**Model Development Phase**

|  |  |
| --- | --- |
| Date | 20th June 2025 |
| Team ID | LTVIP2025TMID44055 |
| Project Title | Revolutionizing Liver Care : Predicting Liver  Cirrhosis Using Advanced Machine Learning Techniques |
| Maximum Marks | 5 Marks |

**Feature Selection Report**

In the forthcoming update, each feature will be accompanied by a brief description. Users will indicate whether it's selected or not, providing reasoning for their decision. This process will streamline decision-making and enhance transparency in feature selection.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Description** | **Selected (Yes/No)** | **Reasoning** |
| Age | It is a numeric column that represents age of an individual | Yes | This data is more widespread among both the classes and would be efficient in explaining the target variable |
| Quantity of alcohol consumption (quarters/day) | It isa numeric column that has values ranging from 1 to 5 | Yes | Alcohol consumption has achieved a good feature importance and would be a good feature to explain the target. |
| Diabetes Result | It is an object  column which has values YES and NO | Yes | Diabetes provides a good base to  diagnose liver cirrhosis |

|  |  |  |  |
| --- | --- | --- | --- |
| Blood pressure (mmhg) | It is an object column that represent the BP of an individual | Yes | In the final model it was found out that it has an importance score of about 0.04.Which makes it a good feature to assess the target |
| PCV (%):  Polymorphs  Lymphocytes  Platelet Count (lakhs/mm)  Indirect | All these are numeric columns that indicate several lab results provided by an individual | Yes | All these features had a relatively good importance score of more than 0.07 in the final model which states that they influence the output pretty well |
| Haemoglobin | It is a numeric column that represents the total  Haemoglobin  levels | Yes | Liver disease is associated with a wider range of Haemoglobin levels.    No liver disease shows more consistent Haemoglobin levels centered around 11.5 g/dl.    This makes it a good feature to be taken |
| Total Protein | It is a numeric column that represents the total Protein levels | Yes | Patients with liver cirrhosis ("yes") have a wider distribution of total protein levels ranging from approximately 3 g/dl to 9 g/dl.  Patients without liver cirrhosis ("no") have a slightly narrower distribution, with total protein levels ranging from approximately 4.5 g/dl to 8 g/dl.    This make it a good feature to include |

|  |  |  |  |
| --- | --- | --- | --- |
| AL.Phosphatase            USG Abdomen | It is a numeric column that represents the phosphate levels.      It is an object column that states whether a person has diffused liver or not | Yes | Both of these features had the highest importance score of 0.1 and 0.2 which makes them a good feature to be taken to predict the target. |
| Type of alcohol consumed Gender  Direct  MCH  MCHC  Obesity  Family history of cirrhosis/ hereditary  TCH  LDL  HDL  MCV  Total Count  Monocytes  Basophils (%)  SGOT/AST | Combination of numerical and  categorical columns representing lifestyle,lab results taken. | No | All of these features either had negligible importance score or were highly inefficient . Thescores would range from 0.00 – 0.003 which makes them highly inefficient to predict the target. Hence they were removed |
| SGPT/ALT  RBC  Quantity of alcohol consumption  Eosinophils  TG  Hepatitis B infection  Hepatitis C infection |  |  |  |
| Duration of alcohol consumption  Total Bilirubin | Both these are numerical which depict lab results | No | Both of them had a very high score which made the model completely biased. The model only took these two rows without giving importance to any other features. Hence these were dropped. |