READ ME

Project overview and goals:

The goal of this project is to identify and evaluate the best classification model for detecting individuals at risk for developing pancreatic cancer, as well as what features, such as biomarkers or patient information have strong importance in driving the models. I will be training and tuning binary classification models to classify a patient as susceptible or not susceptible to developing pancreatic cancer. The data references urinary biomarker levels such as LYVE1, REG1A, TFF1, CA19-9 which will help to define feature importance as well as show trends.

Expected Results

Once preprocessing and EDA are complete, the key features (biomarkers) most strongly associated with pancreatic cancer will be identified through LASSO regression. Following feature selection, the machine learning model (such as Logistic Regression or Random Forest) will predict whether a patient is likely to develop pancreatic cancer. Cross-validation will ensure the model's stability, and model evaluation metrics (precision, accuracy, F1-score) will give insights into its predictive power. The best-case scenario is a highly accurate model that can reliably predict a patient's likelihood of developing pancreatic cancer, which could be used for early intervention.

Rationale, why is this question important?

Pancreatic cancer is one of the deadliest cancers, with a 5-year survival rate of less than 10% due to late diagnosis. Each year, about 60,000 people are diagnosed in the U.S., and over 90% of patients die within a year of diagnosis. Early detection is critical to improving survival rates, but current methods are often invasive or unreliable. This model aims to provide a non-invasive, cost-effective solution by predicting cancer risk based on urinary biomarkers, which could enable earlier diagnosis and treatment.

From a business perspective, the healthcare industry could use such a model to develop new diagnostic tools or improve existing ones, potentially reducing healthcare costs associated with late-stage cancer treatments. Pharmaceutical companies could also benefit by identifying at-risk individuals for clinical trials. Additionally, insurers could use such models to better assess patient risk and improve preventive care strategies, ultimately leading to better health outcomes and reduced costs.

Data Sources and Information

I used the "Urinary biomarkers for pancreatic cancer" dataset from Kaggle. This dataset is well-suited because it includes relevant biomarkers (LYVE1, REG1A, TFF1, CA19-9), patient demographics, and diagnostic data, allowing for both feature importance analysis and disease classification. The dataset has both control and diagnosed patients, which is ideal for training a classification model to distinguish between those at risk of developing pancreatic cancer and those who are not.

- Sample_id
    - Unique identifier for each subject
- Patient_cohort:
    - cohort1: previously used samples
    - cohort2: newly added sample
- Sample_origin:
    - BPTB: Barts Pancreas Tissue Bank, London, UK
    - ESP: Spanish National Cancer research Centre, Madrid, Spain
    - LIV: Liverpool University, UK
    - UCL: University Collee London, UK
- Age:
    - patient age
- Sex:
    - Male / Female
- Diagnosis:
    - 1 = control (no pancreatic disease)
    - 2 = benign hepatobiliary disease (119 of which are chronic pancreatitis)
    - 3 = Pancreatic ductal adenocarcinoma, i.e. pancreatic cancer
- Stage:
    - For those with pancreatic cancer, what stage it was at
    - IA, IB, IIA, IIIB, III, IV
- Benign Sample:
    - For those with a benign, non-cancerous diagnosis, what was the diagnosis?
- Plasma_CA19_9:
    - Blood plasma levels of CA 19–9 monoclonal antibody that is often elevated in patients with pancreatic cancer.
- Creatine:
    - Urinary biomarker of kidney function
- LYVE1:
    - Urinary levels of Lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis
- REG1B:
    - Urinary levels of a protein that may be associated with pancreas regeneration.
- TFF1:
    - Urinary levels of Trefoil Factor 1, which may be related to regeneration and repair of the urinary tract
- REG1A:
    - Urinary levels of a protein that may be associated with pancreas regeneration. Only assessed in 306 patients (one goal of the study was to assess REG1B vs REG1A)

- 0 sample_id 590 non-null object
- 1 patient_cohort 590 non-null object
- 2 sample_origin 590 non-null object
- 3 age 590 non-null int64
- 4 sex 590 non-null object
- 5 diagnosis 590 non-null int64
- 6 stage 199 non-null object
- 7 benign_sample_diagnosis 208 non-null object
- 8 plasma_CA19_9 350 non-null float64
- 9 creatinine 590 non-null float64
- 10 LYVE1 590 non-null float64
- 11 REG1B 590 non-null float64
- 12 TFF1 590 non-null float64
- 13 REG1A 306 non-null float64

Cleaning and Preparation

From the data

- The columns 'stage' and the 'benign_sample_diagnosis' have empty values for their first 200 rows. This is because half the dataset contains patients who have 'stage 0' or 'no diagnosis', so these empty cells should be filled in with '0' and 'No Diagnosis'.

