HEALTHCARE DATA INTEGRATION FOR HEP-C PREDICTION

- Team members
  - Siyona Behera, Gina Seo
- Contact Member
  - Gina Seo
- Project ID
  - A9

BACKGROUND

You can provide

- The topic chosen for this project is Hepatitis C. Hepatitis C is a viral infection that causes inflammation within the liver. The virus can cause both acute and chronic hepatitis, ranging in severity from a mild illness to a serious, lifelong illness including liver cirrhosis and cancer. The bloodborne virus is spread through exposure to infected blood from unsafe injection practices, unsafe health care, unscreened blood transfusions, injection drug use, and sexual practices that lead to exposure to blood. When it comes to symptoms, most people do not show in the first weeks after infection. It can take between two weeks to six months for the symptoms to appear.

Hypothesis

  - Suspected blood donors are more likely to contract Hep-C as compared to regular blood donors because of their possible elevated ALT, AST, and bilirubin levels.

- This project is to create a predictive model that can determine whether or not a blood donor can develop HepC

DATA PREPROCESSING, INTEGRATION, AND FUSION

Data Preprocessing
-Explore Class Distribution: Check for imbalance in target classes.
-Missing Values: Handle missing values
-Data Cleaning: Remove `Unnamed: 0` as it's just an index, Convert categorical columns like **Sex** into numerical format.
-Outlier Detection: Use z-score or IQR methods on continuous variables.
Data Integration and Fusion
•Data Fusion: Synthesize features: e.g., liver enzyme ratios (**ALT/AST**, **ALP/ALT**, etc.).
•External Integration: Merged dataset packages together to include protein enzymes. Add derived features like Sex

	Category	Age	Sex	ALT	AST	BIL
0	0=Blood Donor	32	m	7.7	22.1	7.5
1	0=Blood Donor	32	m	18.0	24.7	3.9
2	0=Blood Donor	32	m	36.2	52.6	6.1
3	0=Blood Donor	32	m	30.6	22.6	18.9
4	0=Blood Donor	32	m	32.6	24.8	9.6
...	...	...	...	...	...	...
559	1=Hepatitis	58	m	12.2	63.2	13.0
560	1=Hepatitis	33	f	3.8	16.7	6.0
561	1=Hepatitis	41	f	8.2	38.3	7.0
562	1=Hepatitis	50	f	9.0	46.0	10.0
563	1=Hepatitis	61	f	27.4	114.4	22.0

564 rows × 6 columns

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106	12.1	69
2	0=Blood Donor	32	m	38.5	70.3	18	24.7	3.9	11.17	4.8	74	15.6	76.5
3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.2	86	33.2	79.3
4	0=Blood Donor	32	m	43.2	52	30.6	22.6	18.9	7.33	4.74	80	33.8	75.7
5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76	29.9	68.7
6	0=Blood Donor	32	m	41.6	43.3	18.5	19.7	12.3	9.92	6.05	111	91	74
7	0=Blood Donor	32	m	46.3	41.3	17.5	17.8	8.5	7.01	4.79	70	16.9	74.5
8	0=Blood Donor	32	m	42.2	41.9	35.8	31.1	16.1	5.82	4.6	109	21.5	67.1
9	0=Blood Donor	32	m	50.9	65.5	23.2	21.2	6.9	8.69	4.1	83	13.7	71.3
10	0=Blood Donor	32	m	42.4	86.3	20.3	20	35.2	5.46	4.45	81	15.9	69.9
11	0=Blood Donor	32	m	44.3	52.3	21.7	22.4	17.2	4.15	3.57	78	24.1	75.4
12	0=Blood Donor	33	m	46.4	68.2	10.3	20	5.7	7.36	4.3	79	18.7	68.6
13	0=Blood Donor	33	m	36.3	78.6	23.6	22	7	8.56	5.38	78	19.4	68.7
14	0=Blood Donor	33	m	39	51.7	15.9	24	6.8	6.46	3.38	65	7	70.4
15	0=Blood Donor	33	m	38.7	39.8	22.5	23	4.1	4.63	4.97	63	15.2	71.9

	Category	Age	Sex	ALT	AST	BIL
0	0=Blood Donor	32	0	7.7	22.1	7.5
1	0=Blood Donor	32	0	18.0	24.7	3.9
2	0=Blood Donor	32	0	36.2	52.6	6.1
3	0=Blood Donor	32	0	30.6	22.6	18.9
4	0=Blood Donor	32	0	32.6	24.8	9.6

