**Assignment 1**

INF2178 Experimental Design for Data Science

Mar/28/2022

Instructor: Shion Guha

Benjamin Kelly 1003192662

Faculty of Information

Information Systems & Design

**Introduction**

Artificial intelligence is increasingly prevalent as researchers and corporations aim to increase the efficacy of certain activities using machines and algorithms. Healthcare is one sector in which AI use has exploded to reduce the need for highly specialized personnel and devices and even improve the accuracy of health care like AI-enabled MRI scan readings (Mehralivand et al., 2020; Wahl et al., 2018). Many groups have focussed these efforts on the prediction of diabetes diagnoses through various techniques like speech recognition and pre-existing health data (Faruque et al., 2019; Kandhasamy & Balamurali, 2015; Spänig et al., 2019; Zou et al., 2018). However, these diabetes prediction models are not yet applied in practice.

Further, many diabetes prediction models use the same relatively small database, the “Pima Indians Diabetes Database” (Smith et al., 1988). This data stems from the willing scientific participation of the Pima Indians of Arizona and Mexico. Unfortunately, due to colonialism and changes to their livelihoods, they have developed into a high-risk community for Type 2 diabetes (Schulz & Chaudhari, 2015). Since the dataset is fundamental for diabetes prediction and based on one specific population, I thought it would be essential to check if generally agreed upon biological relationships hold in this smaller population (Kaufman et al., 1997; Mungreiphy et al., 2011). Therefore, my hypothesis is that age and BMI predict blood pressure.

**Method**

**Data**

The Pima Indians Diabetes Database is pulled from the National Institutes of Diabetes and Digestive and Kidney Disease. Yet, the study of the Pima Indians began in 1965, when they were discovered to have the highest prevalence of diabetes ever recorded (Schulz & Chaudhari, 2015). The condensed for of the dataset used for this paper has been reduced to 9 columns: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome (diabetes or not). The dataset consists of 768 females, all at least 21 years of age. The data was also confirmed to be tidy as the column headers only include vital information and all data descriptors are clear. Each column represents a variable, each row is an observation (an individual), and each cell represents a value (Wickham, 2014).

The pregnancy column contains the number of times each individual has been pregnant. Glucose represents the plasma glucose concentration 2 hours after an oral glucose tolerance test. The measure for Blood Pressure (BP) is the Diastolic blood pressure in mm Hg. Skin Thickness is the thickness of the skin fold created by squishing the back of a relaxed tricep in millimetres. The Insulin column is the insulin level (mu U/ml) taken from a serum sample at the same 2-hour interval as the glucose tolerance test. Body Mass Index (BMI) is a general measure of body mass comparing weight and height, calculated as weight in kg/(height in m)^2. Diabetes Pedigree Function (DPF) is a unitless measure that calculates the likelihood of getting Type 2 diabetes based on family history (Smith et al., 1988). Finally, age and Outcome are the individual's age and if the individual has or does not have diabetes.

**Cleaning and EDA**

After first finding that all values exist for each column and individual and none of them were NULL, NA, or NaN, describe was used to view summary statistics for each column. A minimum value of 0 was found for all columns except Age and DBF. Glucose, BP, Skin Thickness, Insulin and BMI were all suspicious as 0’s based on the variable, histogram distributions, and lying outside the boxplot interquartile ranges. Based on their measurements, glucose, BP, and BMI could not be 0 as the individuals would be dead, meaning the 0’s represented errors, each accounting for less than 10% of the total. They were corrected by imputing the median of each column grouped by Outcome to account for any group differences. Following some research, insulin levels can’t be 0, but they can be lower than the tools they would have been using at data collection could detect, so the 0’s could represent very low levels. Along with accounting for almost half of all insulin values, the 0’s in insulin were left untouched. Finally, though 0 Skinfold thickness seems odd, I could not find a reputable source saying the value could not be 0 and based on how it is collected, very tight skin for whatever reason could not be measured by the tool. Just under 30% of the Skinfold Thickness values were 0, and according to some, imputing more than 10% a variable may be more detrimental than leaving them (Bennett, 2001). Overall, Skinfold Thickness values of 0 were left, with the confusion confirming the decision to leave it out of further evaluation in my primary statistical models (ANOVA’s and Linear Regression).

A correlation matrix (Figure 1) informed me of which variables had the most substantial relationships with others with the imputed values. Some variables had easily explainable correlations like Pregnancies and Age (moderate), and of the variables to be analyzed, Age and BP (0.33) along with BMI and BP (0.29) had weak correlations but larger than many of the other variable combinations. Interestingly, Age and BMI (0.028) did not have a correlation, which will be of interest in the Two-Way ANOVA.

