Final Project 1 - Lightning Talk

Carolyn Nohejl

Criteria for project selection

1. Healthcare focus - relevant to my domain expertise and what we do at Verily
2. Leverage supervised learning - the focus of this course
3. If a classification problem, confirm a decent class distribution
4. Data must provide an opportunity to practice the following techniques:
    a. Data clean up/wrangling
    b. EDA, opportunity to leverage plots to gather additional insights from the data
    c. K-nearest neighbors algorithm for classification
    d. Linear regression of attributes
5. From Kaggle - leverage inputs to enhance learning experience

#1 - Predicting type 2 diabetes leveraging diagnostic tests

Background: Type 2 diabetes (T2D) is a preventable condition that is highly prevalent, associated with severe comorbidities, and is very costly to the healthcare system. It would be very valuable to predict whether a patient is at risk leveraging data from simple diagnostic tests. This information would be valuable to the patient, provider, and payer, as it would help identify who could benefit from early intervention and result in better disease prevention.

Problem statement: Predict the likelihood of onset of type 2 diabetes within a 5 year period based on diagnostic measures, leveraging a subset of the Pima Indians Diabetes Database.*

Data: Data subset includes 768 cases, of which all are women at least 21 years old with Pima Indian heritage.

- **Attributes:** 1 classifier (whether the woman has diabetes) and 8 features: pregnancies, glucose, blood pressure, skin thickness, triceps skinfold thickness, insulin, BMI, diabetes pedigree function, age
- **Class distribution:** ~⅓ of subjects have diabetes, ~⅔ do not
- **Data quality:** Pretty clean, with missing values for two features (skin thickness, insulin)

Hypothesis: Expect that glucose, insulin test, and BMI will be the greatest predictors of T2D.

Success looks like: identifying which of these variables/inputs have the greatest correlation of diabetes so they can be leveraged as an early test to identify high risk patients for early intervention/behavior changes.

**data is originally from National Institute of Diabetes and Digestive and Kidney Diseases*

#2 - Predicting mental illness & attitudes toward mental health in tech

Background: Nearly 50% of people experience mental illness, only around 20% of people with mental illness receive professional help, and stigma around mental health persists.[1] Better understanding attitudes towards and prevalence of mental illness in the workplace could be leveraged to drive programs to provide employees the support they need to lead happy and productive lives and drive cultural change. Valuable to employees, employers, psychologists, and payers.

Problem statement: Predict the likelihood of having a mental illness and attitudes toward mental health based on employee info, geography, and workplace information, leveraging the Mental Health Tech survey data provided by Open Sourcing Mental Illness in 2014.

Data: Data includes 1260 responses from both genders, all over the world.

- **Attributes:** 25 attributes covering employee information including gender and geography, workplace info, and openness about mental health, and one comments section.

- **Class distribution:** Would leverage the "work interfere[2]" feature as a class indicator telling us if the employee struggles with mental health (960/1260 responses).

- **Data quality:** seems pretty clean, with exception of gender and number of employees

Hypothesis: Expect attitudes about mental health to vary between tech/non, geographic location, and company size. **Success looks like:** Understand predictors for employment environments which lack support for employees with mental illness such that these situations can be improved. Also, determine predictors for mental health to better understand which employees may be at risk.

1. From Mental illness in society
2. work_interfere: If you have a mental health condition, do you feel that it interferes with your w

#3 - Predict malignancy of a breast tumor

Background: Breast cancer is a threatening illness, and early diagnosis leads to greater chances of survival.

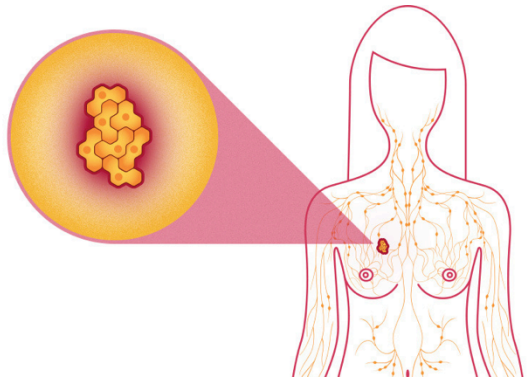
The Problem: Predict the malignancy of a tumor based on cell nucleus attributes identified via digitized images breast tissue extracted via fine needle aspiration via the Diagnostic Wisconsin Breast Cancer Database.

Data: Features of the cell nucleus are computed from a digitized image of a fine needle aspirate of a breast mass. There are 569 samples.

- **Attributes:** 10 features for each cell nucleus: radius, texture, perimeter, areas, smoothness, compactness, concavity, concave points (# of), symmetry, fractal dimension.
- **Class distribution:** Classified as malignant (212) or benign (357).
- **Quality:** very clean - no missing attribute values

Hypothesis: More than one (or a combination of) attributes will predict malignancy.

Success looks like: Ability to identify with confidence if a tumor is malignant or benign.



Recommendation

Pursue project #1 - predicting type 2 diabetes leveraging diagnostic tests

- Highly relevant to my work and interests
- Simple dataset → simple strategy
- Acceptable class distribution (⅓ / ⅔)
- Data to clean

Project 2: predicting mental illness and attitudes toward mental health

- Very interesting, but many variables increases the complexity. Will be more challenging to develop a strategy. Would like to tackle as a future project.

Project 3: predicting breast tumor malignancy

- Less relevant to my work. Attributes are all dimensions (attributes of other datasets are more diverse, which I find more interesting).