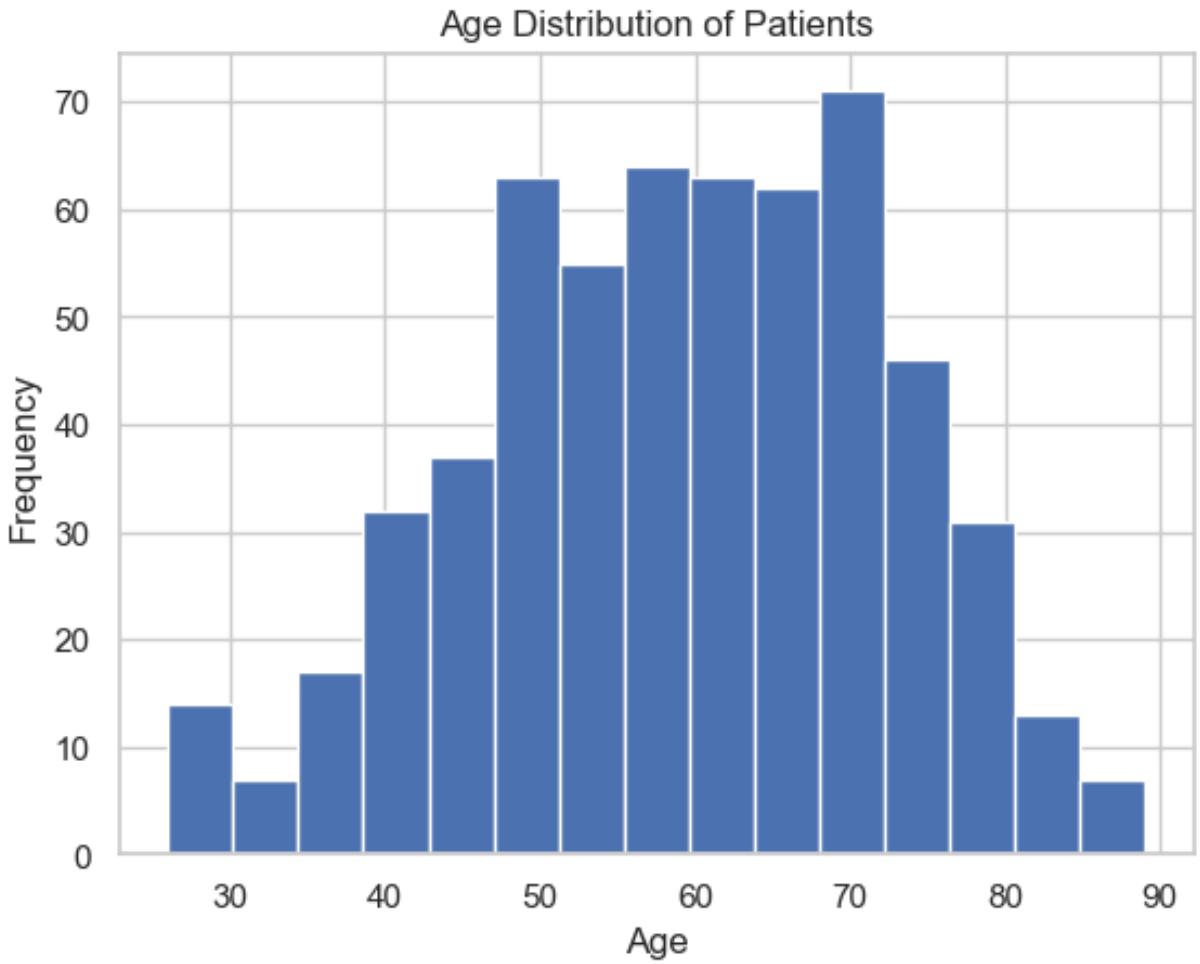
- The columns 'Plasma_CA_19_9' and 'REG1A' both have missing values at random points in the dataset. We could consider to drop these columns or fill in the missing values. After doing some online research, it was determined that dropping these columns would not be beneficial for modeling as both columns are important indicators of pancreatic cancer. To fill in the missing values, a mean and std was found for 'Plasma CA19 9' and 'REG1A' based on stage level, and randomly generated to fill in the empty cells.

- Lastly we removed 'sample origin' as it is not needed.

