1 Introduction and Data Description

The dataset used in my analysis is the *Diabetes prediction* dataset, which consists of various patient attributes and the final diabetes diagnosis. The data was obtained from Kaggle, a popular online data science resource platform, and can be found here. The dataset contains 8 attributes, including gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level, which were used to explore the relationship between these factors and the likelihood of developing diabetes. There are 100 000 instances in the dataset, each representing a unique patient record.

2 Preliminary Data Analysis

To automatically create the network, it was necessary to discretize some attributes. This was accomplished through the GeNIe 4.0 Academic software. The final form of the attributes and the possible values they take are shown in table 1.

Gender	Age	Hypertension	Heart Disease	Smoking History	BMI	HbA1c Level	Blood Glucose Level
Male, Female, Other	s1_below_10, s2_10_20, s3_20_30, s4_30_40, s5_40_50, s6_50_60, s7_60_70, s8_70_up	0, 1	0, 1	current, ever, former, never, no_info, not_current	b1_below_20, b2_20_31, b3_31_42, b4_42_52, b5_52_63, b6_63_74, b7_74_84, b8_84_up	h1_below_4, h2_4_6, h3_6_7, h4_7_up	State80, State85, State90, State100, State126, State130, State140, State145, State155, State158, State159, State160, State200, State220, State240, State260, State280, State300

Table 1: Attributes and their possible values.

When trying to generate a Bayesian network using GeNIe, counter-intuitive relationships were created between some attributes. As a result, it was decided to create the network manually to better reflect the relationships and dependencies between the attributes in the dataset. The created network is shown in next section in the fig. 3.1.

