘College/Bachelor degree’, ‘doctorate degree’ or

‘masters degree’. The list of 7 features below were statistically significant in predicting. Our KNN model and neural network models performed the best in predicting educational achievement with both having an accuracy of 63.3%. The Logistic Regression model was a little less accurate, correctly predicting education 62.0% of the time. We decided to use the KNN model over the neural network model due its advantages in simplicity and replicability. The KNN model represents a 46.6% increase in accuracy over a random guess.

**Heights:**

Respondents could either be Afraid, Not Afraid or Neutral about their phobia of Heights. The KNN model was the most accurate in predicting fear of heights, with a score of 55.0%. Logistic regression was the second most accurate at 53.1%, followed by the neural network with an accuracy of 52.1%. The KNN model yields a 21.7% improvement over a random guess in predicting fear of heights. A model that predicts a phobia that is somewhat irrational in nature is going to achieve high levels of accuracy.

**Snakes:**

Respondents could either be Afraid, Not Afraid or Neutral about their phobia of Snakes. The neural network model was the most accurate in predicting phobia of snakes, with an accuracy of 52.7%. The KNN model was the second most accurate at 52.1%, followed by logistic regression with an accuracy of 49.9%. The neural network model yields a 19.4% improvement over a random guess.

**Spiders:**

Respondents could either be Afraid, Not Afraid or Neutral about their phobia of Spiders. Logistic regression was the most accurate model in predicting fear of spiders, with an accuracy of 55.6%. KNN was the second most accurate at 55.0%, followed by the neural network model at 50.9%. The logistic regression model yielded a 22.3% improvement over a random guess.

**Overall Analysis:**

Gender was the most accurate model (84.6%) followed by Education (63.3%), Spiders (55.6%), Heights (55.0%) and Snakes (52.7%). Interestingly, the three types of models for each target yielded accuracies within 5% of each other, outside of Gender. Which suggests that the parameters that we chose for the neural network were very close to perfectly optimal. Relative to random chance, our Education model was the best with an improvement of 46.6% compared to randomly predicting the target state. Gender (34.6%) was the second best followed by Spiders (22.3%), Heights (21.7%) and Snakes (19.4%). Phobias are by nature somewhat irrational and hard to explain why they occur, which makes predicting them inherently challenging. Predicting gender is inherently easier due to there only being two target states in the dataset so it was not surprising that this was our most accurate model. There were 6 different education levels, making this the hardest to accurately predict, but it was our second most accurate model. The k-nearest neighbor model was most accurate for three of the questions and the neural network and logistic regression were each most accurate for one. Logistic regression and KNN models were very easy to work with, whereas the neural network model was volatile, where the accuracy of the model with the same parameters could change by around 4% with the same parameters.