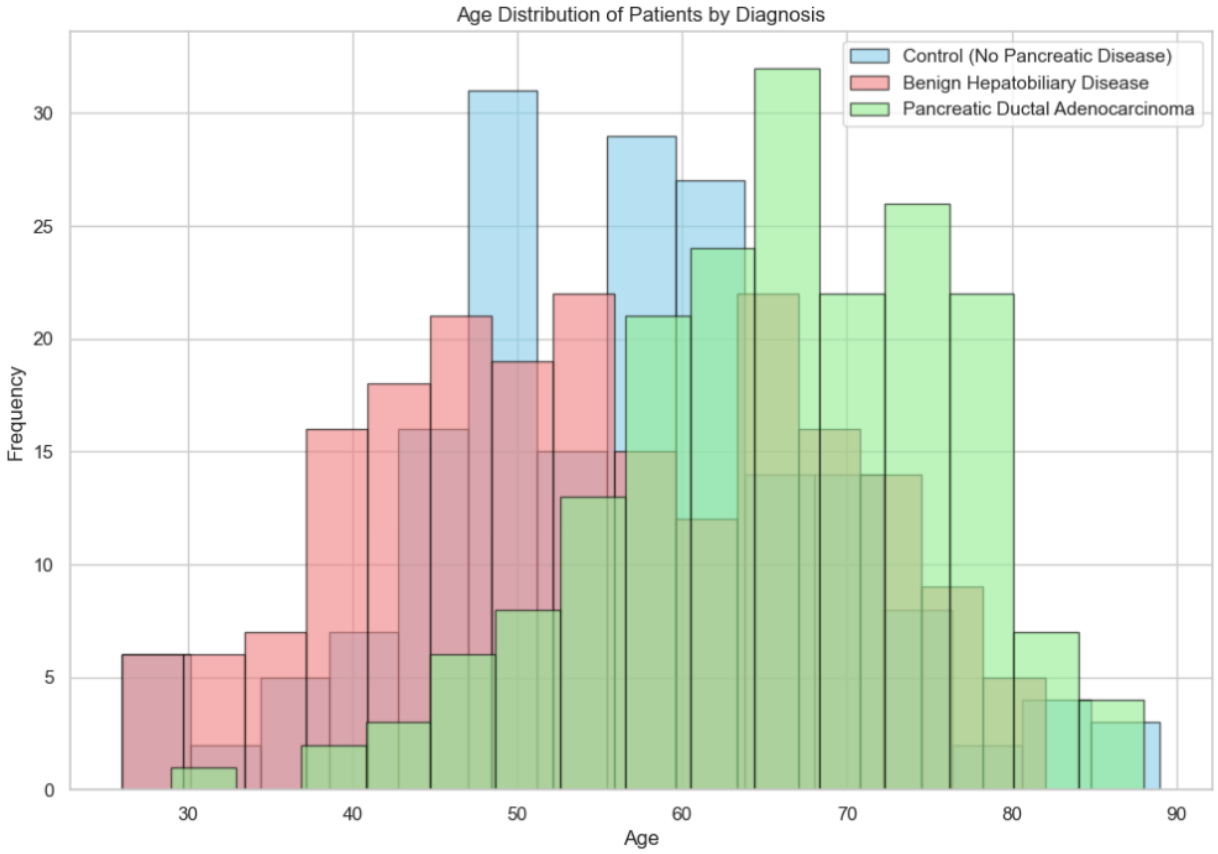
EDA Findings

1. **Age Distribution**:

```
The average age of the patients in the dataset is 59.0 years.
```

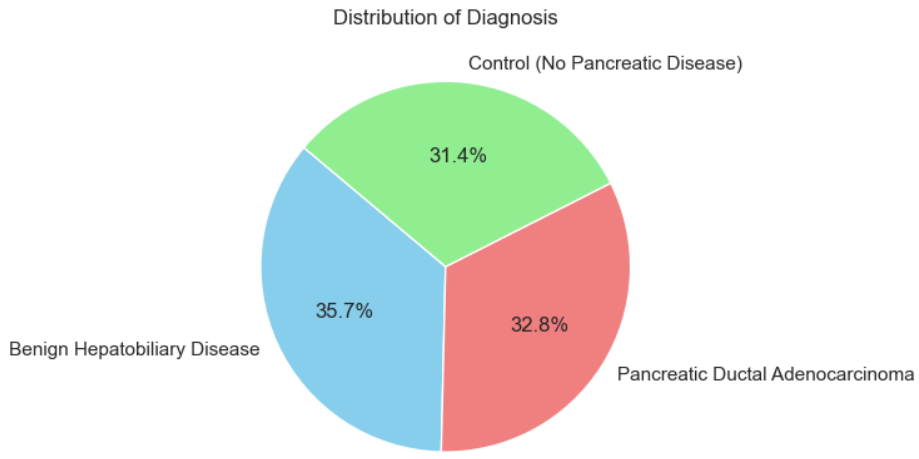


We will group patients by age and diagnosis. Diagnosis 1 means control (no pancreatic disease). 2 means benign hepatobiliary disease (119 of which are chronic pancreatitis). And 3 means Pancreatic ductal adenocarcinoma, i.e. pancreatic cancer.
The average age of patients with diagnosis 1 is 56.3 years.
The average age of patients with diagnosis 2 is 54.7 years.
The average age of patients with diagnosis 3 is 66.1 years.