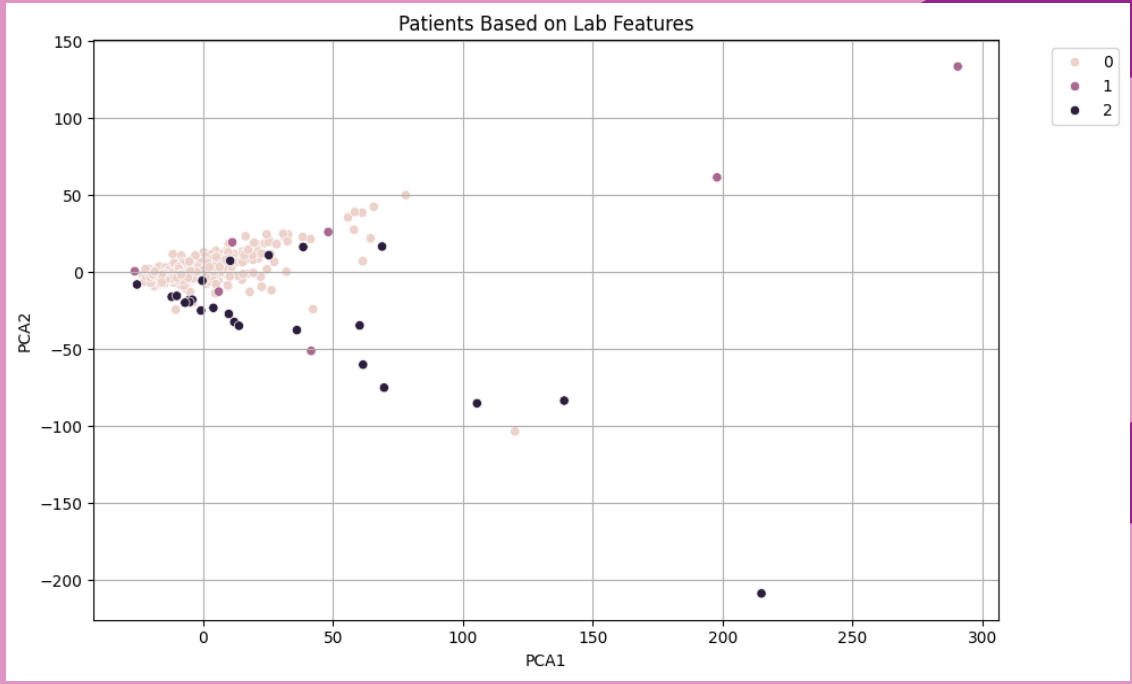
STATISTICAL ANALYSIS

Kruskal Wallis

- **Category (Target Variable):** Very strong separation with a test statistic of 563.0 and an extremely low p-value (5.57×10^{-123}), confirming significant class differences overall
- **Age:** p-value = 0.0001
- **Sex:** p-value = 0.027
- **AST :** p-value = 5.66×10^{-12}
- **BIL :** p-value = 1.13×10^{-5}
- **ALT :** p-value = 0.084

PCA

- **Clusters:** Class 0 forms a dense cluster near the origin, indicating consistent and lower variability in their lab values.
- **Overlap:** Classes 1 and 2 are more dispersed. There's partial overlap with class 0, but also outliers spread further along both PCA axes
- **Outliers:** A few extreme outliers likely correspond to patients with severe lab profiles
- **Insight:** While Blood Donors can often be separated from other categories, classes 1 and 2 are harder to distinguish linearly– supporting the need for more complex or non-linear models




```
{'Category': {'Statistic': np.float64(563.0000000000001),
  'P-value': np.float64(5.573183520129081e-123)},
 'Age': {'Statistic': np.float64(18.24990038584148),
  'P-value': np.float64(0.00010891419441788837)},
 'Sex': {'Statistic': np.float64(7.216635406567609),
  'P-value': np.float64(0.027097394400992587)},
 'ALT': {'Statistic': np.float64(4.95083471078968),
  'P-value': np.float64(0.08412787170824834)},
 'AST': {'Statistic': np.float64(51.79510851290985),
  'P-value': np.float64(5.660243199908075e-12)},
 'BIL': {'Statistic': np.float64(22.790280902863902),
  'P-value': np.float64(1.1250022251193324e-05)}}
```

MACHINE LEARNING

Logistic Regression:

- **Accuracy:** 96.5%
- **Precision/Recall for Blood Donors (class 0):** Precision was 0.97, recall at 1.0
- **For Hepatitis (class 2):** Precision was 1.0 but recall was 0.4, indicating false negatives
- **For Suspect Donors (class 1):** The model didn't correctly identify the only test sample, resulting in 0 precision and recall

Random Forest :

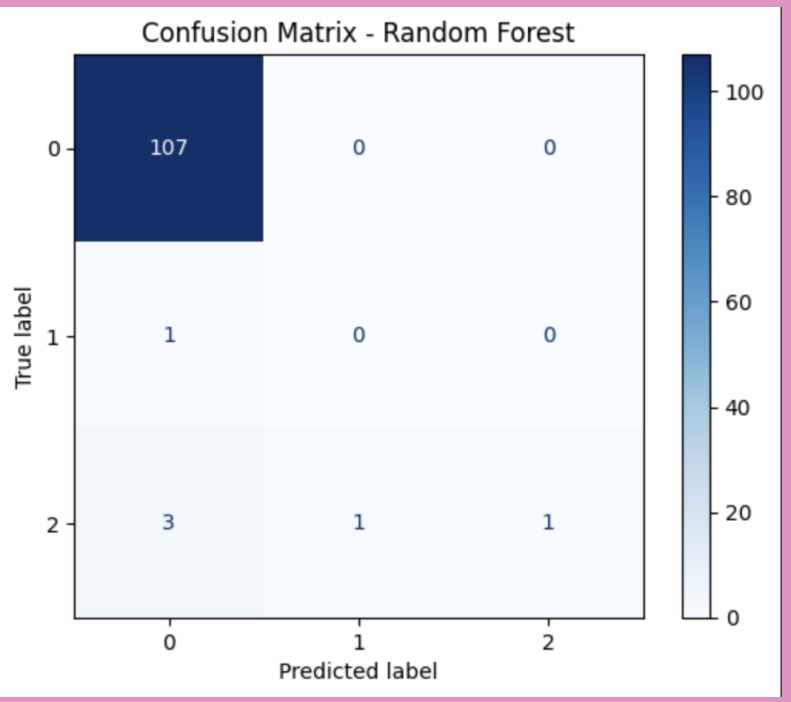
- **Accuracy:** 95.6%
- **Blood Donors (class 0):** excellent performance—recall was 1.0
- **Hepatitis (class 2):** Precision remained at 1.0, but recall dropped to 0.2
- **Suspect Donor (class 1):** Still 0 precision and recall

These results show that while both models perform well on the dominant class, they struggle with underrepresented categories—particularly Suspect Donors and Hepatitis cases. This highlights the importance of addressing class imbalance, possibly through techniques like SMOTE or adjusting class weights.

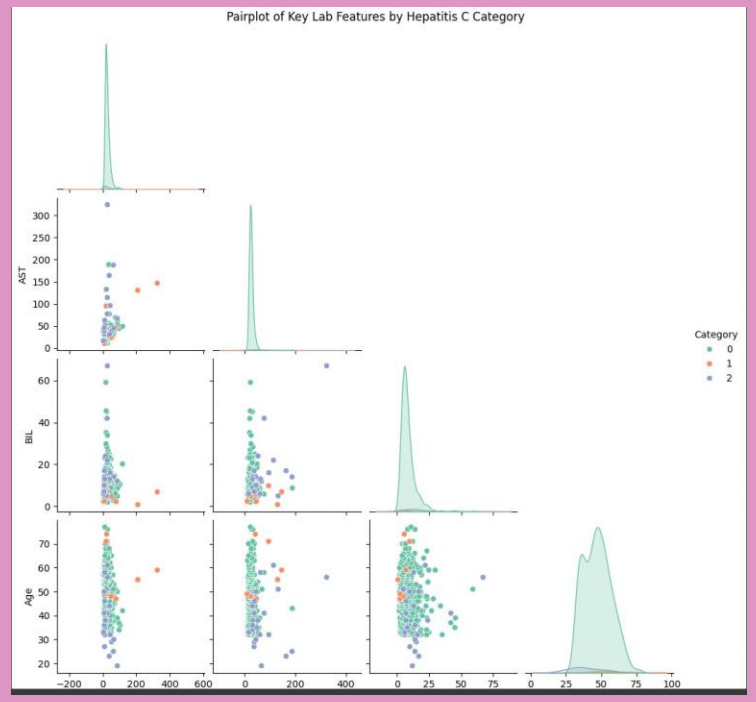

```
{'Logistic Regression': {'Classification Report': {'0': {'precision': 0.9727272727272728,
    'recall': 1.0,
    'f1-score': 0.9861751152073732,
    'support': 107.0},
   '1': {'precision': 0.0, 'recall': 0.0, 'f1-score': 0.0, 'support': 1.0},
   '2': {'precision': 1.0,
    'recall': 0.4,
    'f1-score': 0.5714285714285714,
    'support': 5.0},
   'accuracy': 0.9646017699115044,
   'macro avg': {'precision': 0.6575757575757576,
    'recall': 0.4666666666666666,
    'f1-score': 0.5192012288786482,
    'support': 113.0},
   'weighted avg': {'precision': 0.9653258246178601,
    'recall': 0.9646017699115044,
    'f1-score': 0.9590962848170956,
    'support': 113.0}}},
 'Random Forest': {'Classification Report': {'0': {'precision': 0.963963963963964,
    'recall': 1.0,
    'f1-score': 0.981651376146789,
    'support': 107.0},
   '1': {'precision': 0.0, 'recall': 0.0, 'f1-score': 0.0, 'support': 1.0},
   '2': {'precision': 1.0,
    'recall': 0.2,
    'f1-score': 0.3333333333333333,
    'support': 5.0},
   'accuracy': 0.9557522123893806,
   'macro avg': {'precision': 0.6546546546546547,
    'recall': 0.39999999999999997,
    'f1-score': 0.4383282364933741,
    'support': 113.0},
   'weighted avg': {'precision': 0.9570278242844615,
    'recall': 0.9557522123893806,
    'f1-score': 0.9442775567643638,
    'support': 113.0}}}}
```

