Challenge Project #1:

Detecting cases of hypothyroidism

Dhrithi Guntaka, Nithika Vivek, Sofie Budman

What problem or goal did your project address, and why did you choose this topic?

- Hypothyroidism is characterized by non-specific symptoms, making it difficult for medical professionals to diagnose.
- While it is an endocrine disorder, thyroid dysfunction affects multiple body systems. In fact, 23.3% of patients with coronary artery disease suffer from some form of thyroid dysfunction.
- Our project detects early thyroid dysfunction using lab results to achieve high sensitivity and specificity.

Image from biorender

What steps did you take during the development of your project?

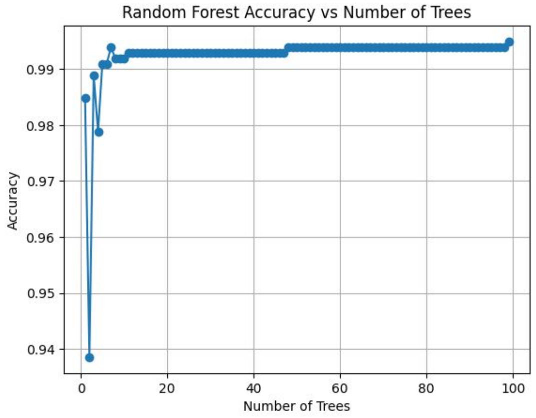
- Data Cleaning
- Chose the 4 features with best correlation (FTI, TT4, T3, TSH)
- Normalized data
- First trained logistic model and zero rule model
- Trained Random Forest Model
- Feature engineering (number of trees, max tree depth, min leaves)
    - 99 trees, max depth = 5, min leaves = 15

	FTI	TT4	T3	TSH	Class
0	109.0	125.0	2.5	1.30	0
1	107.0	102.0	2.0	4.10	0
2	120.0	109.0	2.0	0.98	0
3	107.0	175.0	1.9	0.16	0
4	70.0	61.0	1.2	0.72	0

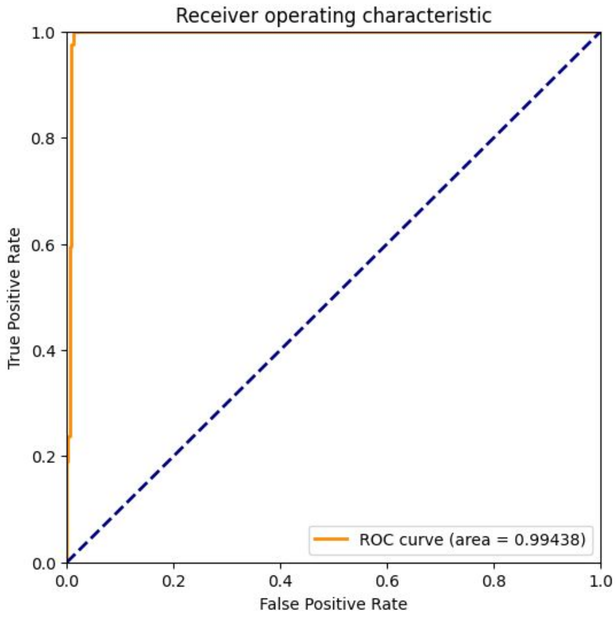
→

	FTI	TT4	T3	TSH
0	-0.042950	0.474170	0.659712	-0.155486
1	-0.107833	-0.196624	-0.022889	-0.005409
2	0.313910	0.007531	-0.022889	-0.172638
3	-0.107833	1.932416	-0.159409	-0.216589
4	-1.308179	-1.392385	-1.115051	-0.186574