2. **Diagnosis Distribution**:

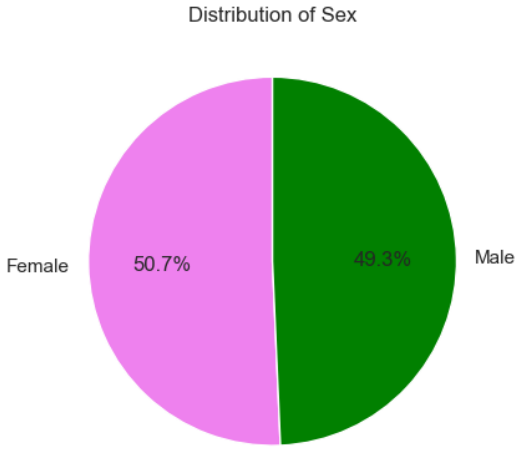
To get a better understanding of the distribution of diagnoses, we will use a pie chart.



Based on this pie chart we can see that the amount of people from each diagnosis level is distributed relatively even.

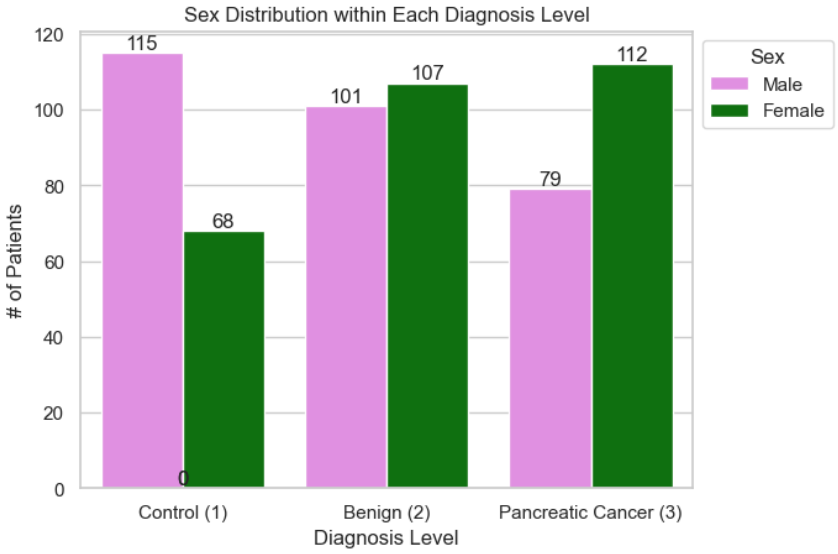
3. **Sex Distribution**:

Now lets take a look at sex. What is the average age for males and females in this study, and what portion of dataset do they make up?
The average age of patients who are F is 58.5 years.
The average age of patients who are M is 59.4 years.



Seeing that the average age of both males and females in this study is relatively close, we can assume there is balance in the data and that both profiles are comparable.

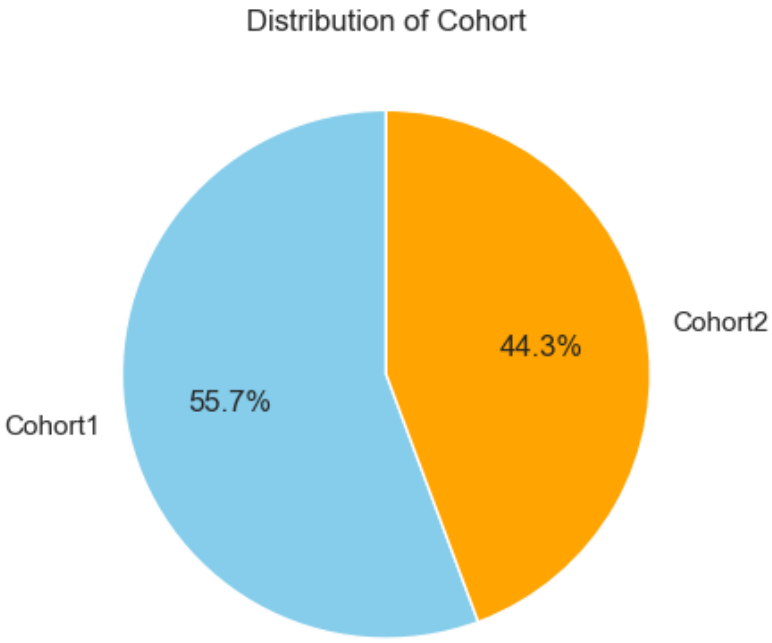
Lets take a look at sex distribution within each diagnosis level to see if there are any gender differences.