EXPLORATORY DATA ANALYSIS

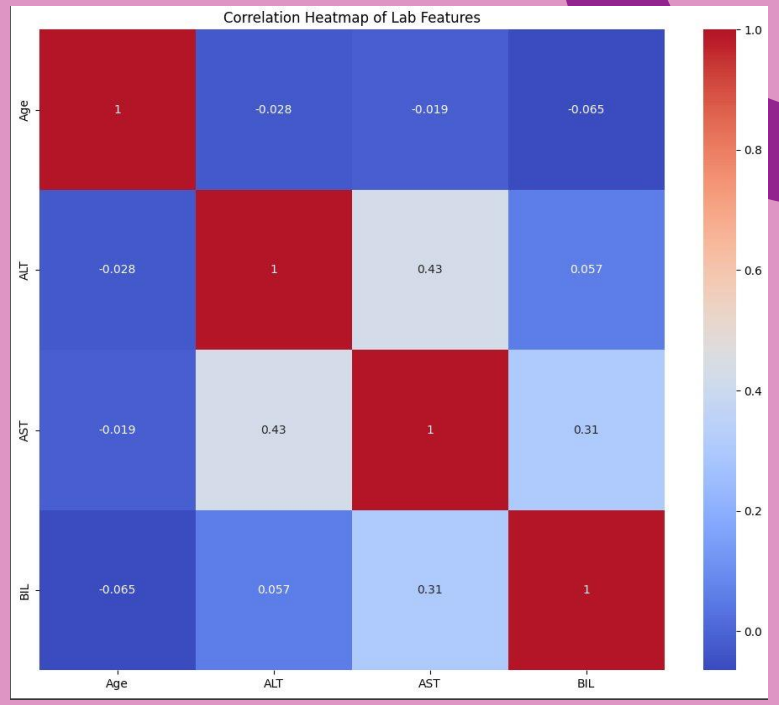
Confusion Matrix



Pair plot



Heat Map



GENERATIVE AI

How we used ChatGPT

- Created a timeline to keep track of our progress
- Debugged any errors on our code
- Overall, called for the organization of the project and kept it neat

RESULTS & CONCLUSION

- What did we see from the model?
  - •Accurately predicted healthy donors
  - •Struggled with Hep-C cases (Classes 1 & 2)
  - •Class imbalance affected performance
- Was it supporting our hypothesis or not?
  - •**Partially yes** – liver enzymes like ALT & AST helped
  - •But not enough to clearly separate Hep-C cases
- Depending on our interpretations, how can this be applied in the real world?
  - •Shows potential for early screening
  - •Could be used as a **clinical decision support tool**
  - •Needs more data & balanced classes for stronger results

LINKS

- https://github.com/ginaaseo/DS320-Final-Project
- https://drive.google.com/file/d/1P1-3vEBO1GmT13VSb-a_mi57dBZxZB2f/view?usp=drive_link