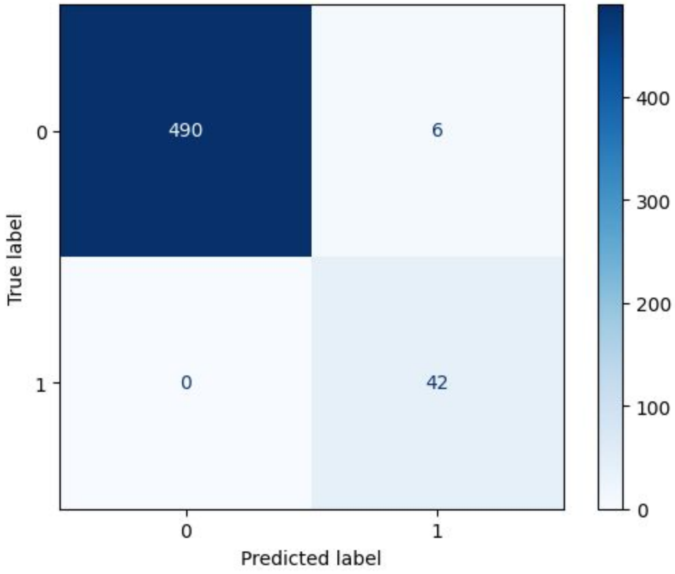
normalization



ROC Curve

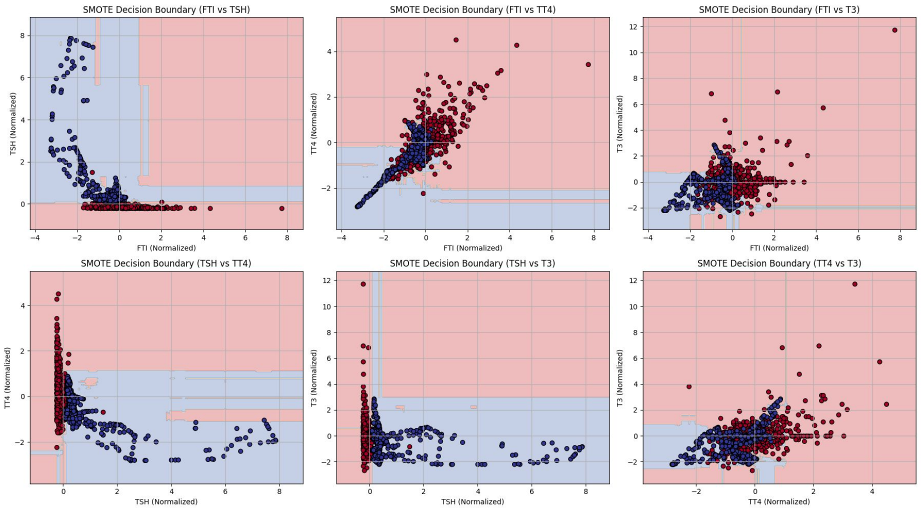


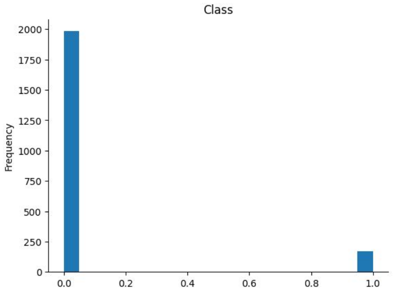
Unbalanced Data Confusion Matrix



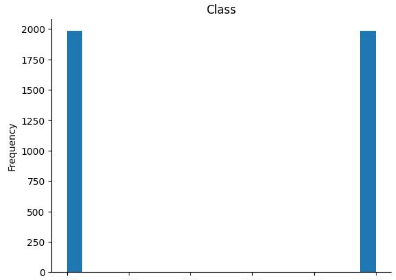
What challenges did you face, and how did you overcome them?

1. Data cleaning and correlation matrices
    a. “TBG Measured” column in the correlation matrix kept giving NaNs when correlated with any other feature
    b. TBG measured had only the “f” variable and no variance, so it could not be correlated
2. Class distributions
    a. We visualized our model predictions, and they were all 0
    b. We visualized the class distributions in the dataset and found that 92% of the data belongs to class 0 (very biased)
        i. That corresponds with the 92% accuracy of our model – the 8% misclassified were all the 1s.
    c. We ran it with parameter class_weight = balanced, and our accuracy was 32%
    d. We tried to solve this in the future model using SMOTE





Before SMOTE



After SMOTE

If you had more time or resources, how would you improve or expand your project?

1. SMOTE
    a. SMOTEENN for noisy barriers
    b. SMOTE variants like Borderline SMOTE, SVM SMOTE
    c. Have SMOTE take into account context when making augmented data (to avoid impossible combinations of data like a patient on thyroxine and a high TSH) through Casual-AWARE
    d. Run SMOTE for each subgroup (male and female to avoid undersampling a certain feature with isn't as prominent)
    e. SMOTE also has limitations so optimizing hyperparameters without SMOTE could be a good idea
2. Automatedly test other hyperparameters other than # of trees as well as other model types like deep learning
3. Try a Random Forest that adjusts it's hyperparameters as it trains to identify ideal hyperparameters rather than iterative training for each combination (saves time, more accurate) - self mutating?
4. Group the features into medical data (TSH, T3, T4U, etc) vs patient profile (age, sex, on_thyroxine) and find feature importance per group to avoid ruling out groups of features that are actually correlated
5. Test different thresholds as well