Diabetes Risk Prediction And Understanding the Results

We built a model to predict the likelihood of diabetes using different machine learning techniques. Below is a simplified explanation of what we found.

1. Logistic Regression (Predicting Diabetes Risk)

1. **Accuracy: 100%**
2. The model correctly predicted every case, meaning it made no mistakes. While this sounds great, such perfect accuracy is highly unusual and could indicate an issue with the dataset or model.
3. **Precision, Recall, and F1-Score: 1.00**
4. These are all perfect scores, meaning the model made no false predictions. Again, this is rare and suggests we should check the dataset for errors, data leakage, or an overly simple dataset.
5. **AUC (Area Under the Curve): 1.00**
6. The model perfectly separates those with diabetes from those without, reinforcing its unusual accuracy.
7. **Log Loss: 0.00**
8. This means the models confidence in its predictions was perfect, another sign of a potential issue.
9. **Intercept: 15.98**
10. This number represents the models baseline prediction when all other factors are zero. However, because our model is too perfect, this number doesnt tell us much by itself.

2. Random Forest (Alternative Prediction Model)

1. **Accuracy: 100%**
2. Just like logistic regression, this model also predicted every case correctly. This further raises concerns about possible data leakage or overfitting.

3. K-Means Clustering (Finding Patterns in the Data)

1. The data was grouped into **three clusters**:
2. **Cluster 2:** 109 responses
3. **Cluster 0:** 105 responses
4. **Cluster 1:** 86 responses

This means the algorithm found natural groupings in the data, but we need to analyze how these clusters relate to diabetes risk.

4. ANOVA Test (Relationship Between Weight and Exercise)

1. **F-Value: 7577.23**
2. This is a very high number, showing a strong relationship between weight and exercise.
3. **P-Value: 0.000**
4. A p-value below 0.05 means the relationship is statistically significant. In simple terms, weight and exercise habits are closely related.

5. Weight Confidence Interval (How Reliable Is the Weight Data?)

1. **Mean Weight:** 70.37 kg
2. The average weight of people in the study.
3. **Standard Deviation:** 13.38 kg
4. This tells us how much individual weights vary from the average.
5. **Confidence Interval:** 68.85 kg to 71.89 kg
6. We are **95% confident** that the true average weight in the population falls within this range.
7. **Margin of Error:** 1.52 kg
8. This is the potential difference between the sample mean and the true population mean.
9. **Sample Size:** 300
10. The number of people used in this analysis.
11. **Population Size:** 1,000,000
12. The total number of people we are trying to make predictions for.

Since the sample size is relatively small compared to the total population, we should be cautious when generalizing the results.

6. Regression Analysis (Relationship Between Height and Weight)

1. **RÂ² (Coefficient of Determination):** 0.516
2. This means height explains **about 51.6%** of the variation in weight. There is a **moderate** relationship between the two.
3. **Slope (Î²â,):** 65.90
4. For every **1-unit** increase in height, weight increases by **65.90 units** on average.
5. **Intercept:** -45.74
6. This is the point where the line crosses the y-axis.
7. **Regression Equation:**

8. **y = 65.90 * x - 45.74**
9. This is the formula we can use to predict weight based on height.
10. **Sum of Squared Errors (SSE):** 25,912.47
11. This measures the total error in the model.
12. **Mean Squared Error (MSE):** 86.37
13. This is the average squared error in predictions.

The model shows a moderate correlation between height and weight, meaning taller people tend to weigh more, but height alone does not fully predict weight.

Key Takeaways & Concerns

Perfect Accuracy in Diabetes Prediction Models (100%)

1. This is highly suspicious and requires investigation.
2. Possible reasons:
3. **Data Leakage** The model may have accidentally "seen" the correct answers during training.
4. **Overfitting** The model may have memorized the data instead of learning useful patterns.
5. **Too Simple Data** If theres an obvious pattern in the dataset, the model might easily separate cases of diabetes and non-diabetes.

Height and Weight Have a Moderate Relationship

1. Height explains about **51.6%** of the variation in weight.
2. Other factors like diet, lifestyle, and genetics likely play a role.

Weight and Exercise Are Strongly Related

1. The ANOVA test confirms that weight and exercise habits are statistically linked.

The Study Sample is Small

1. The dataset has **300** responses, but we are trying to predict for **1,000,000** people.
2. A larger sample size would give more reliable predictions.

Final Thoughts

The results suggest that while our model is technically "perfect," something may be wrong with the data or methodology. Before using this model in real-world applications, we must:

1. **Double-check the dataset** for issues like duplicate entries or leaked information.
2. **Re-evaluate the training process** to ensure proper data splitting.
3. **Test the model on completely unseen data** to confirm its real-world accuracy.

While the height-weight relationship and weight-exercise findings seem reasonable, the diabetes prediction models require further validation before they can be trusted.