Looking at the chart. In the Control diagnosis group less males are prevalent. In the benign group the distribution seems fairly balanced. And in the cancer group males are more prvelant. This could indicate that pancreatic cancer is more common amongst males.
A chi-square test could be conducted to see if there are great differences within the distribution.

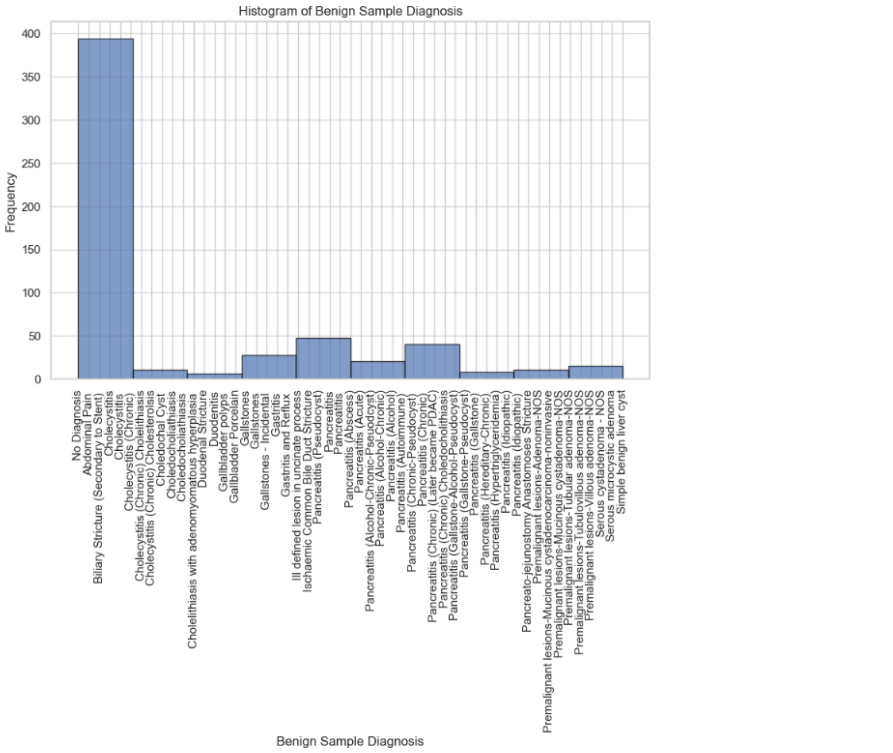
4. **Cohort Distribution**:

We will also take a quick look at the distribution of patient cohort.



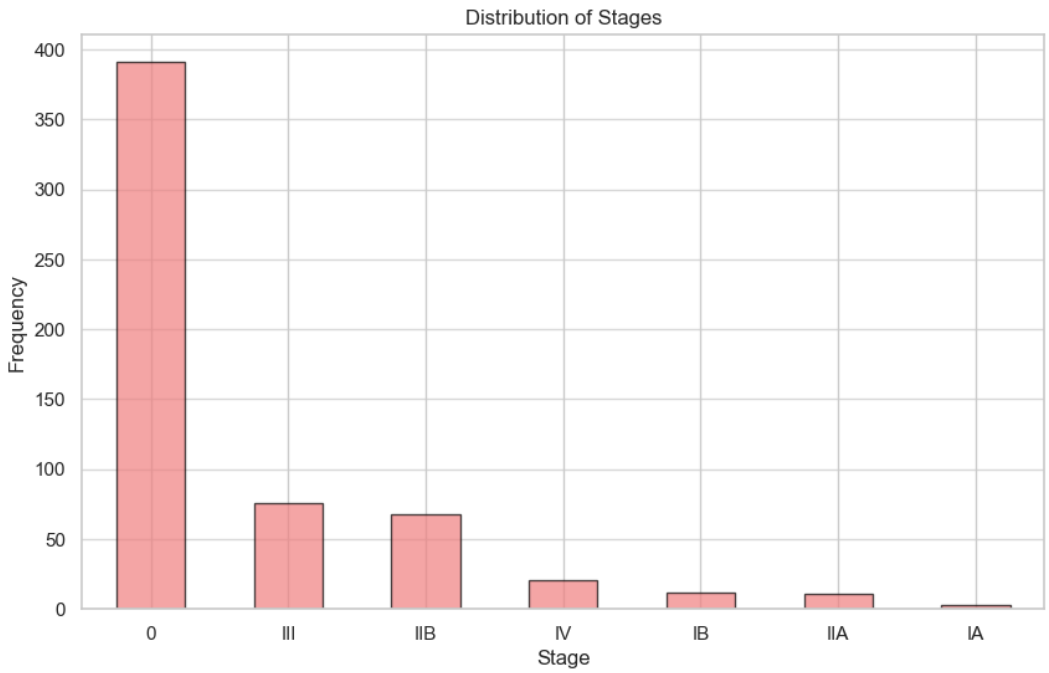
5. **Benign Sample Diagnosis**:

Now that we have an understanding of basic patient demographics. Lets dive into the stage and diagnosis samples that each of these patients have.



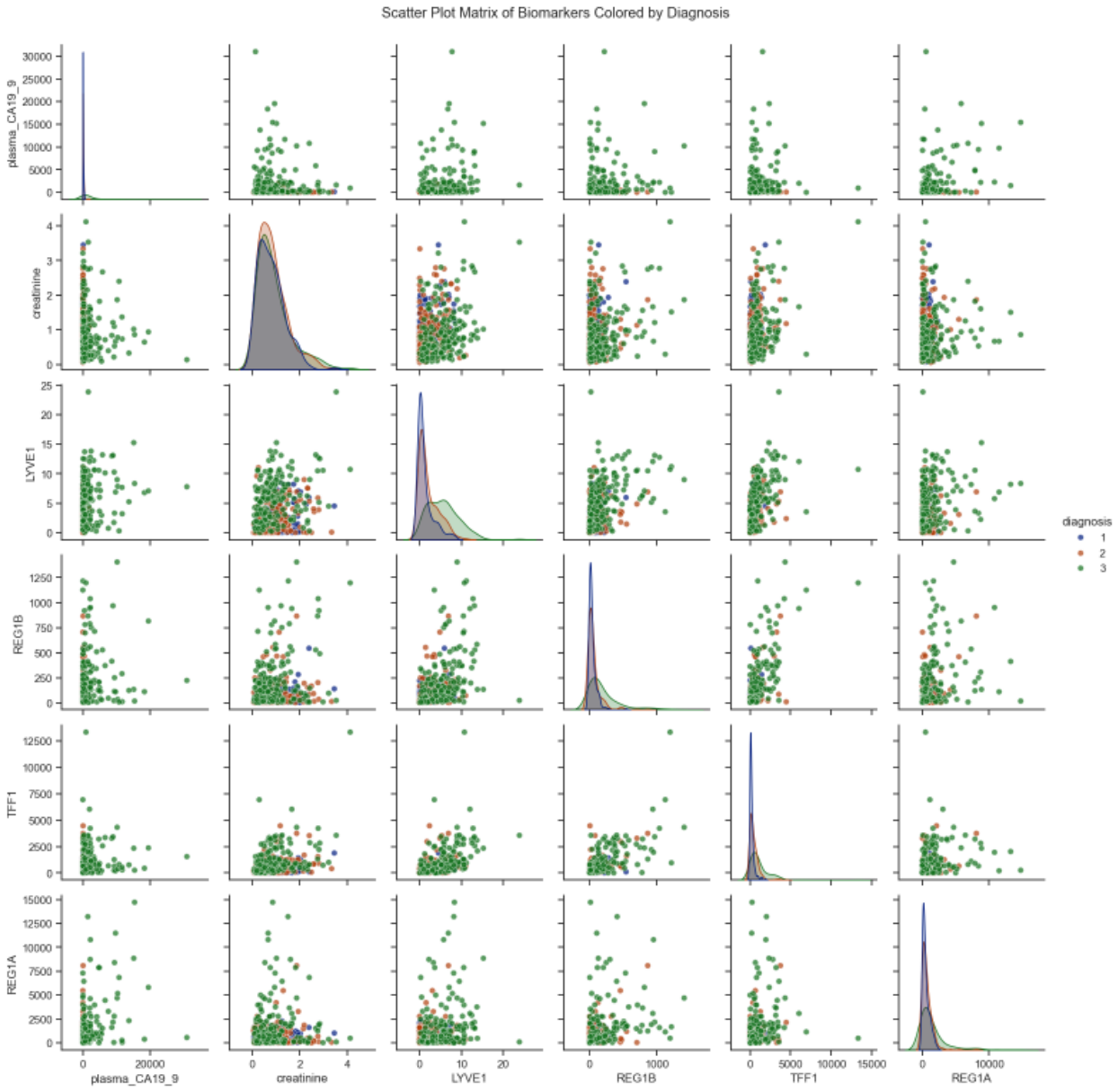
No diagnosis, abdominal pain, biliary stricture, and cholecystitis seem to make up most of the benign sample diagnoses.