A picture containing application

Description automatically generated

**Figure 1:** Correlation matrix with imputed 0 values.

The continuous variables of interest need to be binned into categorical variables in preparation for the ANOVA’s. Age was binned into ‘Young Adult,’ ‘Middle Age,’ and ‘Older Adult’ while BMI was binned into ‘Underweight,’ ‘Normal,’ ‘Overweight,’ and ‘Obese’ following established binning principles for each (Health Canada, 2010; Petry, 2002). These newly binned categorical variables were also turned into dummy variables for the Linear Regression model. Boxplots of the binned data revealed many outliers outside the confidence intervals, let alone the IQRs for, but outlier analysis would be outside the scope of the assignment. Value counts were analyzed within each bin’s groupings, and Older Adults and Underweight individuals were relatively few, potentially influencing further analysis.

           Assumptions of the data were then tested before moving further. Using Shapiro-Wilks tests, no variable achieved normality, breaking that assumption for ANOVA’s and the linear model. The binned data was also non-normal. However, the normality measures can be ignored with our large sample size as the Central Limit Theorem ensures residuals will be distributed normally (Ghasemi & Zahediasl, 2012). Levene’s tests were run between the three variables of interest, all with non-significant p-values, meaning Age, BMI, and BP all have similar variance, even when comparing BP to the binned data. The observations between individuals in the community are also assumed to be independent to satisfy our analysis.

           Finally, t-tests were performed before ANOVAs to explore further the proposed research question of Age and BMI predicting BP. T-tests on the raw values between Age and BP and BMI and BP were both statistically significant at a *p<.05* level with t-statistic values of ~164 and ~159, respectively. Consistent with the correlation matrix, BMI and Age did not have significantly different means, another interesting layer of their interaction in predicting blood pressure (BP). T-tests involving dummy variables were statistically significant but potentially less meaningful as data is effectively lost when comparing the means.

**Analysis Methods**

One-Way ANOVA’s were performed on the difference of BP means between the individual binned groups of Age and BMI. Then, Tukey-HSD tests with Bonferroni corrections followed to identify between which groups the differences were significant. This was followed by a Two-Way ANOVA with all three variables involved, again with Tukey-HSD with Bonferroni correction *post hoc* tests. The Two-Way ANOVA is also accompanied by an interaction plot to visualize how Age and BMI influence BP. Finally, the ANOVA’s were compared to the predictions of a multiple linear regression using Age and BMI to predict BP, evaluated with various error measures and r-squared values.

**Results**

Within the One-Way ANOVA’s, BP was significantly different between Age groups, with the Tukey-HSD *post hoc* test showing these differences are really between Young and Middle-Aged adults and Young and Older Adults with both having *p<.05*significant values. There was no significant difference between Middle-Aged and Older adults. BMI groups were also significantly different for the BP and BMI One-Way. Interestingly, this difference only occurred between Obese and Normal and Obese and Overweight groups with *p<.05*. All relationships with the Underweight group were non-significant, perhaps related to their minimal representation in the dataset.

           The Two-Way ANOVA showed that BMI and the interaction between BMI and Age were significant, but Age in the model was not, with a *p>.05* from two separate packages. It might have been an output error related to significant digits, and the *post hoc­*tests on Age showed the same significant differences between Young and Middle-Aged adults and Young and Older Adults. The Tukey-HSD with Bonferroni correction *post hoc ­*tests also showed the same significant group differences only between Obese and Normal and Obese and Overweight groups, *p<.05*. Finally, the interaction *post hoc*tests were slightly more challenging to understand, tracking every relation between all possible group combinations. Generally, the significant differences occurred between Age and BMI combinations that were previously significant; age differences that were not between Middle Age and Older, and BMIs differing between Obese and Normal/Overweight. NA values were removed from the overview for easier viewing since they represented groupings that did not exist, like the Underweight - Older Adult group. These interactions can be best understood through the interaction plot in Figure 2.

Chart, line chart

Description automatically generated

**Figure 2:** Interaction plot for Age and BMI predicting blood pressure.

The linear regression was first applied to the Age and BMI dummy variables. The intercept occurs at a blood pressure value of around 53, with an Age coefficient of almost 6 and BMI coefficient of almost 3. The model is a poor blood pressure predictor with a mean absolute error of 8.8, root mean squared error (RMSE) of 11.5 and r-squared of 0.12. The coefficients are strange as for each increase in the dummy value, blood pressure increases by those values, but there are only 3-4 levels of dummy values. This assumes that the differences between each bin are considered equal, which is false when binning any continuous data. The difference between individuals aged 29 and 30 is relatively small on a continuous scale, but much more prominent when the 29 is now Young Adult and 30 is Middle Age, especially when there are only three groups. Since the dummy variables lose some data through binning and linear regressions work best with continuous data, I ran the same linear regression on the raw values before binning. Its intercept occurs at a diastolic blood pressure of almost 46, and Age coefficient of around 0.3 and BMI coefficient of almost 0.5. That is a much more understandable 0.3 increase in blood pressure for each age increase of 1 year. This model has a mean absolute error of 8.56, root mean squared error (RMSE) of 11.3, and r-squared of 0.18, making it a slightly more accurate blood pressure predictor than with dummy variables but still an inaccurate model. The differences in prediction for a small sample between the models can be found in Figure 3.