3 Building the Belief Network

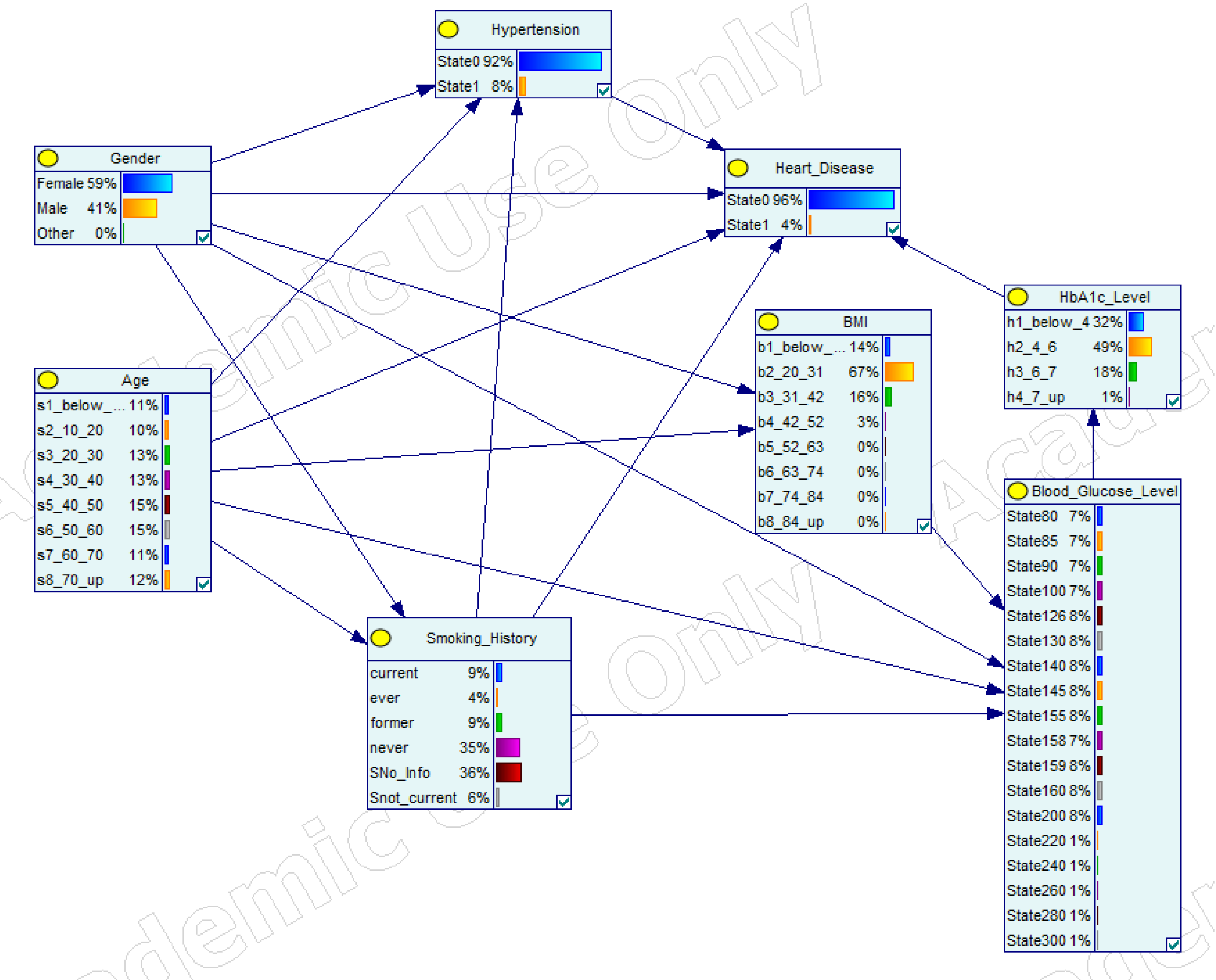

Figure 3.1: Built network.

Due to the fact that these are medical data, attributes such as gender and age have a direct impact on most of the available parameters. Smoking cigarettes has a direct impact on high blood pressure, heart disease, and also blood sugar levels. Ultimately, the HbA1c level, which measures the amount of blood sugar attached to your hemoglobin in past 2–3 months is influenced by the glucose level.

4 Probabilistic Queries

The analysis of the impact of individual attributes on others was made, thanks to the possibility of setting specific evidence in the network.

1. Selecting only male patients translates to an increase in heart disease to **6%** (+2%), current smokers to **11%** (+2%), and a minimal increase in HbA1c in the h3_5_7 to **19%** (+1%) and h4_7 **to 2%** (+1%) ranges, which translates into the risk of developing diabetes. The percentage distribution of BMI also changes a bit, the probability is higher in the b2_20_31 range (+3%) and lower in the b1_below_20 range (-2%). Therefore, it can be concluded that men are more prone to heart disease, more of them currently smoke and have a BMI in the range containing higher values.

2. By setting the glucose level to 220, there was a significant increase in patients with high HbA1c levels in the h3_6_7 (+21%) and h4_7_up range (+24%). Thus, the low, healthy ranges decreased in h1_below_4 (-32%) and h2_4_6 (-12%) respectively. From the data obtained, it can also be concluded that the set glucose level mostly applies to patients over 50 and those who currently smoke or have smoked in the past. High blood sugar also translates into an increased risk of hypertension (+7%) and heart disease (+8%). Thus, it can be concluded that people with such a high level of sugar are usually old people who used to smoke or currently smoke. They have an increased level of HbA1c, which is associated with the risk of diabetes

3. By setting the age of the patients to under 10 years, it was observed that the chance of hypertension (-8%) and heart disease (-6%) was reduced to **0%**. Regarding smoking history, the variable with no information is the most likely one (**86%**), followed by never (**12%**). A change in the distribution of BMI was also observed, where the probability of b1_below_20 level is **73%**, and then b2_20_31 level is **27%**. Thus, it can be concluded that children have a negligible risk of hypertension, heart disease. There is also no information about their smoking history or they do not smoke at all, which is a logical conclusion due to their age. Most of them are in the correct BMI range.

5 Decision Analysis

The next step was to add the Utility and Decsion blocks. The utility block was directly connected to the HbA1c node. The decision itself concerns whether a given patient has diabetes and thus should start treatment, or whether he does not have diabetes and no further action is needed. The definition of the Utility block is presented in the table 2. Figure 5.1 shows the network with added blocks.

HbA1C_Level	h1_below_4		h2_4_6		h3_6_7		h4_7_up	
Decision	*HasDiabetes*	*DoesntHaveDiabates*	*HasDiabetes*	*DoesntHaveDiabates*	*HasDiabetes*	*DoesntHaveDiabates*	*HasDiabetes*	*DoesntHaveDiabates*
Value	2	98	6	94	15	85	85	15

Table 2: Utility block definition.