6. **Stage Distribution**:



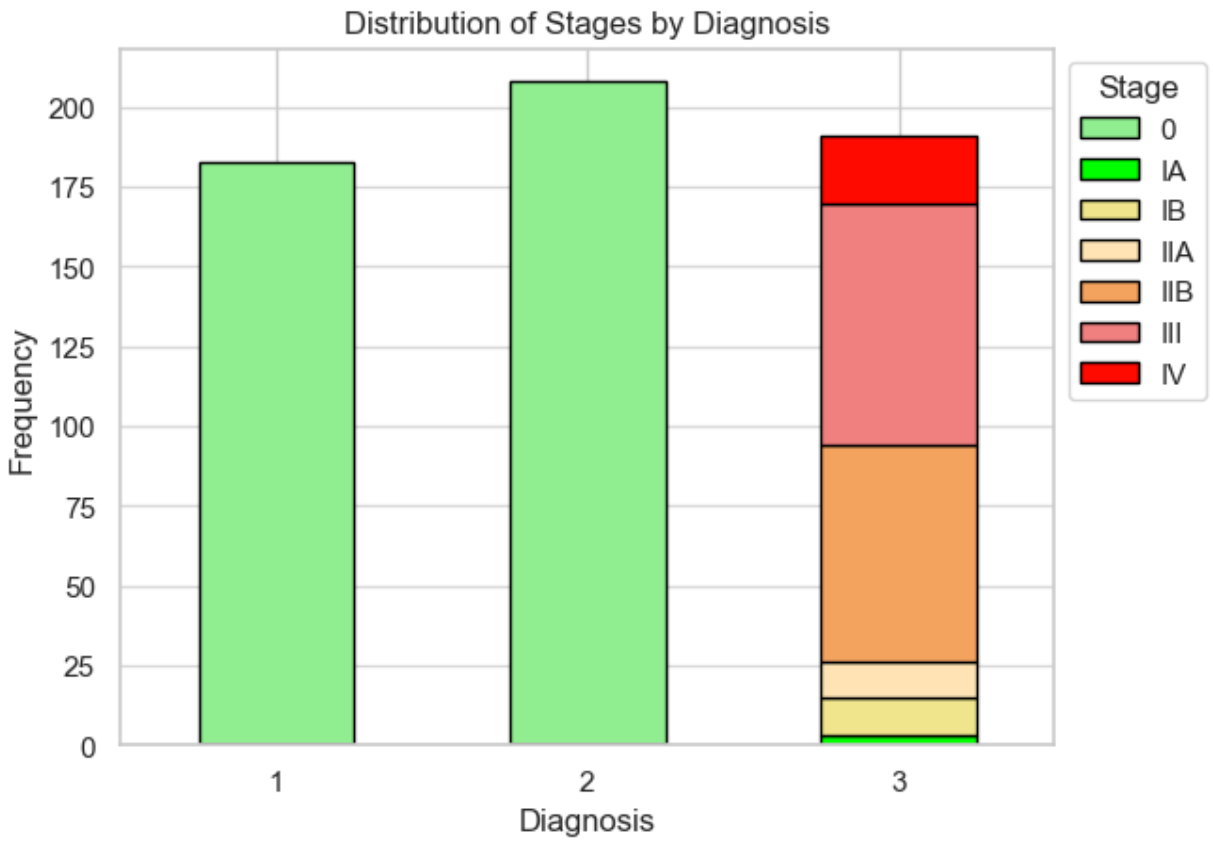
We can see two things here. First thing is the biggest group of people are stage 0 but the other half of the data are patients with a stage designation.
The second thing we can see is that from the half of the patients who have a stage level designation, stage III and IIB are most common amonst people in this study.