1. **b)**

**Table

Description automatically generated** **Table

Description automatically generated**

**Figure 3:** Panel a) Represents the predictions for the Age and BMI binned into dummy variables, b) shows the raw value predictions on the test data.

**Conclusion**

The suite of analyses highlighted in this paper highlight some interesting findings about the relationship between Age, BMI, and Blood Pressure (BP) within the female Pima Indian population. Correlations reveal that Age and BMI both correlate with Blood Pressure and t-tests found significant differences between their means. The methods using ANOVA’s revealed group differences between binned Age, BMI, and their interaction, indicating some predictive value of BP from both. However, not all group differences were significant, mainly related to the Underweight and Older Adult groups. It would be interesting to perform the same analysis on data with a greater representation of those individuals. Finally, the linear regression does not provide an accurate enough method to predict blood pressure for any practical purpose. It was also interesting that the linear regression increased its predictive value when run using continuous values compared to dummy values. Overall, the analyses highlight that the Pima Indian population generally follows accepted relationships between Age, BMI, and Blood Pressure, adding a layer of confidence that models trained on this data could represent broader populations (Kaufman et al., 1997; Mungreiphy et al., 2011). In the future, analysis could treat diabetes outcomes as a covariate to account for group differences, and including more variables with more sophisticated techniques will lead to more accurate models.

**Bibliography**

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, *25*(5), 464–469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

Faruque, Md. F., Asaduzzaman, & Sarker, I. H. (2019). Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus. *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1–4. https://doi.org/10.1109/ECACE.2019.8679365

Ghasemi, A., & Zahediasl, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Health Canada. (2010). Prenatal Nutrition Guidelines for Health Professionals. *Minister of Health Canada*. https://doi.org/: H164-126/2010E-PDF

Kandhasamy, J. P., & Balamurali, S. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia Computer Science*, *47*, 45–51. https://doi.org/10.1016/j.procs.2015.03.182

Kaufman, J. S., Asuzu, M. C., Mufunda, J., Forrester, T., Wilks, R., Luke, A., Long, A. E., & Cooper, R. S. (1997). Relationship Between Blood Pressure and Body Mass Index in Lean Populations. *Hypertension*, *30*(6), 1511–1516. https://doi.org/10.1161/01.HYP.30.6.1511

Mehralivand, S., Harmon, S. A., Shih, J. H., Smith, C. P., Lay, N., Argun, B., Bednarova, S., Baroni, R. H., Canda, A. E., Ercan, K., Girometti, R., Karaarslan, E., Kural, A. R., Pursyko, A. S., Rais-Bahrami, S., Tonso, V. M., Magi-Galluzzi, C., Gordetsky, J. B., Macarenco, R. S. E. S., … Turkbey, B. (2020). Multicenter Multireader Evaluation of an Artificial Intelligence-Based Attention Mapping System for the Detection of Prostate Cancer With Multiparametric MRI. *American Journal of Roentgenology*, *215*(4), 903–912. https://doi.org/10.2214/AJR.19.22573

Mungreiphy, N. K., Kapoor, S., & Sinha, R. (2011). Association between BMI, Blood Pressure, and Age: Study among Tangkhul Naga Tribal Males of Northeast India. *Journal of Anthropology*, *2011*, e748147. https://doi.org/10.1155/2011/748147

Petry, N. M. (2002). A Comparison of Young, Middle-Aged, and Older Adult Treatment-Seeking Pathological Gamblers. *The Gerontologist*, *42*(1), 92–99. https://doi.org/10.1093/geront/42.1.92

Schulz, L. O., & Chaudhari, L. S. (2015). High-Risk Populations: The Pimas of Arizona and Mexico. *Current Obesity Reports*, *4*(1), 92–98. https://doi.org/10.1007/s13679-014-0132-9

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261–265.

Spänig, S., Emberger-Klein, A., Sowa, J.-P., Canbay, A., Menrad, K., & Heider, D. (2019). The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes. *Artificial Intelligence in Medicine*, *100*, 101706. https://doi.org/10.1016/j.artmed.2019.101706

Wahl, B., Cossy-Gantner, A., Germann, S., & Schwalbe, N. R. (2018). Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings? *BMJ Global Health*, *3*(4), e000798. https://doi.org/10.1136/bmjgh-2018-000798

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00515