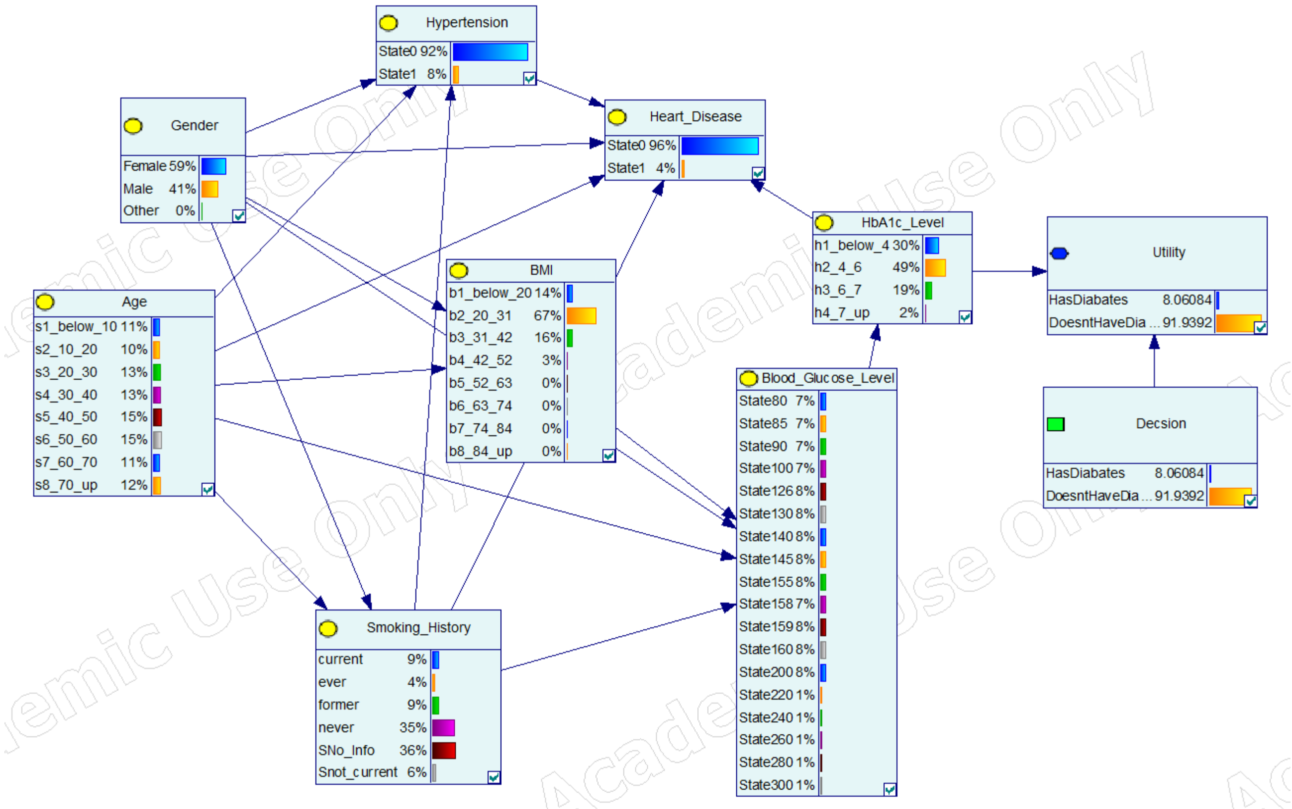

Figure 5.1: Network with added blocks.

The result of the decision based on the network used is that 8.06084% of the patients in the used dataset are diabetic. Considering that the number of such people in the original dataset was 8.5%, it can be concluded that the network correctly deduces decisions based on the data provided to it and the relationship between the attributes.

6 Decision Cases

1. Setting the highest available glucose level in the network, i.e. 300, increases the decision about diabetes to 28.043% (+19.98216%), thus reducing the chance of not having diabetes to 71.957% (-19.822%). Similarly, setting the lowest glucose level, i.e. 80, reduces the decision about diabetes to 6.2582% (-1.80264%). It can be concluded that the network behaves intuitively during this test because it significantly increases the chance of diabetes in people with high blood sugar levels and similarly lowers it for people with low blood sugar levels.

2. Similarly, the network behaves in the case of setting the highest age of the patient, i.e. 70 years and above. Here, the decision on the presence of diabetes is 9.02323% (+0.96239%). Setting the smallest available age, i.e. under 10 years of age, the chance of diabetes drops to 7.39288% (-0.66796%). Again, the network seems to make logical decisions because older people tend to have a higher risk of having diabetes than younger people.

7 Value of Perfect Information

The VPI was calculated using the *Value of Information* function built into GeNie. In the beginning, I analyzed the results for individual nodes, which I placed in table 3.

Node	*HbA1c_Level*	*BMI*	*Hypertension*	*Blood_Glucose_Level*	*Smoking_History*	*Gender*	*Age*	*Heart_Disease*
VPI	1.3995644	2.8421709e-14	1.4210855e-14	1.4210855e-14	1.4210855e-14	0	0	0

Table 3: VPI for individual nodes.

According to the network model, the HbA1c_Level node has the highest VPI value because it enters directly into the Utility block. Then I considered the VPI from the perspective of HbA1c by changing the states in the BMI node. The results are presented in the table 4.

HbA1c_Level	*no set state*	*b1_below_20*	*b2_20_31*	*b3_31_42*	*b4_42_52*	*b5_52_63*	*b6_63_74*	*b7_74_84*	*b8_84_up*
VPI	1.3995644	0.97896232	1.2965684	1.9894069	2.3876637	3.0104074	4.462898	5.4324572	4.6259085

Table 4: VPI for HbA1c_Level node with changing BMI state.

It seems that from the perspective of the decision block, knowing the BMI value from certain ranges increases the VPI value in the HbA1c_Level node.