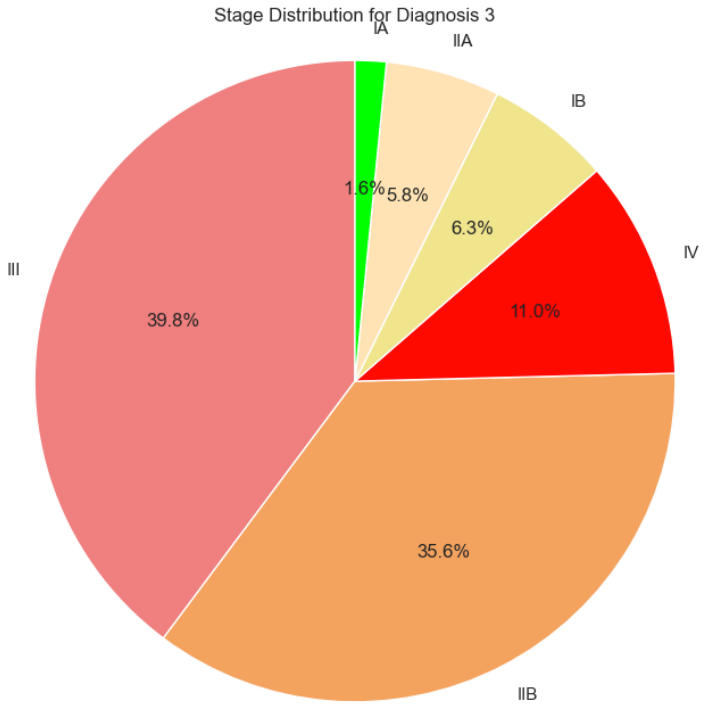
7. **Biomarkers Analysis**:



The correlation patterns show that biomarker relationships strengthen from control to benign conditions, and are most pronounced in pancreatic cancer, with particularly strong correlations among LYVE1, REG1B, and TFF1. Additionally, a unique correlation between plasma_CA19_9 and REG1A in pancreatic cancer suggests a distinct biomarker interaction specific to this disease.

8. **Grouped Analysis**:





Here we have a pie chart showing the various percentages of people and what stage of ductal adenocarcinoma they have been classified with.

Preprocessing

The initial plan is to make two models. 1 using binary classification, and 1 using multiclass classification. Both models will use different target columns. For binary we will add a new column called ‘cancer_present’ that designates anyone with ‘stage’ 0 as ‘N’, and others with ‘Y’.

Drop sample_id column as it is not needed since its values are unique identifiers for each row and does not help with training models.

After adding the new column, we will need to perform encoding on the categorical data types.

Patient_cohort, sex, cancer_present, stage.

Dropped benign_sample_diagnosis since the column contained more than 60% empty values which were changed to No diagnosis.

Final Dataset information

<class 'pandas.core.frame.DataFrame'>
Index: 582 entries, 0 to 589
Data columns (total 12 columns):

#	Column	Non-Null Count	Dtype
0	patient_cohort	582 non-null	int64

```
 1  age             582 non-null    int64
 2  sex             582 non-null    int64
 3  diagnosis       582 non-null    int64
 4  stage           582 non-null    int64
 5  plasma_CA19_9   582 non-null    float64
 6  creatinine      582 non-null    float64
 7  LYVE1           582 non-null    float64
 8  REG1B           582 non-null    float64
 9  TFF1            582 non-null    float64
 10 REG1A           582 non-null    float64
 11 Cancer_Present  582 non-null    int64
dtypes: float64(6), int64(6)
memory usage: 59.1 KB
```

Findings

I trained my dataset to be used two ways, binary classification and multiclass classification. In binary classification I saw a perfect performance of the model, but this is most likely due to the dataset being relatively small. For my multiclass models I chose Random Forest and gradient Boosting for a couple reasons. First, these models are good at handling and capturing complex non-linearly related datasets. The are also good at overfitting which is helpful to my type of data. And these models can help distinguish important biomarkers in the data. Upon running the multiclass models we saw an accuracy of around 89-91% for both with gradient boosting performing slightly better, and they had similar precision, recall, and F1 score. Both models effectively distinguished between the three classes (no pancreatic disease, benign hepatobiliary, and ductal adenocarcinoma), with particularly high precision and recall for the cancer class, indicating reliable identification of this critical condition.

Best Parameters for Random Forest: {'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100}
Best Cross-Validation Accuracy: 0.81
Looking at these results we can see the minimum samples to be at a leaf node to prevent overfitting is 2, the minimum number of samples required to split a node is 2 which allows for deeper trees. This is the average accuracy across the cross-validation folds using the best hyperparameters. It provides a more reliable estimate of the model's performance compared to a single train-test split.

Best Parameters for Gradient Boosting: {'learning_rate': 0.1, 'max_depth': 3, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 50}
Best Cross-Validation Accuracy: 0.81
For gradient boosting we see that first off we have a low learning rate meaning more trees would be required to achieve good performance. The max_depth is 3 samples to limit it to prevent overfitting. The biggest difference is the minimum amount of samples required to split is 10.

Comparing the two random forest uses larger number of trees and can capture complex patterns at the risk of overfitting. Gradient boosting uses less amount trees which helps with preventing overfitting and can be a better performing model for multiclass with the only downside being the processing speed is significantly higher than random forest.

Results and Conclusion

Future Research and Development

Next Steps and Recommendations